INTELIGENCIA ARTIFICIAL APLICADA A LA INVESTIGACIÓN CIENTÍFICA EN CIENCIAS VETERINARIAS

Fundamentos, métodos, herramientas y aplicaciones para la generación de evidencia científica



Víctor Alejandro Bosquez Barcenes
Joscelito Bolívar Solano Gaibor
Danilo Fabian Yanes Silva
Jonnathan Vinicio Arias Real

ISBN: 978-9907-818-68-0

INTELIGENCIA ARTIFICIAL APLICADA A LA INVESTIGACIÓN CIENTÍFICA EN CIENCIAS VETERINARIAS

Fundamentos, métodos, herramientas y aplicaciones para la generación de evidencia científica

Autores

Víctor Alejandro Bosquez Barcenes
abosquez@ueb.edu.ec
https://orcid.org/0000-0001-7679-6023
Universidad Estatal de Bolívar

Joscelito Bolívar Solano Gaibor
jsolano@ueb.edu.ec
https://orcid.org/0000-0002-5109-2713
Universidad Estatal de Bolívar

Danilo Fabian Yanes Silva
jyanez@ueb.edu.ec
https://orcid.org/0009-0007-0451-6337
Universidad Estatal de Bolívar

Jonnathan Vinicio Arias Real
joarias@ueb.edu.ec
https://orcid.org/0009-0008-2518-7365
Universidad Estatal de Bolívar

Editorial “ACACFESA SAS”

La presente obra fue revisada por 2 pares académicos externos ciegos conforme al proceso editorial de ACACFESA SAS.

Los rigurosos procedimientos editoriales de ACACFESA SAS garantizan la selección de manuscritos por sus aportes significativos al conocimiento y cualidades científicas.

Editorial: ACACFESA SAS.

Sello editorial:

Teléfono: (+593) 998955446

Web: https://acacfesa.com/editorial/index.php/1

ISBN digital: 978-9907-818-68-0

Doi: https://doi.org/10.70577/drxagd55/ACACFESA.EDITORIAL

Año: Octubre 2026

Aviso Legal

El contenido de esta obra, que incluye ilustraciones, textos, tablas, gráficos, cuadros y referencias bibliográficas, es responsabilidad única del autor o autores. Lo que se ha dicho, los criterios y los datos no reflejan necesariamente la posición institucional ni el pensamiento de la Editorial ACACFESA SAS.

Derechos de Autor ©

Este documento se publica conforme los términos y condiciones de la EDITORIAL ACACFESA SAS

Esta obra está bajo una licencia internacional



Creative Commons Atribución-NoComercial-CompartirIgual 4.0

La "Editorial ACACFESA SAS " y todos sus autores tienen la propiedad única de los derechos de autor y de propiedad intelectual e industrial relacionados con el contenido de esta publicación. La reproducción total o parcial de esta obra, su almacenamiento en sistemas informáticos, su tratamiento digital y cualquier tipo de distribución, transmisión o comunicación pública por medios electrónicos, ópticos, mecánicos, químicos, fotográficos o de grabación están prohibidos. Esto se encuentra bajo las sanciones que establece la legislación vigente a menos que se cuente con la autorización previa y por escrito de los titulares del copyright.

Queda excluido únicamente el uso con fines académicos o de investigación científica, siempre que no busque objetivos comerciales y se ejecute sin remuneración, debiendo mencionarse en todo momento a la fuente editorial correspondiente. Los autores son los únicos responsables de las opiniones expresadas en los diferentes capítulos, las cuales no necesariamente representan el punto de vista institucional de la editorial.

Constancia de Arbitraje

La Editorial ACACFESA SAS, certifica que este libro es el resultado de un estudio llevado a cabo por los autores, el cual fue evaluado por jurados expertos bajo el sistema de doble ciego, tanto en contenido como en forma. Asimismo, se llevó a cabo un análisis del método de investigación, paradigma y enfoque; a partir de la matriz epistémica adoptada por los autores, utilizando las normas APA, Séptima Edición y el proceso de antiplagio en línea turnitinedu para asegurar que la obra fuera científica.

DEDICATORIA

A mi hijo Miguel, porque tu sonrisa es el motor que impulsa cada uno de mis pasos y la razón más poderosa para no rendirme jamás. A mis padres, quienes con su amor, sacrificio y ejemplo me enseñaron que la educación es el camino más noble para transformar la vida. Gracias por creer en mí incluso cuando el camino parecía incierto, por cada desvelo compartido y por haberme dado las raíces firmes desde las cuales hoy puedo crecer. Este libro es, ante todo, un tributo a ustedes tres, porque sin su presencia nada de esto tendría sentido.

Víctor Alejandro

A mis hijas, Romina y Omara lo más preciado de mi vida: ustedes son el latido que me impulsa a imaginar y construir nuevos espacios para una convivencia más armónica y para la verdadera realización de los seres humanos.

Les dedico estas páginas sobre inteligencia artificial aplicada a las ciencias veterinarias con la esperanza de que la tecnología amplíe nuestras posibilidades sin apartarnos de lo esencial: la sensibilidad, la conciencia, la compasión y el respeto por toda forma de vida. Antes de cualquier inteligencia artificial, está el corazón humano; y mientras ese corazón sepa cuidar, comprender y amar, habrá futuro para la ciencia y para la humanidad.

Joscelito

A mis amadas hijas e hijo, por ser inspiración y fortaleza en cada paso de mi vida. A mis queridos estudiantes de Medicina Veterinaria, por permitirme compartir conocimientos y aprender junto a ustedes. Dedico este esfuerzo con profundo cariño, esperando que siempre prevalezcan en sus vidas la vocación, ética y perseverancia.

Danilo

Dedico esta obra, en primer lugar, a Dios, por ser la fuente de fortaleza, sabiduría y esperanza que me acompaña cada día. Gracias por concederme el don de la vida, la salud y la oportunidad de

ejercer una profesión que representa una parte fundamental de mi propósito personal. Agradezco también por poner en mi camino a personas que, de una u otra manera, han contribuido a mi crecimiento, dejando enseñanzas, experiencias y aportes que enriquecen tanto mi vida personal como mi trayectoria profesional. De manera muy especial, dedico esta obra a mis hijos, Emilio y Ezequiel, quienes constituyen mi mayor motivación y mi más grande inspiración. Ellos son el motor que me impulsa a seguir adelante, a superar cada desafío y a trabajar con perseverancia para construir un mejor futuro. Cada logro alcanzado tiene también el significado de un legado que deseo entregarles como ejemplo de esfuerzo, dedicación y constancia.

A mi padre, por ser una fuente permanente de inspiración y un ejemplo de vida, por sus enseñanzas y por mostrarme, con su ejemplo, el valor del trabajo, la responsabilidad y la perseverancia. Muchas de las metas que hoy alcanzo llevan consigo parte de aquello que aprendí de él.

Jonnathan

PRÓLOGO

La investigación en Ciencias Veterinarias atraviesa un periodo de profunda expansión metodológica. Historias clínicas electrónicas, imágenes diagnósticas, registros productivos, dispositivos portátiles, sensores ambientales, datos genómicos y sistemas de vigilancia epidemiológica generan volúmenes de información que difícilmente pueden estudiarse mediante procedimientos manuales. Este escenario ha favorecido la incorporación de herramientas de inteligencia artificial capaces de reconocer patrones, clasificar observaciones, estimar riesgos y apoyar el análisis de problemas complejos relacionados con la salud, la producción y el bienestar animal.

Sin embargo, disponer de algoritmos cada vez más sofisticados no garantiza investigaciones científicamente sólidas. Un modelo predictivo puede alcanzar métricas aparentemente sobresalientes y, aun así, carecer de utilidad clínica o biológica. También puede aprender sesgos de selección, aprovechar información disponible únicamente después del desenlace, confundir características del centro veterinario con señales de enfermedad o producir probabilidades mal calibradas. La capacidad computacional no compensa una pregunta imprecisa, una muestra inadecuada, la pseudorreplicación, la ausencia de controles o una definición deficiente de la variable de resultado.

Esta obra parte de una premisa esencial: la inteligencia artificial debe integrarse al método científico y permanecer subordinada a sus principios. No constituye un procedimiento alternativo para generar conocimiento ni reemplaza la metodología de la investigación, la bioestadística, el diseño experimental o el juicio profesional. Su aporte depende de la calidad de los datos, de la coherencia entre la pregunta y el diseño, de la selección razonada del método analítico y de la interpretación crítica de los resultados.

Inteligencia Artificial Aplicada a la Investigación Científica en Ciencias Veterinarias: Fundamentos, métodos, herramientas y aplicaciones para la generación de evidencia científica fue concebido para estudiantes universitarios, docentes, investigadores, médicos veterinarios y profesionales de las ciencias animales interesados en comprender estas tecnologías sin perder de vista la realidad clínica, productiva y epidemiológica en la que serán utilizadas. El libro no presenta

la inteligencia artificial como una colección de soluciones automáticas. La examina como un conjunto de métodos cuyo valor debe demostrarse frente a alternativas más sencillas, prácticas profesionales existentes y necesidades concretas de animales, responsables, productores y comunidades.

Los primeros capítulos establecen las bases conceptuales necesarias para diferenciar inteligencia artificial, aprendizaje automático, aprendizaje profundo, procesamiento del lenguaje natural e inteligencia artificial generativa. Esta distinción evita denominar inteligencia artificial a cualquier análisis computacional y permite reconocer qué tipo de problema puede resolverse mediante cada enfoque. También se analiza la relación entre descripción, inferencia, predicción y causalidad, conceptos que suelen confundirse cuando el rendimiento algorítmico se interpreta como explicación biológica.

La conexión con el método científico ocupa un lugar central. El texto recorre la delimitación del problema, la formulación de preguntas e hipótesis, la revisión de literatura, la planificación de la recolección de datos, la selección del diseño, el cálculo del tamaño muestral y la definición del análisis. De este modo, la inteligencia artificial se incorpora dentro de una secuencia de decisiones científicas y no como una intervención aislada al final del estudio.

El tratamiento del diseño experimental resulta especialmente pertinente para las Ciencias Veterinarias. En esta disciplina conviven animales, muestras biológicas, lesiones, imágenes, mediciones repetidas, corrales, hatos, explotaciones, hospitales y regiones. Identificar correctamente la unidad experimental y la unidad observacional es indispensable para evitar inferencias artificialmente precisas. La aleatorización, el bloqueo, el enmascaramiento, los controles, las réplicas independientes y el manejo de factores de confusión conservan toda su importancia cuando se emplean modelos de aprendizaje automático. Ningún algoritmo puede reconstruir la validez perdida por un experimento mal diseñado.

El libro dedica una atención cuidadosa a la gestión y preparación de los datos. Los registros veterinarios contienen valores faltantes, mediciones realizadas con diferentes equipos, errores de codificación, categorías poco frecuentes y observaciones dependientes. Las historias clínicas agregan abreviaturas, negaciones, expresiones locales y variaciones en la documentación

profesional. Las imágenes dependen de la técnica de adquisición, la posición del animal, el operador y el equipo utilizado. Por ello, la limpieza, la trazabilidad, la imputación, la normalización, el control de calidad y la prevención de fugas de información se presentan como componentes sustantivos de la investigación.

La exposición de los modelos mantiene una orientación comparativa. Regresión lineal y logística, árboles de decisión, random forest, gradient boosting, XGBoost, LightGBM, máquinas de vectores de soporte y métodos de vecinos cercanos son examinados según la naturaleza de la pregunta, el tamaño de la muestra, la estructura de los datos y la necesidad de interpretación. Esta organización evita la preferencia automática por algoritmos complejos y reconoce que una regresión correctamente especificada puede superar a métodos más flexibles cuando la muestra es limitada o la relación entre variables es relativamente simple.

Los capítulos sobre aprendizaje no supervisado, series temporales, visión computacional y procesamiento del lenguaje natural amplían el horizonte analítico hacia problemas frecuentes en la investigación veterinaria. La identificación de grupos clínicos, la reducción de dimensionalidad, la detección de anomalías en sensores, la predicción de variables fisiológicas, la clasificación de imágenes radiográficas y la extracción de información de historias clínicas se presentan con sus posibilidades y restricciones. En cada caso se insiste en documentar el estándar de referencia, separar adecuadamente los conjuntos de entrenamiento, validación y prueba, y evaluar el rendimiento en poblaciones diferentes de aquellas utilizadas durante el desarrollo.

La inteligencia artificial generativa recibe un tratamiento deliberadamente crítico. Los modelos de lenguaje pueden apoyar la exploración de conceptos, la formulación inicial de estrategias de búsqueda, la programación, la traducción y la revisión de la claridad de un texto. No obstante, producen respuestas a partir de regularidades estadísticas del lenguaje y no mediante una comprobación autónoma de la evidencia. Pueden inventar referencias, alterar datos, proponer análisis incompatibles con el diseño o redactar afirmaciones convincentes pero falsas. El libro establece, por tanto, una separación necesaria entre asistencia lingüística y autoridad científica. La responsabilidad sobre cada afirmación, referencia, resultado y decisión metodológica permanece en las personas autoras.

Las aplicaciones veterinarias abarcan medicina clínica, cirugía, anestesiología, evaluación del dolor, oncología, cardiología, neurología, dermatología, oftalmología, medicina regenerativa y rehabilitación. A estas áreas se suman la ganadería de precisión, la reproducción, la nutrición, la producción de leche, la avicultura, la porcicultura, la acuicultura y la evaluación automatizada del comportamiento y del bienestar animal. La epidemiología veterinaria, la vigilancia de enfermedades, la resistencia a los antimicrobianos, la inocuidad alimentaria y el enfoque One Health completan una visión que conecta al paciente individual con las poblaciones animales, los ecosistemas y la salud pública.

Los casos didácticos cumplen una función que supera la ejemplificación. En ellos se relacionan la pregunta científica, la población, la unidad experimental, las variables, el diseño, el método estadístico o algoritmo, la validación, las métricas y las limitaciones. Problemas como la predicción de mastitis bovina, la evaluación automatizada del dolor, la clasificación de radiografías caninas, la vigilancia de zoonosis y el análisis de historias clínicas permiten apreciar que el desarrollo de un modelo constituye solo una parte del proceso. Su utilidad depende de cómo se mide el desenlace, quién utilizará la predicción, en qué momento se aplicará y qué consecuencias producirán los errores.

Otro aporte relevante reside en la evaluación crítica de los modelos. Sensibilidad, especificidad, valores predictivos, área bajo la curva ROC, precisión, recall, puntuación F1, error absoluto y coeficiente de determinación se explican desde su significado clínico y biológico. La calibración, la selección de umbrales, el análisis de utilidad y la validación externa reciben la atención que merecen. Una métrica elevada no demuestra, por sí sola, que una herramienta sea segura, generalizable o beneficiosa. La evaluación debe considerar prevalencia, costos de error, incertidumbre, equidad entre subgrupos y comparación con la práctica habitual.

La ética atraviesa toda la obra. La información veterinaria puede incluir datos del responsable del animal, ubicación de explotaciones, registros económicos, resultados productivos y antecedentes clínicos. Su utilización plantea obligaciones de confidencialidad, seguridad, consentimiento y gobernanza. También deben considerarse el bienestar animal, la distribución de responsabilidades, la transparencia de los modelos y el riesgo de automatizar decisiones sin supervisión profesional suficiente.

El cierre del libro sitúa la reproducibilidad y la comunicación científica como condiciones inseparables de la calidad metodológica. Organizar datos, conservar código, registrar versiones, documentar transformaciones y utilizar guías de reporte permite que otras personas comprendan, evalúen y, cuando sea posible, reproduzcan los análisis. Los principios FAIR, la ciencia abierta y el control de versiones se analizan junto con sus límites éticos y legales. Abrir la ciencia no significa exponer información sensible; implica hacerla tan accesible como sea posible y tan protegida como resulte necesario.

Para América Latina, estos contenidos poseen una relevancia particular. La diversidad productiva, las diferencias en infraestructura, la limitada interoperabilidad de los sistemas y la escasez de conjuntos de datos representativos exigen modelos adaptados al contexto. Importar un algoritmo entrenado en otra población, especie, sistema productivo o red hospitalaria sin validación local puede generar conclusiones equivocadas. El desarrollo regional requiere fortalecer competencias en metodología, estadística, programación, gestión de datos y evaluación crítica, además de promover colaboraciones que respeten la soberanía y la diversidad de la información veterinaria.

Este libro invita a trabajar con inteligencia artificial desde una posición científica exigente. Su mensaje no consiste en rechazar la innovación ni en aceptarla por entusiasmo tecnológico. Propone formular mejores preguntas, diseñar estudios defendibles, utilizar métodos proporcionales al problema, interpretar los resultados con prudencia y evaluar si la herramienta mejora realmente la generación de evidencia o la atención de los animales.

La inteligencia artificial encontrará un espacio legítimo en las Ciencias Veterinarias cuando sus resultados sean verificables, reproducibles, clínicamente pertinentes y éticamente responsables. Alcanzar ese objetivo depende menos del prestigio del algoritmo que de la calidad del razonamiento científico que orienta su uso. Esa es la perspectiva que articula las páginas siguientes y la que convierte esta obra en una guía académica oportuna para investigar, enseñar y tomar decisiones en una disciplina cada vez más vinculada con los datos.

CONTENIDO

DEDICATORIA v

PRÓLOGO vii

INTRODUCCIÓN 1

UNIDAD 1 9

FUNDAMENTOS DE INTELIGENCIA ARTIFICIAL EN CIENCIAS VETERINARIAS 9

Introducción 10

1.1 Concepto y evolución de la inteligencia artificial 11

1.2 Inteligencia artificial ciencia de datos y aprendizaje automático 13

1.3 Machine learning deep learning e inteligencia artificial generativa 14

1.4 Aprendizaje supervisado no supervisado y por refuerzo 16

1.5 Diferencias entre bioestadística inferencia predicción y causalidad 18

1.6 Tipos de datos empleados en investigación veterinaria 20

1.7 Principales aplicaciones en medicina y producción animal 22

1.8 Alcances y limitaciones de la inteligencia artificial 24

1.9 Mitos frecuentes sobre el uso de algoritmos 26

1.10 Terminología fundamental 27

1.11 Caso didáctico predicción de mastitis subclínica en bovinos 28

1.12 Síntesis autoevaluación y actividad práctica 33

CAPÍTULO 2 36

LA INTELIGENCIA ARTIFICIAL DENTRO DEL MÉTODO CIENTÍFICO 36

Introducción 37

2.1. El problema de investigación veterinaria 38

2.2. Formulación de preguntas científicas e hipótesis 40

2.3. Objetivos de explicación, estimación, clasificación y predicción 42

Coherencia entre objetivo, datos y método 43

2.4. Uso de IA en la delimitación del problema 43

2.5. Búsqueda, selección y evaluación crítica de literatura 44

2.6. Construcción del marco teórico y estado del arte 46

2.7. Identificación de vacíos de conocimiento 47

Procedimiento para justificar un vacío 48

2.8. Elaboración de protocolos de investigación 49

2.9. IA en las fases del ciclo científico 50

2.10. Riesgos de automatizar decisiones metodológicas 52

2.11. Caso didáctico: formulación de un estudio sobre dolor posoperatorio canino 53

2.11.2. Pregunta PICOT 54

2.12. Síntesis, autoevaluación y actividad práctica 57

CAPÍTULO 3 61

ÉTICA INTEGRIDAD CIENTÍFICA Y GOBERNANZA DE DATOS VETERINARIOS 61

Introducción 62

3.1 Principios éticos aplicados a la investigación con animales 63

3.2 Bienestar animal y principio de las 3R 64

3.3 Consentimiento informado de tutores y propietarios 66

3.4 Privacidad y confidencialidad de historias clínicas 67

3.5 Propiedad custodia y reutilización de datos 69

3.6 Sesgos algorítmicos y desigualdad entre poblaciones animales 71

3.7 Transparencia y responsabilidad profesional 73

3.8 Autoría plagio e integridad científica 74

3.9 Declaración del uso de herramientas de IA 75

3.10 Regulación de sistemas de inteligencia artificial 76

3.11 Riesgos de la IA generativa en investigación 78

3.12 Caso didáctico uso secundario de expedientes clínicos veterinarios 79

3.13 Síntesis autoevaluación y actividad práctica 82

CAPÍTULO 4 88

DISEÑO DE INVESTIGACIONES VETERINARIAS ASISTIDAS POR INTELIGENCIA ARTIFICIAL 88

Introducción 89

4.1. Estudios observacionales y experimentales 90

4.2. Unidad experimental y unidad observacional 92

4.3. Tratamientos, controles y comparadores 94

4.4. Aleatorización, bloqueo y enmascaramiento 94

4.5. Réplicas biológicas y técnicas 96

4.6. Pseudorreplicación e independencia de observaciones 97

4.7. Diseños completamente aleatorizados y en bloques 98

4.8. Diseños factoriales y medidas repetidas 99

4.9. Estudios longitudinales y ensayos clínicos veterinarios 100

4.10. Muestreo y representatividad 102

4.11. Tamaño muestral para inferencia y predicción 103

4.12. Confusión, sesgo de selección y fuga de información 104

4.13. Uso responsable de IA en la planificación experimental 106

4.14. Caso didáctico: comparación de tratamientos para cicatrización posquirúrgica 107

4.15. Síntesis, autoevaluación y actividad práctica 112

CAPÍTULO 5 117

BIOESTADÍSTICA INFERENCIA Y CAUSALIDAD EN LA ERA DE LA INTELIGENCIA ARTIFICIAL 117

Introducción 118

Mapa metodológico del capítulo 119

5.1 Descripción y exploración de datos veterinarios 120

5.2 Distribuciones variabilidad e incertidumbre 122

5.3 Estimación e intervalos de confianza 123

5.4 Pruebas de hipótesis y valor p 124

5.5 Tamaño del efecto y relevancia clínica 125

5.6 Pruebas t y análisis de varianza 127

5.7 ANCOVA y control estadístico de covariables 128

5.8 Métodos no paramétricos 129

5.9 Regresión lineal y logística 130

5.10 Modelos lineales generalizados 130

5.11 Modelos mixtos y datos agrupados 132

5.12 Análisis de supervivencia Kaplan Meier y modelo de Cox 132

5.13 PCA PERMANOVA y análisis multivariado 133

5.14 Modelos aditivos generalizados 134

5.15 Predicción frente a explicación causal 135

5.16 Caso didáctico supervivencia en oncología veterinaria 137

5.17 Síntesis autoevaluación y actividad práctica 141

Rúbrica breve 145

CAPÍTULO 6 147

GESTIÓN CALIDAD Y PREPARACIÓN DE DATOS VETERINARIOS 147

6.1. Fuentes de datos en Ciencias Veterinarias 149

6.2. Historias clínicas electrónicas 150

6.3. Datos de laboratorio sensores imágenes y producción 152

6.4. Creación de diccionarios de datos 154

6.5. Estandarización y codificación de variables 155

6.6. Evaluación de calidad de los datos 156

6.7. Detección de duplicados errores y valores atípicos 157

6.8. Datos faltantes y mecanismos de ausencia 159

6.9. Imputación de datos 160

6.10. Transformación normalización y escalamiento 161

6.11. Codificación de variables categóricas 162

6.12. Ingeniería y selección de características 163

6.13. Datos desbalanceados 165

6.14. Interoperabilidad y estándares veterinarios 165

6.15. Trazabilidad metadatos y control de versiones 167

6.16. Caso didáctico depuración de una base de historias clínicas caninas 168

6.17. Síntesis autoevaluación y actividad práctica 173

CAPÍTULO 7 178

APRENDIZAJE AUTOMÁTICO SUPERVISADO EN INVESTIGACIÓN VETERINARIA 178

Introducción 179

Objetivos de aprendizaje 180

7.1 Formulación de un problema predictivo 180

7.2 Variables predictoras y variable objetivo 182

7.3 Regresión lineal y logística como modelos predictivos 183

7.4 Árboles de decisión 185

7.5 Random forest 186

7.6 Gradient boosting XGBoost y LightGBM 187

7.7 Máquinas de vectores de soporte 189

7.8 Métodos de vecinos cercanos 189

7.9 Selección del algoritmo según el problema científico 190

7.10 Entrenamiento validación y prueba 192

7.11 Validación cruzada 193

7.12 Ajuste de hiperparámetros 194

7.13 Sobreajuste y subajuste 196

7.14 Comparación justa entre modelos 197

7.15 Caso didáctico predicción de hospitalización en pacientes veterinarios 199

7.16 Síntesis autoevaluación y actividad práctica 204

CAPÍTULO 8 210

APRENDIZAJE NO SUPERVISADO, SERIES TEMPORALES Y DETECCIÓN DE PATRONES 210

Introducción 211

Objetivos de aprendizaje 211

8.1. Fundamentos del aprendizaje no supervisado 212

8.2. Análisis de componentes principales 213

8.3. Agrupamiento jerárquico 214

8.4. K-means y otros métodos de clustering 215

8.5. Reducción de dimensionalidad 216

8.6. Detección de anomalías 217

8.7. Minería de patrones en datos clínicos 218

8.8. Fundamentos de series temporales 219

8.9. Predicción de producción y variables fisiológicas 221

8.10. Datos longitudinales procedentes de sensores 221

8.11. Análisis espacial y espaciotemporal 222

8.12. Limitaciones de la interpretación de grupos algorítmicos 223

8.13. Caso didáctico: identificación de patrones conductuales en animales 224

8.14. Caso didáctico: predicción de producción lechera y estrés térmico 226

8.15. Síntesis, autoevaluación y actividad práctica 227

Código orientativo en R 231

CAPÍTULO 9 232

DEEP LEARNING, VISIÓN COMPUTACIONAL Y DIAGNÓSTICO POR IMÁGENES ... 232

Introducción 233

9.1. Fundamentos de las redes neuronales 234

9.2. Arquitectura y entrenamiento de modelos profundos 235

9.3. Redes neuronales convolucionales 236

9.4. Clasificación, detección y segmentación de imágenes 237

9.5. Radiografía, ecografía, tomografía y resonancia magnética 239

9.6. Imágenes histopatológicas y citológicas 241

9.7. Reconocimiento visual de lesiones y parásitos 241

9.8. Aumento de datos y transferencia de aprendizaje 242

9.9. Anotación de imágenes y concordancia entre evaluadores 243

9.10. Calidad de la imagen y variabilidad entre equipos 245

9.11. Interpretabilidad mediante mapas de activación 245

9.12. Sesgo de espectro y generalización entre centros 246

9.13. Caso didáctico: clasificación de radiografías torácicas caninas 248

9.14. Caso didáctico: segmentación de lesiones cutáneas 250

9.15. Síntesis, autoevaluación y actividad práctica 251

CAPÍTULO 10 256

PROCESAMIENTO DEL LENGUAJE NATURAL E INTELIGENCIA ARTIFICIAL GENERATIVA 256

Introducción 257

10.1. Fundamentos del procesamiento del lenguaje natural 258

10.2. Representación computacional del lenguaje científico 259

10.3. Transformers y modelos de lenguaje 260

10.4. Extracción de información de historias clínicas 260

10.5. Clasificación y resumen de textos veterinarios 261

10.6. ChatGPT y otros asistentes generativos 262

10.7. Búsqueda orientativa y revisión de literatura 263

10.8. Apoyo en programación y análisis estadístico 263

10.9. Construcción de prompts científicos 264

10.10. Verificación de afirmaciones y referencias 265

10.11. Alucinaciones, sesgos y errores metodológicos 266

10.12. Confidencialidad de la información clínica 266

10.13. Traducción y redacción científica asistidas 267

10.14. Criterios para declarar el uso de IA generativa 268

10.15. Caso didáctico: extracción de signos clínicos desde textos narrativos 269

10.16. Caso didáctico: evaluación crítica de una respuesta generada por IA 270

10.17. Síntesis, autoevaluación y actividad práctica 271

CAPÍTULO 11 274

INTELIGENCIA ARTIFICIAL EN MEDICINA CLÍNICA CIRUGÍA Y ESPECIALIDADES VETERINARIAS 274

Introducción 275

11.1 Sistemas de apoyo a la decisión clínica 276

11.2 Predicción diagnóstica y pronóstica 278

11.3 Medicina interna de perros y gatos 279

11.4 Cirugía y seguimiento posoperatorio 281

11.5 Anestesiología y monitorización 282

11.6 Evaluación automatizada del dolor 283

11.7 Oncología veterinaria 284

11.8 Cardiología y neurología 285

11.9 Dermatología y oftalmología 286

11.10 Medicina regenerativa y rehabilitación 287

11.11 Farmacología y medicina personalizada 288

11.12 Integración de datos clínicos multimodales 289

11.13 Caso didáctico predicción del riesgo anestésico 291

11.14 Caso didáctico evaluación del dolor posoperatorio mediante datos multimodales 294

11.15 Síntesis autoevaluación y actividad práctica 296

Conclusiones del capítulo 299

CAPÍTULO 12 300

GANADERÍA DE PRECISIÓN PRODUCCIÓN REPRODUCCIÓN Y BIENESTAR ANIMAL 300

Introducción 301

12.1 Fundamentos de la ganadería de precisión 302

12.2 Sensores dispositivos portátiles e internet de las cosas 304

12.3 Producción y calidad de leche 306

12.4 Predicción de mastitis y enfermedades metabólicas 307

12.5 Reproducción y detección automatizada del celo 308

12.6 Nutrición y alimentación de precisión 309

12.7 Crecimiento y eficiencia productiva 310

12.8 Avicultura y porcicultura inteligentes 311

12.9 Acuicultura de precisión 312

12.10 Evaluación automatizada del comportamiento 313

12.11 Detección de dolor estrés y alteraciones del bienestar 314

12.12 Visión computacional para identificación y seguimiento animal 315

12.13 Impacto ambiental y sostenibilidad 316

12.14 Caso didáctico alerta temprana de mastitis bovina 317

12.15 Caso didáctico clasificación del comportamiento en porcinos 320

12.16 Síntesis autoevaluación y actividad práctica 322

CAPÍTULO 13 326

EPIDEMIOLOGÍA SALUD PÚBLICA VETERINARIA Y ENFOQUE ONE HEALTH 326

Introducción 327

13.1 Inteligencia artificial en epidemiología veterinaria 328

13.2 Vigilancia y detección temprana de enfermedades 330

13.3 Modelos de riesgo y distribución espacial 332

13.4 Predicción de brotes 333

13.5 Zoonosis y enfermedades emergentes 334

13.6 Resistencia a los antimicrobianos 335

13.7 Inocuidad y trazabilidad de alimentos 337

13.8 Integración de datos animales humanos y ambientales 338

13.9 Datos climáticos movilidad y sistemas productivos 340

13.10 Modelos de redes de contacto 341

13.11 Genómica metagenómica y vigilancia molecular 341

13.12 Sesgos incertidumbre y comunicación del riesgo 342

13.13 Caso didáctico predicción espacial de una enfermedad zoonótica 344

13.14 Caso didáctico clasificación de perfiles de resistencia antimicrobiana 347

13.15 Síntesis autoevaluación y actividad práctica 349

CAPÍTULO 14 354

VALIDACIÓN INTERPRETACIÓN Y EVALUACIÓN CRÍTICA DE MODELOS 354

Introducción 355

14.1 Validación interna y externa 356

14.2 Validación temporal geográfica y entre especies 357

14.3 Sensibilidad especificidad y valores predictivos 358

14.4 Curvas ROC y área bajo la curva 359

14.5 Precision recall y puntuación F1 360

14.7 Calibración y utilidad clínica 362

14.8 Desbalance de clases y selección de umbrales 363

14.9 Importancia de variables 365

14.10 SHAP y otros métodos de explicabilidad 365

14.11 Robustez y análisis de sensibilidad 367

14.12 Comparación con el desempeño profesional 368

14.13 Validez clínica y biológica 369

14.14 Evaluación de impacto en condiciones reales 369

14.15 Monitoreo y deterioro del modelo 370

14.16 Caso didáctico validación externa de un modelo de riesgo clínico 372

14.17 Síntesis autoevaluación y actividad práctica 375

CAPÍTULO 15 379

REPRODUCIBILIDAD, COMUNICACIÓN CIENTÍFICA Y FUTURO DE LA IA VETERINARIA 379

Introducción 380

15.1. Investigación reproducible y ciencia abierta 381

15.2. Organización de proyectos de análisis 382

15.3. Uso de R y RStudio en flujos reproducibles 384

15.4. Scripts, cuadernos y reportes dinámicos 386

15.5. Control de versiones y documentación 387

15.6. Datos y código FAIR 388

15.7. Comunicación de métodos y resultados algorítmicos 389

15.8. Tablas, figuras y visualización científica 390

15.9. Redacción de artículos con modelos predictivos 391

15.10. Guías para reportar estudios de inteligencia artificial 391

15.11. Revisión crítica por pares 393

15.12. Transferencia del modelo a la práctica veterinaria 394

15.13. Costos, infraestructura y competencias profesionales 395

15.14. Modelos multimodales y medicina veterinaria de precisión 396

15.15. Gemelos digitales, robótica y sistemas autónomos 397

15.16. Agenda de investigación para América Latina 398

15.17. Formación del investigador veterinario ante la IA 399

15.18. Proyecto integrador: desarrollo de una investigación veterinaria asistida por IA 400

15.19. Síntesis y reflexión final 404

Preguntas de autoevaluación 405

ANEXOS 430

Glosario de términos 430

Anexo 2. Guía para seleccionar métodos estadísticos y algoritmos 440

Anexo 3. Banco de casos didácticos veterinarios 447

Anexo 4. Plantillas para protocolos de investigación 464

Anexo 5. Prompts científicos aplicados a Ciencias Veterinarias 471

Anexo 6. Respuestas orientativas a las preguntas de autoevaluación 480

INTRODUCCIÓN

La investigación científica en Ciencias Veterinarias se desarrolla en un entorno caracterizado por la diversidad de especies, la heterogeneidad de los sistemas productivos y la creciente disponibilidad de información clínica, biológica, epidemiológica y ambiental. Una consulta veterinaria puede generar antecedentes, resultados de laboratorio, imágenes diagnósticas, registros terapéuticos y notas clínicas. En una explotación pecuaria, los sensores pueden registrar producción, consumo, rumia, actividad, temperatura, desplazamiento y comportamiento durante semanas o meses. A escala poblacional, los sistemas de vigilancia integran datos sanitarios, climáticos, geográficos, genómicos y de movilidad. Esta multiplicación de fuentes plantea oportunidades científicas relevantes, pero también problemas metodológicos que no pueden resolverse mediante acumulación indiscriminada de datos.

La inteligencia artificial ofrece procedimientos para clasificar imágenes, extraer información de textos, reconocer patrones, detectar anomalías y estimar la probabilidad de desenlaces futuros. En medicina veterinaria puede contribuir a identificar animales con riesgo de deterioro, apoyar la lectura de estudios radiográficos, cuantificar lesiones histopatológicas, reconocer signos de dolor o anticipar eventos productivos. En epidemiología, permite explorar relaciones espaciales, integrar fuentes heterogéneas y construir sistemas de alerta. En producción animal, los dispositivos conectados posibilitan el seguimiento continuo de variables que antes se medían de forma ocasional.

Estas posibilidades han generado un interés creciente por incorporar aprendizaje automático, aprendizaje profundo, visión computacional, procesamiento del lenguaje natural e inteligencia artificial generativa en proyectos veterinarios. Sin embargo, el entusiasmo tecnológico suele acompañarse de simplificaciones conceptuales. Con frecuencia se confunde predicción con causalidad, se interpreta una elevada exactitud como prueba de utilidad clínica o se supone que un algoritmo complejo es necesariamente superior a un modelo estadístico convencional. También se emplean términos como inteligencia artificial, machine learning y deep learning como si fueran equivalentes, aunque describen niveles y métodos diferentes.

Este libro se construye alrededor de una idea rectora: la inteligencia artificial adquiere valor científico cuando responde a una pregunta veterinaria bien formulada, se apoya en datos de calidad, respeta el diseño de la investigación y produce resultados interpretables dentro de un contexto clínico, biológico, productivo o epidemiológico. El algoritmo no constituye el punto de partida. Antes de elegirlo deben definirse el problema, la población, la unidad de análisis, el desenlace, el propósito del estudio y las condiciones en las que se espera utilizar sus resultados.

Inteligencia artificial y método científico

La incorporación de inteligencia artificial no modifica las exigencias fundamentales de la investigación. Una pregunta imprecisa seguirá produciendo resultados ambiguos, aunque el análisis se realice con redes neuronales. Una muestra sesgada continuará limitando la generalización, incluso si el modelo presenta una elevada área bajo la curva. La pseudorreplicación no desaparece porque se utilice validación cruzada, y la fuga de información puede crear una impresión de rendimiento que se desvanece al aplicar el modelo en otra población.

En Ciencias Veterinarias, estas dificultades adquieren características particulares. Una base puede contener varias consultas del mismo animal, múltiples animales pertenecientes a un solo hato, varias imágenes obtenidas durante el mismo estudio o numerosas mediciones realizadas en una misma lesión. Si estas dependencias se ignoran al dividir los datos, las observaciones relacionadas pueden distribuirse entre entrenamiento y prueba. El algoritmo reconoce entonces características del individuo, del centro o del procedimiento de adquisición, en lugar de aprender una relación generalizable.

La definición de la unidad experimental merece la misma atención. En un ensayo de alimentación, el tratamiento puede aplicarse al corral, aunque el peso se mida en cada animal. En un estudio de mastitis, varias lactancias pueden proceder de una misma vaca y varias vacas compartir condiciones del hato. En investigación por imágenes, cada paciente puede aportar múltiples proyecciones. Considerar todas estas observaciones como independientes incrementa artificialmente el tamaño muestral y produce estimaciones demasiado optimistas.

Por estas razones, la obra relaciona cada herramienta computacional con las etapas del método científico: identificación del problema, revisión de literatura, formulación de preguntas e hipótesis,

elección del diseño, cálculo del tamaño muestral, recopilación de datos, control de calidad, análisis, validación, interpretación y comunicación. La inteligencia artificial no se aborda como un componente aislado, sino como parte de un proceso en el que las decisiones previas al modelado determinan buena parte de la validez posterior.

Bioestadística, predicción y causalidad

La bioestadística y el aprendizaje automático comparten procedimientos y fundamentos, pero no persiguen siempre los mismos objetivos. La inferencia busca estimar parámetros, contrastar hipótesis y cuantificar incertidumbre respecto de una población. La predicción pretende anticipar resultados en nuevas observaciones. La explicación estudia relaciones entre variables, mientras que la inferencia causal intenta estimar qué ocurriría bajo una intervención o exposición diferente. Estas finalidades pueden coexistir dentro de un proyecto, pero requieren diseños, supuestos y criterios de evaluación distintos.

Una asociación predictiva no demuestra que la variable utilizada por el modelo sea causa del desenlace. Si la administración de un medicamento aparece vinculada con mortalidad, el algoritmo puede estar capturando la mayor gravedad de los pacientes tratados. Del mismo modo, un modelo que predice mastitis a partir de disminuciones en la producción de leche no establece que esa disminución cause la enfermedad. Puede tratarse de una manifestación temprana del mismo proceso biológico.

El libro mantiene esta distinción durante el desarrollo de los métodos. Pruebas t, ANOVA, ANCOVA, modelos lineales generalizados, modelos mixtos, regresión logística, Kaplan-Meier, modelos de Cox, métodos no paramétricos, análisis de componentes principales, PERMANOVA y modelos aditivos generalizados conservan un lugar relevante en la investigación veterinaria. Random forest, XGBoost, LightGBM, máquinas de vectores de soporte, clustering, redes neuronales, CNN y transformers se incorporan cuando la naturaleza del problema y los datos justifican su utilización.

No existe mérito científico en emplear un algoritmo complejo cuando una técnica más sencilla responde adecuadamente a la pregunta. Por el contrario, la comparación con un modelo de referencia interpretable permite determinar si el aumento de complejidad aporta una mejora real.

La elección debe considerar rendimiento, estabilidad, calibración, tamaño de la muestra, costo computacional, necesidad de explicación y consecuencias de los errores.

Los datos como fundamento de la investigación

La calidad de los datos condiciona la calidad de cualquier modelo. Los registros veterinarios se generan inicialmente para atender pacientes, documentar procedimientos, gestionar explotaciones o cumplir funciones administrativas. Rara vez fueron diseñados para responder de forma directa a una pregunta de investigación. Por ello contienen valores faltantes, cambios de unidad, errores de digitación, códigos locales, mediciones dependientes del operador y variaciones en la disponibilidad de pruebas.

Los datos de hospitales de referencia pueden sobrerrepresentar enfermedades complejas, pacientes derivados y responsables con capacidad para asumir procedimientos especializados. Los registros de campo dependen de la infraestructura, la conectividad y las prácticas de documentación. Los sensores producen secuencias extensas, pero también pérdidas de señal, desplazamientos, fallos de calibración y observaciones imposibles. Las imágenes pueden revelar diferencias entre equipos, protocolos, proyecciones y centros antes que características de la lesión estudiada.

La preparación de datos no debe reducirse a una operación técnica. Cada transformación modifica la representación del fenómeno. Imputar, excluir valores, agrupar categorías, normalizar variables o seleccionar ventanas temporales implica supuestos que deben declararse y evaluarse. La depuración exige conservar los datos originales, documentar los cambios y evitar utilizar información del conjunto de prueba durante el preprocesamiento.

La separación entre entrenamiento, validación y prueba constituye una protección frente al optimismo. Cuando existen animales repetidos, hatos, centros o periodos relacionados, la partición debe respetar esa estructura. En determinados proyectos será necesario separar por animal; en otros, por explotación, hospital, región o periodo. Una división aleatoria simple puede ser insuficiente si permite que información estrechamente relacionada aparezca en ambos lados de la evaluación.

Validación y utilidad

El desempeño de un modelo no puede resumirse en una única cifra. Sensibilidad y especificidad describen comportamientos diferentes. Los valores predictivos dependen de la prevalencia. El área bajo la curva ROC evalúa discriminación, pero no informa si las probabilidades están bien calibradas ni si el uso del modelo mejora una decisión. Precision, recall y F1 pueden resultar útiles ante clases desbalanceadas, aunque tampoco sustituyen la interpretación clínica de los errores.

En regresión, el error absoluto, el error cuadrático y el coeficiente de determinación describen dimensiones distintas del rendimiento. Una predicción de producción lechera debe evaluarse considerando la magnitud de los errores y sus consecuencias para la gestión. En supervivencia, el modelo debe respetar el seguimiento y la censura. En imágenes, la clasificación, la detección y la segmentación requieren métricas acordes con la tarea y comparación con la variabilidad entre evaluadores.

La validación interna estima estabilidad dentro de los datos disponibles. La validación temporal examina el comportamiento en otro periodo. La validación geográfica o institucional analiza la transportabilidad hacia otros centros. La validación entre especies enfrenta diferencias aún mayores. Un modelo desarrollado con perros no debe aplicarse automáticamente a gatos; uno entrenado en bovinos de alta producción puede no conservar su rendimiento en sistemas extensivos.

La utilidad también depende del momento en que se genera la predicción y de la acción que desencadena. Una alerta de mastitis debe emitirse con anticipación suficiente para permitir una evaluación o intervención. Un sistema de apoyo anestésico debe considerar la disponibilidad de monitorización y personal. Un clasificador de radiografías debe indicar cuándo la imagen está fuera del dominio previsto o posee calidad insuficiente. El modelo más preciso no siempre es el más útil si produce alertas tardías, requiere variables inaccesibles o incrementa procedimientos innecesarios.

Alcance de la obra

Los quince capítulos siguen una progresión que comienza con los fundamentos y concluye con la transferencia responsable de los modelos a la práctica. Los primeros capítulos delimitan los conceptos de inteligencia artificial, aprendizaje automático y aprendizaje profundo, y los sitúan

dentro del método científico. Se desarrollan la formulación de preguntas, la revisión de literatura, la ética, la integridad científica, la protección de datos y la gobernanza.

El diseño de investigaciones veterinarias ocupa un espacio propio debido a su influencia sobre la validez de los resultados. Se revisan estudios observacionales, experimentos, ensayos clínicos, investigaciones longitudinales y proyectos predictivos. Se presta atención a tratamientos, controles, aleatorización, bloqueo, cegamiento, réplicas, mediciones repetidas, factores de confusión y pseudorreplicación.

Los capítulos dedicados a bioestadística, gestión de datos y preprocesamiento establecen la transición hacia el modelado. Se explican los datos estructurados y no estructurados, los registros clínicos electrónicos, la calidad de las mediciones, los valores faltantes, la codificación, la normalización, la selección de variables, el desbalance de clases y la prevención de fugas de información.

La sección metodológica desarrolla el aprendizaje supervisado y no supervisado, las series temporales, la detección de anomalías, el análisis espacial, el deep learning, la visión computacional y el procesamiento del lenguaje natural. La inteligencia artificial generativa se examina en relación con la búsqueda orientativa, la formulación de preguntas, la programación, la traducción y la redacción científica. Se analizan también sus riesgos: referencias inexistentes, afirmaciones falsas, errores estadísticos, sesgos y exposición de información confidencial.

Las aplicaciones incluyen medicina clínica, cirugía, anestesiología, dolor, oncología y diferentes especialidades veterinarias. En producción animal se desarrollan la ganadería de precisión, los sensores, la reproducción, la nutrición, la eficiencia productiva, la avicultura, la porcicultura y la acuicultura. El análisis del comportamiento, el bienestar y el impacto ambiental amplía la perspectiva más allá del rendimiento económico.

La epidemiología y el enfoque One Health integran información animal, humana y ambiental. La vigilancia de enfermedades, la predicción de brotes, las zoonosis, la resistencia a los antimicrobianos, la inocuidad alimentaria, la movilidad y las redes de contacto muestran cómo la inteligencia artificial puede utilizarse a escala poblacional. A la vez, se examinan los sesgos de

vigilancia, la desigualdad en la disponibilidad de datos y la dificultad de comunicar probabilidades e incertidumbre.

Los capítulos finales se concentran en validación, interpretación, reproducibilidad y comunicación científica. Se revisan calibración, utilidad clínica, selección de umbrales, explicabilidad, robustez, comparación con profesionales y evaluación de impacto. El cierre incorpora organización de proyectos, uso de R y RStudio, scripts, reportes dinámicos, control de versiones, principios FAIR, ciencia abierta y guías para comunicar investigaciones que emplean inteligencia artificial.

Orientación didáctica

La obra combina fundamentos conceptuales, explicaciones metodológicas, ejemplos veterinarios, casos simulados, actividades prácticas y preguntas de autoevaluación. Los casos didácticos se identifican expresamente cuando utilizan datos ficticios. Su propósito es mostrar cómo se articula una investigación completa y no presentar resultados como evidencia real.

En los ejemplos se describen, según corresponda, el problema científico, la pregunta, la hipótesis, la población, la unidad experimental, las variables, el diseño, el análisis, la validación y las limitaciones. Se abordan situaciones relacionadas con mastitis bovina, dolor posoperatorio, imágenes radiográficas, riesgo anestésico, enfermedades cardiacas, producción de leche, comportamiento animal, registros clínicos y vigilancia epidemiológica.

El código y las herramientas informáticas se presentan como medios para ejecutar y documentar procedimientos. No sustituyen la comprensión de los supuestos. Una instrucción ejecutada sin criterio puede producir resultados numéricamente correctos pero científicamente inadecuados. Por ello, cada procedimiento debe interpretarse desde la pregunta veterinaria y revisarse antes de incorporarlo a una conclusión.

Las plantillas, guías y prompts científicos incluidos al final del libro buscan apoyar el trabajo académico sin convertir la inteligencia artificial generativa en fuente de evidencia. Un prompt bien formulado puede ayudar a organizar una tarea, señalar restricciones y solicitar comprobaciones, pero la salida requiere revisión humana. Las referencias deben localizarse en fuentes verificables, el código debe probarse y los resultados estadísticos tienen que reproducirse a partir de los datos.

Una perspectiva crítica y aplicada

La inteligencia artificial puede ampliar la capacidad para estudiar problemas veterinarios, pero también puede producir nuevas formas de error. Los modelos pueden excluir poblaciones poco representadas, reproducir decisiones históricas, fallar ante cambios de equipos o confundir correlaciones institucionales con señales biológicas. El riesgo aumenta cuando la herramienta se presenta como objetiva, aunque sus datos reflejen prácticas incompletas o desiguales.

La supervisión humana no consiste únicamente en observar la salida de un sistema. Requiere capacidad real para cuestionarla, corregirla y detener su uso. El profesional necesita conocer la población de desarrollo, las variables empleadas, el rendimiento esperado, las limitaciones y las situaciones en las que el modelo debe abstenerse. La responsabilidad no puede desplazarse hacia una herramienta que carece de juicio clínico, obligaciones profesionales y comprensión del bienestar animal.

Este libro propone una aproximación equilibrada. La inteligencia artificial no debe rechazarse por su complejidad ni aceptarse por su novedad. Debe evaluarse mediante los mismos principios que sostienen la investigación científica: claridad de la pregunta, coherencia del diseño, calidad de la medición, transparencia del análisis, cuantificación de la incertidumbre, posibilidad de verificación y relevancia de los resultados.

El propósito final no es formar usuarios de algoritmos, sino investigadores y profesionales capaces de decidir cuándo estas herramientas aportan valor, cómo deben validarse y qué límites no pueden ignorarse. La generación de evidencia científica en Ciencias Veterinarias exige integrar conocimiento disciplinar, razonamiento metodológico, competencia estadística, responsabilidad ética y comprensión computacional. La inteligencia artificial encuentra su lugar dentro de esa integración, no fuera de ella.

UNIDAD 1

FUNDAMENTOS DE INTELIGENCIA ARTIFICIAL EN CIENCIAS VETERINARIAS

Introducción

La inteligencia artificial ha entrado en la investigación veterinaria por vías muy distintas: sistemas de ordeño que producen registros por vaca y por minuto, algoritmos que reconocen patrones en radiografías, modelos que anticipan brotes, sensores que describen conducta y herramientas generativas capaces de producir texto o código. Bajo una misma etiqueta conviven técnicas con propósitos, exigencias y riesgos diferentes. Un clasificador de mastitis, un modelo causal para evaluar una intervención y un asistente de redacción no responden la misma pregunta ni generan la misma clase de evidencia.

El punto de partida científico consiste en tratar la inteligencia artificial como un conjunto de métodos y sistemas, no como una autoridad autónoma. Su utilidad depende de la pregunta, del diseño del estudio, de la calidad de los datos y de la evaluación en la población donde se pretende usar. Los algoritmos flexibles pueden descubrir relaciones no lineales, integrar grandes volúmenes de información y automatizar tareas repetitivas. También pueden aprender errores de medición, prácticas clínicas históricas o sesgos de selección. Ningún incremento de capacidad computacional compensa una unidad experimental mal definida, una variable de desenlace ambigua o una validación contaminada por información del futuro.

Este capítulo construye un vocabulario común para estudiar la IA en medicina veterinaria, producción animal y salud pública. Diferencia descripción, inferencia, predicción y causalidad; presenta las familias principales de aprendizaje; examina datos clínicos, productivos, ómicos, espaciales, textuales y de imagen; y propone criterios para juzgar alcances y límites. El caso final convierte estos fundamentos en un protocolo reproducible para predecir mastitis subclínica sin confundir una alerta algorítmica con un diagnóstico etiológico.

Objetivos de aprendizaje

- Explicar el concepto de inteligencia artificial y reconocer sus principales etapas de desarrollo.
- Distinguir inteligencia artificial, ciencia de datos, aprendizaje automático, aprendizaje profundo e inteligencia artificial generativa.

- Seleccionar el paradigma de aprendizaje adecuado según la pregunta veterinaria, el desenlace y la disponibilidad de etiquetas.
- Separar objetivos descriptivos, inferenciales, predictivos y causales dentro de un protocolo de investigación.
- Clasificar los datos veterinarios según estructura, escala, temporalidad, nivel jerárquico y modalidad.
- Evaluar críticamente desempeño, calibración, transportabilidad, equidad, bienestar animal, privacidad y supervisión profesional.
- Diseñar un flujo metodológico para desarrollar y validar un modelo de riesgo de mastitis subclínica.

1.1 Concepto y evolución de la inteligencia artificial

La inteligencia artificial puede definirse como el campo que estudia y construye sistemas capaces de recibir información del entorno, procesarla y producir resultados orientados a objetivos, tales como clasificar, predecir, recomendar, generar contenido o apoyar decisiones. La definición debe centrarse en funciones observables. Atribuir comprensión, intención o juicio clínico a un programa confunde el comportamiento del sistema con capacidades humanas que no han sido demostradas.

La historia técnica suele organizarse en varias corrientes. La IA simbólica representó conocimiento mediante reglas explícitas y sistemas expertos. En veterinaria, una formulación temprana habría relacionado signos, resultados de laboratorio y diagnósticos mediante reglas del tipo “si-entonces”. Su ventaja era la trazabilidad; su debilidad, la dificultad para abarcar excepciones, incertidumbre y variabilidad biológica. Con el aumento de bases de datos y potencia computacional, el aprendizaje automático desplazó parte del trabajo desde la programación de reglas hacia la estimación de patrones a partir de ejemplos.

El aprendizaje profundo amplió esta tendencia mediante redes neuronales con múltiples capas, especialmente eficaces para imágenes, audio, señales y texto. La radiología veterinaria ofrece un ejemplo directo: una red convolucional puede aprender representaciones de una imagen sin que el investigador programe de manera exhaustiva bordes, texturas o formas. Sin embargo, esa flexibilidad aumenta las exigencias de datos, cómputo, control de sesgo y explicación. La revisión

de Hennessey et al. (2022) muestra que el campo de la imagen diagnóstica veterinaria está creciendo, aunque la heterogeneidad de estudios y la escasez de validación externa limitan la traslación clínica.

Desde 2022 se generalizó el acceso a modelos fundacionales y sistemas generativos. Estos modelos se entrenan con grandes corpus y pueden producir texto, imágenes, audio, código o estructuras moleculares. En investigación veterinaria pueden ayudar a bosquejar una estrategia de búsqueda, revisar la lógica de un script o convertir una explicación técnica en lenguaje para tutores. No deben citarse como fuente científica ni utilizarse para inventar datos faltantes. La respuesta fluida de un modelo de lenguaje no demuestra veracidad, y la información clínica introducida en servicios externos puede quedar fuera del control institucional.

La evolución no debe interpretarse como una secuencia donde cada técnica nueva elimina a la anterior. Un sistema de reglas puede ser preferible cuando existe normativa clara; una regresión logística puede ofrecer mejor calibración e interpretación que una red neuronal; y un modelo profundo puede ser razonable ante miles de imágenes etiquetadas. Russell y Norvig (2021) describen la IA como un campo amplio en el que conviven representación, razonamiento, aprendizaje, percepción y acción. La elección veterinaria debe responder a la tarea y a la evidencia disponible.

Tabla 1.1

Etapas técnicas y ejemplos veterinarios

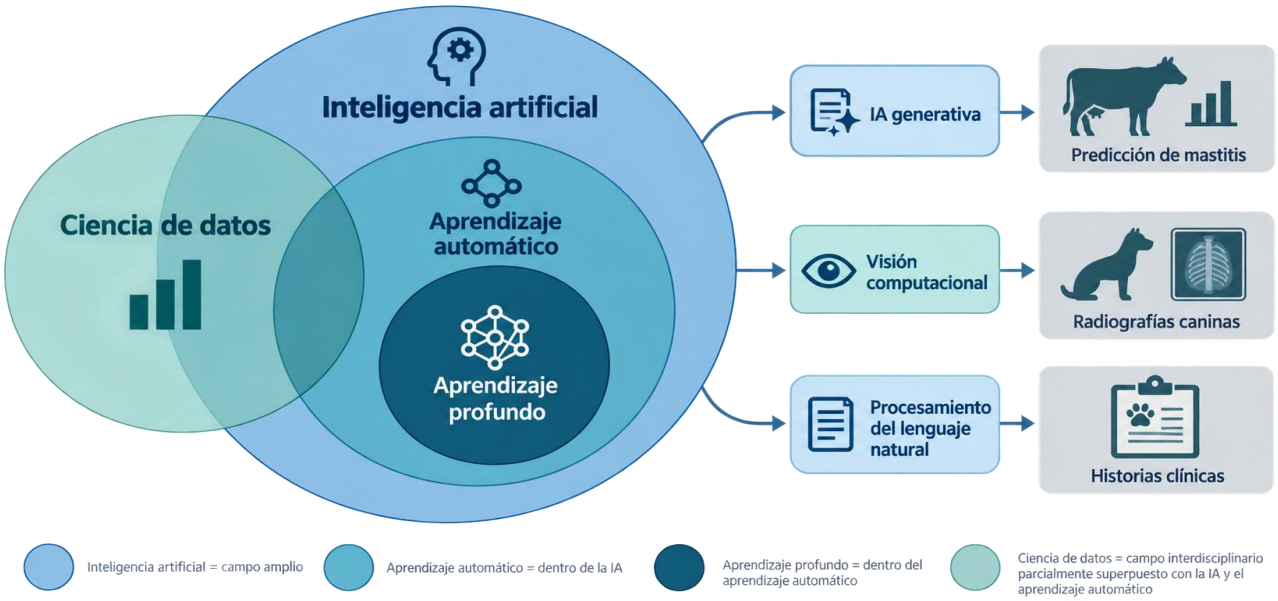
Etapa o enfoque	Idea central	Ejemplo veterinario	Límite principal
Sistemas simbólicos	Reglas explícitas construidas por expertos	Árbol de triaje con signos de urgencia	Rigidez ante casos atípicos
Aprendizaje estadístico	Estimación de relaciones desde datos	Regresión logística para riesgo de hospitalización	Depende de especificación y medición
Machine learning	Optimización predictiva con algoritmos flexibles	Random forest para riesgo de mastitis	Puede sobreajustar y ser difícil de transportar
Deep learning	Aprendizaje jerárquico de representaciones	CNN para clasificar radiografías caninas	Requiere muchos datos y validación rigurosa

IA generativa	Producción de contenido nuevo condicionado por instrucciones	Borrador de código R o resumen para revisión humana	Alucinaciones confidencialidad y atribución

1.2 Inteligencia artificial ciencia de datos y aprendizaje automático

Figura 1.1.

Relación entre la inteligencia artificial ciencia de datos y aprendizaje automático



Inteligencia artificial es la categoría más amplia. Ciencia de datos es una práctica interdisciplinaria que organiza el ciclo completo de obtención, gestión, análisis, comunicación y uso de datos. El aprendizaje automático o machine learning es un conjunto de métodos que ajustan su comportamiento a partir de datos para mejorar una tarea definida. Las tres áreas se superponen, pero ninguna es sinónimo de las otras.

Un proyecto sobre mortalidad neonatal porcina ilustra la diferencia. La ciencia de datos incluye acuerdos de acceso, diccionario de variables, integración de partos y ambiente, auditoría de faltantes, análisis exploratorio, visualización y documentación. El machine learning puede aportar un modelo de clasificación que estime riesgo por camada. La IA aparece cuando ese modelo se integra en un sistema que prioriza inspecciones o recomienda una acción. Si el estudio termina en una tabla descriptiva bien construida, sigue siendo ciencia de datos aunque no utilice IA.

La separación evita dos errores. El primero consiste en llamar IA a cualquier análisis computacional, incluso a un promedio o una prueba t. El segundo supone que el algoritmo constituye el proyecto completo. En realidad, gran parte del trabajo científico ocurre antes y después del ajuste: formular la pregunta, definir la población objetivo, identificar la unidad de análisis, medir el desenlace, prevenir filtración de información, comparar con una referencia sencilla y estudiar consecuencias de los errores.

Ezanno et al. (2021) señalan que la investigación en salud animal puede beneficiarse de la integración de epidemiología, modelización, datos heterogéneos y conocimiento del dominio. Esa integración exige colaboración real entre especialistas. El profesional veterinario aporta fisiopatología, contexto productivo y significado clínico; el bioestadístico define estimandos e incertidumbre; el especialista en datos implementa procesos reproducibles; y quienes trabajan con los animales valoran factibilidad y bienestar.

Tabla 1.2

Diferencias entre campos relacionados

Campo	Pregunta típica	Producto	Criterio de calidad
Bioestadística	¿Cuál es la magnitud y la incertidumbre de una asociación o efecto?	Estimación intervalo prueba o modelo	Validez inferencial y supuestos
Ciencia de datos	¿Cómo convertir datos dispersos en información utilizable?	Conjunto depurado flujo reproducible visualización	Trazabilidad calidad y utilidad
Machine learning	¿Qué patrón permite predecir casos nuevos?	Función predictiva y probabilidades	Desempeño fuera de muestra y calibración
Inteligencia artificial	¿Cómo ejecutar o apoyar una tarea orientada a objetivos?	Sistema que percibe predice genera o recomienda	Seguridad utilidad supervisión y gobernanza

1.3 Machine learning deep learning e inteligencia artificial generativa

Machine learning

Machine learning reúne algoritmos que estiman una función a partir de observaciones. En un problema supervisado, el modelo aprende una relación aproximada entre predictores X y un

desenlace Y. Para clasificación, Y puede indicar mastitis sí o no; para regresión, puede ser producción diaria de leche; para supervivencia, tiempo hasta recaída de un mastocitoma. Regresión penalizada, árboles, random forest, máquinas de soporte vectorial y boosting son alternativas con distintas capacidades y supuestos.

La complejidad debe justificarse empíricamente. Si edad, condición corporal y creatinina explican de forma aproximadamente lineal un riesgo renal, una regresión puede ser suficiente y más fácil de calibrar. XGBoost gana interés cuando existen interacciones y no linealidades relevantes, pero exige ajuste de hiperparámetros y controles estrictos contra sobreajuste. La comparación debe incluir un modelo de referencia y realizarse con los mismos datos de evaluación.

Deep learning

Deep learning utiliza redes neuronales profundas para aprender representaciones sucesivas. Las redes convolucionales son frecuentes en imágenes; las recurrentes y los transformers pueden procesar secuencias; las arquitecturas multimodales combinan texto, imagen y señales. Bouhali et al. (2022) describen el potencial de la radiómica y la IA para extraer rasgos cuantitativos de imágenes veterinarias, pero advierten retos de estandarización, tamaño muestral y reproducibilidad.

En una investigación radiográfica canina, dividir imágenes al azar puede producir una evaluación engañosa si distintas proyecciones del mismo perro quedan en entrenamiento y prueba. La partición correcta se realiza por paciente y, cuando se desea probar transportabilidad, por hospital o periodo. El modelo no debe aprender marcas del equipo, etiquetas incrustadas o protocolos propios de un centro como atajos para la clase diagnóstica.

Inteligencia artificial generativa

La IA generativa aprende una distribución de datos y produce nuevas salidas condicionadas por una instrucción. En investigación puede apoyar la exploración lingüística, la traducción preliminar, la explicación de código y la elaboración de listas de comprobación. La responsabilidad científica permanece en los autores. Cada afirmación, referencia, número y decisión analítica debe verificarse en la fuente original o en los datos.

Un uso razonable sería pedir a un modelo que proponga términos MeSH y palabras libres para una revisión sobre osteoartritis canina, revisar después cada término y ejecutar la búsqueda en bases bibliográficas. Un uso inaceptable sería solicitar “diez referencias recientes” y copiarlas sin comprobar DOI, título, autores y correspondencia con la afirmación. UNESCO (2021) sitúa transparencia, responsabilidad, privacidad, no discriminación y supervisión humana entre los principios centrales de una gobernanza ética de la IA.

Tabla 1.3

Comparación entre familias de métodos

Familia	Datos frecuentes	Fortaleza	Riesgo metodológico	Ejemplo
ML convencional	Tablas con variables clínicas y productivas	Buen desempeño con muestras moderadas	Fugas de datos y ajuste excesivo	Riesgo de mastitis
Deep learning	Imágenes audio texto y series extensas	Aprende representaciones complejas	Opacidad y demanda de datos	Detección radiográfica
IA generativa	Corpus masivos y datos multimodales	Genera borradores y contenido	Alucinación sesgo y divulgación de datos	Apoyo a código y comunicación

1.4 Aprendizaje supervisado no supervisado y por refuerzo

Aprendizaje supervisado

El aprendizaje supervisado utiliza ejemplos con desenlace conocido. La etiqueta debe corresponder a la pregunta y haberse medido con un procedimiento defendible. Para predecir mastitis subclínica, una etiqueta basada solo en una alarma del mismo sensor utilizado como predictor puede circularmente premiar al modelo. Un estándar de referencia puede combinar recuento de células somáticas, cultivo y criterios definidos por protocolo, reconociendo que cada opción representa un constructo distinto.

Clasificación y regresión son las tareas básicas. En clasificación no basta informar exactitud. Si la prevalencia es 10 %, predecir siempre “sin mastitis” logra 90 % de exactitud y no detecta ningún caso. Sensibilidad, especificidad, valores predictivos, curvas ROC y precisión-recall deben acompañarse de calibración y umbrales clínicamente justificados. En regresión se valoran el error

absoluto, el error cuadrático, la distribución de residuos y la magnitud del error en unidades biológicas.

Aprendizaje no supervisado

El aprendizaje no supervisado busca estructura sin una etiqueta de desenlace. Clustering puede agrupar patrones de actividad en perros con dolor crónico; PCA puede resumir variables correlacionadas de microbiota o producción; métodos de detección de anomalías pueden identificar cambios inusuales en consumo de agua. Los grupos obtenidos no son diagnósticos naturales por defecto. Su estabilidad debe evaluarse y su interpretación requiere variables externas, conocimiento biológico y análisis de sensibilidad.

En producción animal, un clúster denominado “bajo bienestar” sería una interpretación, no una propiedad descubierta automáticamente. Antes de usar esa etiqueta habría que demostrar coherencia con lesiones, conducta, ambiente u otros indicadores. Cambiar la escala de las variables o el número de grupos puede modificar la solución.

Aprendizaje por refuerzo

El aprendizaje por refuerzo estudia agentes que aprenden una política de acción mediante recompensas acumuladas. Sus aplicaciones veterinarias directas son menos maduras que las del aprendizaje supervisado. Puede investigarse en control ambiental de galpones, asignación dinámica de alimento o robótica, siempre que la función de recompensa incorpore salud y bienestar y no solo rendimiento productivo.

Experimentar mediante ensayo y error sobre animales plantea límites evidentes. Por ello, el desarrollo suele apoyarse en simuladores o datos históricos, seguido de pruebas escalonadas y supervisadas. Una recompensa mal definida puede inducir conductas técnicamente óptimas pero biológicamente indeseables. Maximizar conversión alimenticia sin restricciones de salud es un ejemplo de objetivo incompleto.

Tabla 1.4

Selección del paradigma de aprendizaje

Situación de investigación	Paradigma	Unidad de análisis	Evaluación mínima
Predecir si una vaca superará un umbral de RCS	Supervisado clasificación	Vaca en una fecha índice	Sensibilidad especificidad calibración y validación por hato
Identificar perfiles de actividad en perros geriátricos	No supervisado clustering	Perro o ventana temporal	Estabilidad separación e interpretación clínica
Regular ventilación de un galpón con acciones secuenciales	Por refuerzo	Estado ambiental y decisión	Seguridad recompensa acumulada y pruebas simuladas
Generar un resumen de una historia clínica	Generativo	Documento o episodio	Fidelidad omisiones privacidad y revisión veterinaria

1.5 Diferencias entre bioestadística inferencia predicción y causalidad

Figura 1.2.

Diferencias entre bioestadística inferencia predicción y causalidad



Bioestadística es la disciplina que aplica razonamiento estadístico a fenómenos biológicos, clínicos y sanitarios. Incluye diseño, muestreo, estimación, contraste, modelización y cuantificación de incertidumbre. Machine learning comparte herramientas con la estadística, pero suele concentrarse en desempeño predictivo fuera de muestra. La diferencia más útil no se encuentra en el nombre del algoritmo, sino en la pregunta y en el criterio con que se juzga la respuesta.

Descripción

La descripción resume lo observado: prevalencia de mastitis en el hato, mediana de días en leche, distribución de razas o tasa de mortalidad. No generaliza automáticamente a otra población ni explica causas. Intervalos de confianza pueden expresar precisión de una estimación descriptiva cuando existe un diseño de muestreo apropiado.

Inferencia

La inferencia utiliza una muestra para estimar parámetros o contrastar hipótesis sobre una población o proceso. Un ensayo aleatorizado puede estimar la diferencia media de días de cicatrización entre dos apósitos. El valor p no mide la magnitud clínica, la probabilidad de la hipótesis ni la capacidad predictiva. Deben informarse tamaño del efecto, intervalo de confianza, supuestos y relevancia biológica.

Predicción

La predicción estima un resultado desconocido en un individuo o unidad nueva. Un modelo puede predecir el riesgo de mastitis en las próximas 48 horas sin identificar la causa de cada evento. Su evaluación exige separar desarrollo y evaluación, evitar filtración, medir discriminación y calibración, y probar el desempeño en condiciones distintas. TRIPOD+AI recomienda informar de forma transparente la procedencia de datos, participantes, predictores, desenlace, manejo de faltantes, evaluación y disponibilidad del modelo (Collins et al., 2024).

Causalidad

La causalidad pregunta qué ocurriría con un desenlace si se interviniera sobre una exposición o tratamiento. Una asociación entre menor rumia y mastitis no demuestra que aumentar artificialmente la rumia prevenga la enfermedad. La rumia puede actuar como marcador temprano de enfermedad o compartir causas con el desenlace. Para estimar efectos causales se necesita un diseño que identifique la comparación contrafactual: aleatorización cuando es viable o, en estudios observacionales, supuestos explícitos sobre confusión, selección y medición.

Predicción y causalidad pueden complementarse. Un modelo de riesgo identifica vacas que requieren evaluación; un ensayo posterior determina si una intervención guiada por la alerta reduce enfermedad, uso de antimicrobianos o pérdidas. La primera pregunta es "¿quién desarrollará el evento?"; la segunda, "¿qué efecto tiene actuar?". El éxito en una no garantiza el éxito en la otra.

Tabla 1.5

Cuatro propósitos analíticos

Propósito	Pregunta	Resultado principal	Error frecuente
Descriptivo	¿Qué se observó?	Frecuencias medias tasas y distribución	Generalizar sin diseño muestral
Inferencial	¿Qué magnitud tiene la asociación o diferencia?	Estimación efecto IC y prueba	Equiparar p menor a 0.05 con importancia clínica
Predictivo	¿Qué ocurrirá en una unidad nueva?	Probabilidad clase o valor futuro	Evaluar con datos vistos durante el ajuste
Causal	¿Qué ocurriría bajo una intervención?	Efecto causal bajo supuestos	Interpretar asociación como causa

1.6 Tipos de datos empleados en investigación veterinaria

Los datos veterinarios se diferencian por modalidad, escala de medición, estructura temporal y nivel jerárquico. Estas propiedades determinan el modelo y la validación. Una fotografía de una lesión, una secuencia genómica, un registro de ordeño y una nota clínica son observaciones de naturaleza distinta. Convertirlas en columnas de una matriz no elimina sus dependencias.

Datos clínicos y de laboratorio

Incluyen anamnesis, signos, diagnósticos, tratamientos, hemogramas, bioquímica, cultivos, histopatología y desenlaces. Las historias clínicas suelen contener faltantes informativos: una prueba no solicitada puede indicar que el clínico consideró improbable una enfermedad o que el tutor no pudo asumir el costo. Imputar sin estudiar el mecanismo de ausencia puede alterar la predicción.

Datos productivos y de sensores

Sistemas de ordeño, collares, podómetros, cámaras, balanzas y estaciones ambientales generan series temporales de alta frecuencia. Las observaciones están anidadas: mediciones dentro de animal, animales dentro de corral o hato, y hatos dentro de región. Ignorar esta estructura produce intervalos demasiado estrechos y particiones optimistas. Para evaluar un modelo destinado a nuevas fincas, el conjunto de prueba debe contener fincas completas no utilizadas en entrenamiento.

Imágenes señales y audio

Radiografías, ecografías, tomografías, resonancias, imágenes térmicas, grabaciones de tos y señales electrofisiológicas requieren control de adquisición, metadatos y etiquetado. La imagen puede revelar información accidental sobre centro, equipo o protocolo. La anotación por un solo observador incorpora error y variabilidad; los estudios deben documentar experiencia, consenso y concordancia.

Datos ómicos microbiológicos y genéticos

Genómica, transcriptómica, proteómica, metabolómica y microbioma se caracterizan por alta dimensionalidad. A menudo hay más variables que unidades biológicas. Selección de características, regularización y reducción dimensional deben ocurrir dentro de cada ciclo de validación para evitar fuga de información. La independencia se define por la unidad biológica, no por el número de secuencias o genes.

Texto espacio y redes

Las notas clínicas y reportes diagnósticos pueden analizarse mediante procesamiento de lenguaje natural, con especial cuidado de anonimización y ambigüedad terminológica. Los datos espaciales de brotes incorporan autocorrelación y cambios en vigilancia. Los datos de redes de movimientos animales permiten estudiar conectividad y propagación, pero la ausencia de un movimiento registrado no equivale siempre a ausencia de contacto.

Tabla 1.6

Modalidades de datos y controles de calidad

Modalidad	Ejemplos	Dependencia relevante	Control prioritario
Tabular clínica	Edad peso laboratorio diagnóstico	Episodios repetidos por paciente	Diccionario unidades faltantes y temporalidad
Serie temporal	Rumia actividad temperatura producción	Ventanas dentro de animal	Sincronización deriva del sensor y corte temporal
Imagen	Radiografía ecografía histología	Varias imágenes por paciente	Partición por animal control de equipo y anotación
Ómica	SNP expresión microbioma	Genes dentro de muestra y lotes	Efecto de lote normalización y dimensionalidad
Texto	Historia clínica informe patológico	Notas por episodio y profesional	Anonimización negación abreviaturas y contexto
Geoespacial y red	Brote movimiento de animales granjas	Vecindad y conectividad	Resolución espacial sesgo de vigilancia y tiempo

1.7 Principales aplicaciones en medicina y producción animal

Las aplicaciones pueden agruparse por tarea científica. En diagnóstico asistido, los modelos clasifican imágenes o combinan señales clínicas. En pronóstico, estiman recaída, mortalidad o respuesta. En vigilancia, detectan anomalías temporales y espaciales. En producción de precisión, anticipan enfermedad, cambios reproductivos o alteraciones de bienestar. Akinsulie et al. (2024) revisan usos en práctica clínica y en investigación biomédica, incluidos diagnóstico por imágenes, epidemiología, resistencia antimicrobiana, oncología, desarrollo de fármacos y genómica.

Medicina de animales de compañía

En perros y gatos, la IA se investiga para interpretar imágenes, priorizar diagnósticos diferenciales, reconocer patrones de laboratorio y anticipar deterioro. Reagan et al. (2022) mostraron la posibilidad de usar algoritmos de machine learning para apoyar la detección temprana de leptospirosis canina a partir de información clínica. El valor de estos sistemas reside en complementar la evaluación, no en sustituir examen, confirmación diagnóstica o juicio profesional.

En oncología, una aplicación posible es predecir tiempo hasta recaída combinando grado histológico, márgenes, marcadores y tratamiento. Si el desenlace es tiempo a evento con censura, una clasificación binaria a plazo fijo puede perder información; modelos de Cox, supervivencia

aleatoria o enfoques de riesgos competitivos pueden ser más adecuados. La métrica debe reflejar horizonte temporal y calibración.

Producción animal y bienestar

La ganadería de precisión integra sensores de movimiento, consumo, rumia, leche y ambiente. Bao y Xie (2022) documentan aplicaciones de IA en producción animal y también limitaciones relacionadas con datos, costo, infraestructura y adopción. Las alertas deben valorarse según la respuesta operativa: un sistema sensible, pero con demasiados falsos positivos puede saturar al personal y terminar ignorado.

Visión computacional puede estimar condición corporal, postura, cojeras o densidad de animales. Audio puede apoyar la detección de tos en porcinos o aves. Modelos ambientales pueden relacionar índice temperatura-humedad con producción o estrés. La investigación debe impedir que la productividad se convierta en sustituto automático de bienestar; se requieren indicadores directos de salud, conducta y estado afectivo.

Epidemiología salud pública y Una Salud

Los modelos pueden integrar vigilancia, clima, movilidad y genómica para priorizar territorios o estimar riesgo de propagación. La utilidad depende de oportunidad y representatividad. Zonas con menor capacidad diagnóstica pueden aparecer artificialmente como de bajo riesgo. En zoonosis y resistencia antimicrobiana, combinar fuentes animales, humanas y ambientales requiere interoperabilidad, acuerdos de datos y definiciones compatibles.

Investigación experimental

La IA también participa en fenotipado, análisis automatizado de conducta, cuantificación histológica y selección de variables. Automatizar la medición puede reducir carga y ampliar frecuencia, pero no elimina error. Un sistema de visión que cuente lesiones debe validarse frente a observadores y examinar concordancia, no solo correlación. Si se utiliza dentro de un experimento, tratamientos, bloqueo, aleatorización y enmascaramiento continúan siendo decisiones del diseño.

1.8 Alcances y limitaciones de la inteligencia artificial

El alcance real de un algoritmo se define por su uso previsto. Un modelo entrenado para detectar una alteración en radiografías torácicas caninas de dos hospitales no queda validado para gatos, imágenes móviles, otras proyecciones ni decisiones terapéuticas. Población, contexto, usuario, momento, entrada, salida y acción posterior deben quedar especificados antes del desarrollo.

Calidad y representatividad de los datos

Los datos retrospectivos reflejan cómo se prestó atención y cómo se registró. Animales sin acceso a pruebas, fincas con conectividad limitada o especies menos estudiadas quedan subrepresentados. Un modelo puede reproducir esas desigualdades. El muestreo de conveniencia también altera prevalencia y valores predictivos. Evaluar subgrupos por especie, raza, edad, sexo, centro y sistema productivo ayuda a detectar fallos, aunque los subgrupos pequeños producen estimaciones imprecisas.

Validez interna y fuga de información

La fuga ocurre cuando el entrenamiento recibe información que no estaría disponible al momento de uso o que pertenece al conjunto de evaluación. Ejemplos: incluir una prueba confirmatoria obtenida después de la alerta, normalizar usando toda la base, seleccionar variables antes de la validación cruzada o colocar registros del mismo animal en ambos conjuntos. La fuga puede producir resultados extraordinarios que desaparecen en uso real.

Generalización calibración y deriva

Discriminación describe la capacidad para ordenar casos de mayor y menor riesgo; calibración compara probabilidades predichas con frecuencias observadas. Un AUC alto no asegura que una probabilidad de 0.70 corresponda a cerca de 70 casos por cada 100. Además, alimentación, genética, equipos, prevalencia y protocolos cambian. Esta deriva exige monitoreo, reevaluación y, cuando se justifique, recalibración.

Explicabilidad y responsabilidad

Una explicación local puede mostrar qué variables influyeron en una predicción, pero no demuestra causalidad ni garantiza que el modelo razone de modo clínicamente válido. Las importancias pueden ser inestables ante variables correlacionadas. La responsabilidad no se transfiere al algoritmo: institución, desarrolladores y profesionales deben definir quién revisa, quién puede anular la recomendación y cómo se registran daños o incidentes. Bellamy (2023) sostiene que el uso veterinario de IA necesita marcos regulatorios y profesionales acordes con sus riesgos.

Privacidad propiedad y bienestar

Los datos veterinarios pueden identificar tutores, explotaciones, ubicación, rendimiento económico y prácticas sanitarias. Deben establecerse base legal, consentimiento cuando corresponda, minimización, control de acceso, plazos de conservación y condiciones de reutilización. En producción animal, la gobernanza también debe atender propiedad y poder de negociación sobre datos. El bienestar exige analizar errores de omisión, intervenciones innecesarias y efectos de automatizar decisiones sobre individuos o poblaciones.

Tabla 1.7

Riesgos y medidas de control

Riesgo	**Ejemplo**	**Consecuencia**	**Medida de control**
Sesgo de selección	Solo hospitales de referencia	Peor desempeño en atención primaria	Muestreo multicéntrico y evaluación externa
Fuga de datos	Registros del mismo animal en ambos conjuntos	Desempeño inflado	Partición agrupada por animal y tiempo
Desbalance	Pocos casos de enfermedad	Exactitud engañosa	Métricas sensibles a prevalencia y ponderación
Mala calibración	Riesgos predichos demasiado altos	Exceso de pruebas o tratamientos	Curvas de calibración y recalibración
Deriva	Cambio de sensor o manejo	Pérdida silenciosa de desempeño	Monitoreo y reevaluación periódica
Automatización acrítica	Aceptar salida sin revisión	Daño clínico o de bienestar	Supervisión humana y protocolo de escalamiento

1.9 Mitos frecuentes sobre el uso de algoritmos

Los mitos persisten porque atribuyen a la técnica propiedades que pertenecen al diseño, a los datos o a la implementación. Desmontarlos es parte de la alfabetización científica.

Mito 1 Más datos garantizan mejor evidencia. La cantidad no corrige mediciones sistemáticamente erróneas ni una muestra sesgada. Millones de registros del mismo sistema pueden representar menos diversidad que una muestra menor y multicéntrica.

Mito 2 El algoritmo más complejo es superior. La complejidad solo se justifica si mejora desempeño relevante fuera de muestra, mantiene calibración y aporta utilidad suficiente para compensar costo y opacidad.

Mito 3 Una exactitud alta demuestra utilidad clínica. La exactitud depende de la prevalencia y del umbral. Deben considerarse falsos negativos, falsos positivos, capacidad operativa y beneficio neto.

Mito 4 Machine learning reemplaza la bioestadística. Diseño, incertidumbre, estructura jerárquica, tamaño muestral y sesgo siguen siendo problemas estadísticos. ML amplía herramientas; no elimina fundamentos.

Mito 5 La predicción identifica causas. Un predictor puede ser útil sin ser causal. Cambiarlo mediante una intervención puede no modificar el desenlace.

Mito 6 El modelo es objetivo porque usa matemáticas. Las decisiones sobre población, etiqueta, variables, función de pérdida y umbral incorporan valores y prioridades.

Mito 7 Balancear clases resuelve todos los problemas. El remuestreo puede ayudar durante el entrenamiento, pero no sustituye una evaluación en prevalencia real ni corrige etiquetas defectuosas.

Mito 8 Un modelo publicado está validado. Un resultado interno no garantiza transporte. Se requiere evaluación independiente y contextual.

Mito 9 Explicar una predicción la vuelve causal. SHAP u otras explicaciones describen la conducta del modelo. No identifican por sí mismas efectos de intervención.

Mito 10 La IA generativa consulta una base verdadera. Un modelo genera secuencias plausibles y puede fabricar referencias o datos. Toda salida científica necesita verificación.

1.10 Terminología fundamental

Tabla 1.8

Glosario operativo

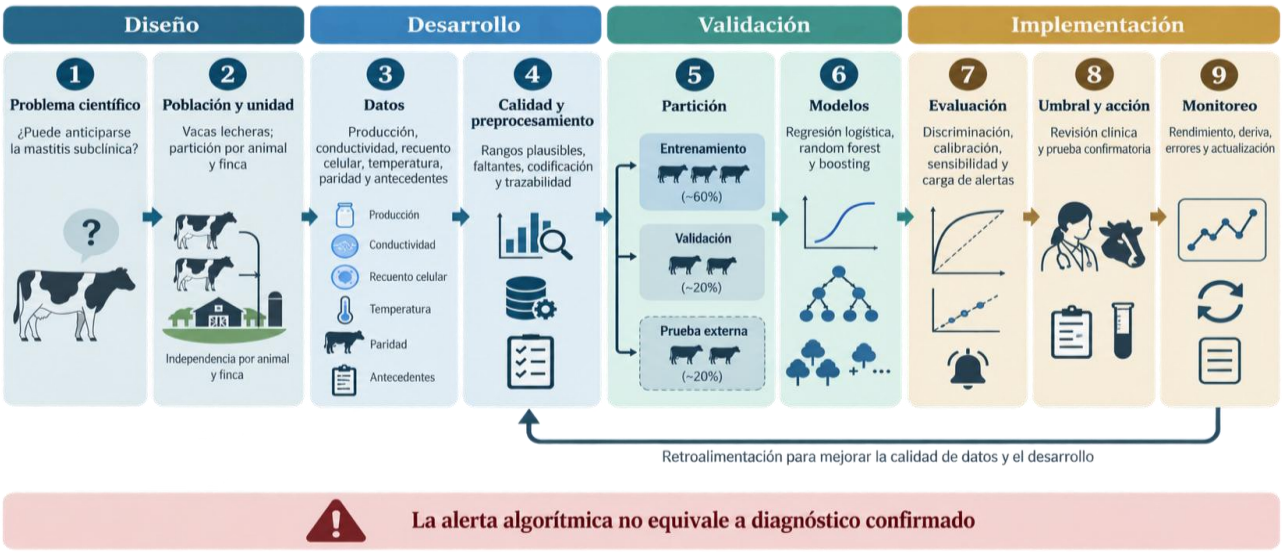
Término	Definición aplicada
Algoritmo	Conjunto definido de operaciones para resolver una tarea.
Característica o predictor	Variable de entrada disponible en el momento de la predicción.
Etiqueta o desenlace	Resultado que el modelo supervisado intenta estimar.
Instancia	Unidad o fila analizada; debe corresponder a la unidad científica definida.
Entrenamiento	Proceso de estimar parámetros del modelo con datos de desarrollo.
Hiperparámetro	Configuración definida fuera del ajuste directo, como profundidad de un árbol.
Validación cruzada	Repetición de particiones de desarrollo para estimar variabilidad y ajustar hiperparámetros.
Conjunto de evaluación	Datos reservados para estimar desempeño final fuera de muestra.
Sobreajuste	Aprendizaje de particularidades del entrenamiento que no generalizan.
Subajuste	Modelo demasiado limitado para capturar señal relevante.
Discriminación	Capacidad para ordenar o separar unidades con distintos desenlaces.
Calibración	Concordancia entre probabilidades predichas y frecuencias observadas.
Umbral	Punto que convierte una probabilidad en una decisión o categoría.
Sensibilidad	Proporción de casos verdaderos que el modelo identifica.
Especificidad	Proporción de no casos que el modelo descarta correctamente.
Valor predictivo positivo	Proporción de alertas positivas que son casos; depende de prevalencia.
AUC ROC	Probabilidad de ordenar por encima un caso respecto de un no caso elegido al azar.
Fuga de información	Entrada accidental de datos indisponibles o pertenecientes a evaluación.
Validación externa	Evaluación en datos independientes por lugar, tiempo o población.
Deriva	Cambio de datos o relación predictor desenlace después del despliegue.
Explicabilidad	Métodos que describen factores asociados a la salida del modelo.
Equidad algorítmica	Evaluación y mitigación de diferencias injustas de desempeño o impacto.

Modelo fundacional	Modelo entrenado a gran escala y adaptable a diversas tareas.
Alucinación	Salida generativa plausible pero falsa o no sustentada.
Reproducibilidad	Capacidad de obtener resultados coherentes con datos código y procedimientos documentados.

1.11 Caso didáctico predicción de mastitis subclínica en bovinos

Figura 1.3.

Caso didáctico predicción de mastitis subclínica en bovinos



Caso didáctico simulado. Los números de este apartado se construyen para enseñar el proceso y no corresponden a resultados de una explotación real. El planteamiento se inspira en la literatura que ha usado recuento de células somáticas, producción, composición láctea y variables del animal para predecir salud de ubre. Bobbo et al. (2021) compararon ocho métodos con registros de control lechero y mostraron que el rendimiento debe juzgarse fuera de muestra. Tian et al. (2024) evaluaron nueve algoritmos para mastitis clínica usando producción, rumia y conductividad eléctrica.

Problema científico y propósito

Una cooperativa lechera desea priorizar vacas para revisión y muestreo antes de que la alteración sea evidente. La pregunta es: ¿puede un modelo estimar, al final del ordeño del día t, la

probabilidad de que una vaca presente mastitis subclínica confirmada durante las siguientes 48 horas? El propósito es predictivo. El sistema emitiría una alerta para evaluación, no un diagnóstico definitivo ni una recomendación automática de antimicrobianos.

Población objetivo: vacas Holstein y cruzadas en lactación, pertenecientes a hatos con ordeño automatizado de la región estudiada. Unidad de predicción: vaca-día índice. Unidad biológica: vaca. Estructura jerárquica: días dentro de vacas y vacas dentro de hatos. Horizonte: 48 horas. Momento de uso: después del último ordeño del día índice.

Hipótesis y desenlace

Hipótesis predictiva: la integración de cambios longitudinales de conductividad eléctrica, producción, rumia, temperatura y antecedentes mejora la discriminación y la calibración frente a un modelo de referencia que utiliza solo el último recuento de células somáticas. Esta hipótesis no afirma que las variables sean causas de mastitis.

Desenlace: mastitis subclínica durante las 48 horas posteriores, definida en el protocolo mediante recuento de células somáticas superior al umbral adoptado y confirmación según el estándar veterinario seleccionado. El estudio debe justificar el umbral, el tipo de muestra, el manejo de cuartos y la repetición. Etiquetas dudosas se mantienen como categoría para análisis de sensibilidad o se excluyen mediante una regla previa al análisis.

Diseño y datos simulados

Se plantea una cohorte prospectiva en 12 hatos durante 10 meses. El conjunto simulado contiene 1 800 vacas y 126 000 vaca-días elegibles, con 6 300 eventos positivos a 48 horas, equivalentes a una prevalencia por ventana de 5 %. Para impedir pseudorreplicación, todas las observaciones de una vaca permanecen en la misma partición. Ocho hatos se destinan al desarrollo, dos a validación interna-temporal y dos se reservan como evaluación externa geográfica.

Tabla 1.9

Variables del caso simulado

Variable	Tipo	Disponibilidad	Riesgo o decisión

Conductividad eléctrica por cuarto	Serie temporal continua	Hasta el día índice	Deriva por equipo y temperatura
Producción diaria y cambio porcentual	Continua longitudinal	Hasta el día índice	Estacionalidad y etapa de lactación
Minutos de rumia y actividad	Serie temporal	Hasta el día índice	Pérdida de señal del collar
Días en leche y paridad	Continua y ordinal	Registro previo	No confundir asociación con causalidad
Último RCS log transformado	Continua	Solo mediciones anteriores	Evitar usar la medición confirmatoria futura
Tratamiento previo y antecedentes	Binaria y conteo	Historia anterior	Sesgo por prácticas del hato
Hato estación y lote	Categóricas	Día índice	Estructura y posible cambio de distribución
Mastitis subclínica 48 horas	Binaria	Después del día índice	Etiqueta con error diagnóstico

Preparación y prevención de fugas

1. Construir una fecha índice por vaca-día y congelar los predictores al final de ese día. Ninguna medición obtenida durante la confirmación entra al modelo.
2. Eliminar duplicados, armonizar unidades y documentar cambios de sensores. Los valores imposibles se contrastan con reglas biológicas antes de decidir su corrección.
3. Resumir ventanas de 1, 3 y 7 días mediante nivel, pendiente y variabilidad. Estas transformaciones se calculan con información pasada.
4. Imputar faltantes dentro de cada pliegue de entrenamiento. Añadir indicadores de ausencia cuando el proceso de pérdida pueda aportar información operativa.
5. Codificar categorías y escalar variables solo cuando el algoritmo lo requiera. Todas las transformaciones se estiman dentro del conjunto de desarrollo.
6. Separar por hato o por vaca según el objetivo. Para el examen final de transporte, conservar hatos completos sin exposición durante el ajuste.

Modelos y comparación

Se comparan cuatro enfoques: regla de alarma del sistema de ordeño, regresión logística penalizada, random forest y XGBoost. La regla representa la práctica vigente y evita proclamar

ventaja contra un estándar débil. Los hiperparámetros se ajustan mediante validación cruzada agrupada por vaca y hato. El umbral operativo se elige en validación, antes de observar la evaluación externa, con prioridad por sensibilidad y un límite de alertas que el personal pueda revisar cada día.

El tamaño muestral no se decide por una regla fija de "eventos por variable" ni por el número total de filas. Debe considerar número de animales y hatos independientes, prevalencia, complejidad del modelo, desempeño esperado y precisión de las métricas. Las 126 000 ventanas no equivalen a 126 000 sujetos independientes.

Resultados simulados de evaluación externa

Tabla 1.10

Desempeño externo del caso simulado

Modelo	AUC ROC	Sensibilidad	Especificidad	VPP	Brier
Regla vigente	0.68	0.61	0.70	0.10	0.052
Regresión penalizada	0.79	0.82	0.67	0.12	0.044
Random forest	0.81	0.84	0.69	0.13	0.042
XGBoost	0.82	0.85	0.70	0.13	0.041

Interpretación del caso simulado. XGBoost discrimina mejor que la regla vigente, pero la diferencia frente a random forest es pequeña. Con prevalencia de 5 %, un valor predictivo positivo de 0.13 significa que alrededor de 13 de cada 100 alertas corresponderían a un evento confirmado. El sistema podría ser útil para cribado si la revisión es poco invasiva y si la reducción de casos omitidos compensa la carga de falsas alertas. No sería suficiente para iniciar tratamiento.

La sensibilidad de 0.85 implica que el modelo detectaría aproximadamente 85 de cada 100 eventos, dejando 15 sin alerta. La especificidad de 0.70 implica 30 falsas alertas por cada 100 ventanas sin evento. Estas consecuencias deben expresarse también por 1 000 vaca-días y por hato para planificar trabajo. El Brier menor sugiere mejor calidad global de probabilidades, pero debe acompañarse de una curva de calibración y de intervalos de confianza agrupados.

Validez utilidad y análisis de errores

El análisis por hato puede mostrar variación que el promedio oculta. Si la sensibilidad cae en hatos con un modelo de sensor diferente, la causa probable es cambio de distribución. Se revisan errores por etapa de lactación, paridad, raza y estación. También se inspeccionan registros alrededor de tratamientos, porque una práctica clínica previa puede modificar predictores y etiquetas.

La utilidad final requiere un ensayo pragmático por conglomerados o un diseño escalonado que compare manejo con y sin alertas. Los desenlaces deben incluir mastitis clínica, días hasta confirmación, uso de antimicrobianos, leche descartada, bienestar, costos y carga de trabajo. El efecto de la intervención se estima con un modelo acorde al diseño, por ejemplo un modelo mixto con hato como conglomerado y periodo como factor. Ese estudio responde una pregunta causal distinta de la evaluación predictiva.

Reproducibilidad y documentación

El protocolo registra versión de datos, diccionario, decisiones de exclusión, código, semilla, particiones, entorno de software, hiperparámetros, métricas con incertidumbre y análisis de subgrupos. La publicación sigue TRIPOD+AI en lo aplicable. Antes del uso se prepara una ficha del modelo con población, uso previsto, exclusiones, fecha, variables, umbral, rendimiento, límites y procedimiento de suspensión.

Consideraciones éticas del caso

- La alerta no reemplaza examen veterinario ni confirmación microbiológica cuando esté indicada.
- Los datos de productores se usan bajo acuerdos que definen finalidad acceso conservación y reutilización.
- No se inicia tratamiento antimicrobiano solo por una predicción.
- Los falsos negativos se vigilan mediante protocolos clínicos paralelos.
- Los efectos sobre bienestar y carga de trabajo se evalúan como desenlaces del sistema.
- El modelo se suspende si la calibración o la sensibilidad caen por debajo de límites predefinidos.

1.12 Síntesis autoevaluación y actividad práctica

Síntesis del capítulo

La IA aplicada a Ciencias Veterinarias reúne sistemas capaces de clasificar, predecir, generar o apoyar decisiones. Machine learning aprende patrones desde datos; deep learning es una familia de modelos de representación profunda; la IA generativa produce contenido y requiere verificación. Ciencia de datos cubre el ciclo completo de datos, mientras la bioestadística aporta diseño, inferencia e incertidumbre.

Descripción, inferencia, predicción y causalidad responden preguntas diferentes. Un modelo predictivo útil necesita un desenlace válido, separación temporal, evaluación fuera de muestra, calibración, comparación con la práctica y análisis de errores. La complejidad no es una virtud aislada. Las restricciones de bienestar, privacidad, equidad y responsabilidad forman parte del método, no de un anexo posterior.

Los datos veterinarios presentan dependencia por paciente, hato, tiempo y dispositivo. Partir filas al azar suele ser incorrecto cuando esas filas comparten animal o centro. La validación externa, el monitoreo de deriva y la evaluación de impacto permiten pasar de un resultado computacional a evidencia de utilidad. El caso de mastitis muestra que una alerta puede priorizar observación, pero no equivale a diagnóstico ni justifica tratamiento por sí sola.

Preguntas de autoevaluación

1. Defina inteligencia artificial sin atribuir al sistema conciencia o intención.
2. Explique por qué un proyecto de ciencia de datos puede no contener machine learning.
3. Compare machine learning convencional y deep learning mediante un problema veterinario.
4. ¿Qué diferencia existe entre una herramienta generativa y una fuente científica?
5. Formule un ejemplo veterinario de aprendizaje supervisado y otro no supervisado.
6. ¿Por qué el aprendizaje por refuerzo exige especial cautela cuando las acciones afectan animales?
7. Diferencie una pregunta inferencial de una predictiva utilizando un ensayo posquirúrgico.

8. Explique por qué una variable predictiva no debe interpretarse automáticamente como causa.
9. Identifique la unidad experimental y la unidad observacional en un estudio con mediciones diarias de vacas agrupadas en hatos.
10. ¿Qué partición usaría para evaluar un modelo destinado a una clínica nunca observada?
11. Explique cómo la prevalencia modifica el valor predictivo positivo.
12. ¿Qué información aporta la calibración que no aporta el AUC?
13. Proponga dos formas de fuga de información en un estudio radiográfico canino.
14. ¿Qué riesgos plantea imputar resultados de laboratorio no solicitados?
15. ¿Por qué una exactitud de 95 % puede ser inútil en una enfermedad con prevalencia de 5 %?
16. Distinga explicación algorítmica y explicación causal.
17. ¿Qué variables consideraría para monitorear deriva en un sistema de ordeño?
18. Describa una consecuencia ética de un falso negativo y otra de un falso positivo.
19. ¿Qué evidencia adicional se necesita después de demostrar buen desempeño predictivo?
20. Justifique por qué ningún algoritmo corrige un experimento mal diseñado.

Actividad práctica

Título de la actividad Auditoría metodológica de una propuesta de IA veterinaria Propósito. Diseñar y defender un protocolo breve de predicción o clasificación en una de estas áreas: dolor posoperatorio canino, cojeras bovinas, mortalidad neonatal porcina, lesiones pulmonares aviares o recaída oncológica. La actividad puede realizarse en equipos de tres o cuatro participantes.

Producto esperado

- Una ficha de dos páginas con problema pregunta población uso previsto unidad de análisis y horizonte temporal.
- Una tabla con desenlace predictores fuente momento de disponibilidad escala y riesgo de sesgo.
- Un esquema de partición que respete animal centro y tiempo.

- Selección de un modelo de referencia y dos candidatos con justificación.
- Plan de métricas que incluya discriminación calibración utilidad y análisis por subgrupos.
- Matriz ética con privacidad bienestar responsabilidad y procedimiento ante errores.
- Un párrafo que separe lo que el estudio podría predecir de lo que no permitiría concluir causalmente.

Secuencia de trabajo

1. Definir la decisión que se pretende apoyar y quién utilizará la salida.
2. Dibujar la línea temporal entre predictores fecha índice desenlace y acción.
3. Marcar cualquier variable que solo estaría disponible después de la decisión.
4. Identificar dependencias y proponer una partición sin fuga.
5. Elegir métricas según prevalencia y costos de error.
6. Redactar un criterio de éxito previo al análisis.
7. Realizar una revisión cruzada con otro equipo y corregir dos debilidades.

Rúbrica orientativa

Tabla 1.11

Criterios de evaluación de la actividad

Criterio	Ponderación	Evidencia esperada
Coherencia científica	25 %	Pregunta desenlace y población alineados
Diseño y datos	25 %	Unidad temporalidad jerarquía y fuga resueltas
Evaluación del modelo	20 %	Referencia métricas calibración y validación externa
Interpretación	15 %	Diferencia entre predicción inferencia y causalidad
Ética y aplicación	15 %	Bienestar privacidad responsabilidad y acciones seguras

CAPÍTULO 2

LA INTELIGENCIA ARTIFICIAL DENTRO DEL MÉTODO CIENTÍFICO

Introducción

La incorporación de inteligencia artificial (IA) a la investigación veterinaria debe comprenderse como una ampliación del instrumental científico y no como un método científico alternativo. Los modelos de lenguaje pueden proponer términos de búsqueda, los algoritmos de aprendizaje automático pueden detectar patrones de riesgo y la visión computacional puede cuantificar rasgos en imágenes; sin embargo, ninguno de estos recursos decide por sí mismo qué problema merece ser investigado, qué comparación es éticamente admisible o qué resultado posee relevancia clínica. Esas decisiones dependen del conocimiento disciplinar, de la teoría, del diseño y de la responsabilidad del equipo investigador.

En ciencias veterinarias, el problema se vuelve especialmente exigente porque conviven pacientes individuales, poblaciones animales, unidades productivas, agentes infecciosos, ecosistemas y personas. Una pregunta sobre mastitis puede referirse al efecto causal de una intervención en vacas, a la prevalencia dentro de un hato, a la predicción de nuevos casos o a la clasificación automática de registros de ordeño. Aunque compartan datos, cada objetivo exige un diseño, un estimando y un criterio de validación distintos. Confundirlos conduce a análisis técnicamente vistosos, pero científicamente débiles.

La literatura reciente reconoce aplicaciones de IA en diagnóstico, epidemiología, vigilancia, patología, genómica y apoyo a la investigación veterinaria, a la vez que advierte sobre alucinaciones, sesgos, privacidad y falta de validación externa (Chu, 2024). La cuestión central, por tanto, no es si se utiliza IA, sino en qué fase, con qué propósito, bajo qué controles y con qué trazabilidad. Este capítulo desarrolla esa integración desde el planteamiento del problema hasta la comunicación de resultados.

Objetivos de aprendizaje

- Distinguir un tema general de un problema científico veterinario delimitado y susceptible de investigación.
- Formular preguntas, objetivos e hipótesis coherentes con propósitos de explicación, estimación, clasificación o predicción.

- Identificar usos legítimos de la IA en la búsqueda bibliográfica, la construcción conceptual, el diseño de protocolos y el análisis.
- Reconocer riesgos de sesgo, automatización acrítica, fuga de información, fabricación de referencias y pérdida de reproducibilidad.
- Diseñar un protocolo básico para evaluar dolor posoperatorio canino, definiendo unidad experimental, desenlace, intervención, análisis y protección del bienestar animal.

Figura 2.1.

Cadena de coherencia de una investigación veterinaria



2.1. El problema de investigación veterinaria

Un problema de investigación es una discrepancia argumentada entre lo que se conoce y lo que se necesita conocer para comprender, estimar o resolver una situación. No equivale al tema. “Dolor en perros” es un campo; “incertidumbre sobre si un protocolo analgésico reduce el dolor agudo durante las primeras 24 horas después de ovariohisterectomía electiva” es un problema potencial. La formulación se vuelve científica cuando identifica población, contexto, fenómeno, consecuencias, conocimiento previo y una incertidumbre verificable.

La relevancia veterinaria puede ser clínica, productiva, epidemiológica, ambiental, económica o ética. En una clínica de pequeños animales, una frecuencia elevada de analgesia de rescate puede señalar un problema de control del dolor. En bovinos, la reducción del rendimiento lechero asociada al estrés térmico puede justificar un estudio de estimación o predicción. En salud pública, la detección tardía de un agente zoonótico puede motivar un sistema de alerta. La disponibilidad de una base de datos no convierte automáticamente cualquiera de estas situaciones en una pregunta valiosa: primero debe establecerse qué decisión podría mejorar con la evidencia.

Un problema bien delimitado responde, al menos, a seis cuestiones: qué ocurre, en quiénes, dónde, durante qué periodo, por qué importa y qué incertidumbre permanece. También anticipa las restricciones. Si las historias clínicas registran dolor de forma irregular, esa carencia es parte del problema metodológico; si los sensores provienen de un solo hato, la transportabilidad a otros sistemas productivos será limitada; si el desenlace depende de una evaluación subjetiva, deben planificarse entrenamiento y enmascaramiento.

Componente	**Pregunta de control**	**Ejemplo veterinario**
Situación observada	¿Qué patrón, evento o decisión genera incertidumbre?	Uso frecuente de analgesia de rescate tras ovariohisterectomía.
Población y contexto	¿En qué animales, establecimientos y periodo?	Perras adultas sanas atendidas en dos clínicas docentes.
Consecuencia	¿Qué daño o pérdida puede producir?	Sufrimiento, recuperación tardía y mayor estancia hospitalaria.
Conocimiento disponible	¿Qué explica ya la literatura?	Existen escalas validadas y recomendaciones multimodales, pero los protocolos locales no se han evaluado.
Vacío específico	¿Qué relación o desempeño se desconoce?	Diferencia en la trayectoria de dolor entre dos protocolos aplicables localmente.
Decisión vinculada	¿Qué decisión cambiaría si se obtiene evidencia?	Adopción o modificación del protocolo analgésico institucional.

De la observación al problema investigable

La observación clínica inicial suele contener ruido: variación entre evaluadores, diferencias de procedimiento, registros incompletos y selección de casos. Antes de atribuir el fenómeno a un tratamiento, conviene construir un mapa causal preliminar. En dolor posoperatorio podrían

intervenir tipo de cirugía, experiencia del cirujano, duración anestésica, edad, peso, ansiedad, analgesia preventiva, momento de la medición y conducta del evaluador. La IA puede ayudar a ordenar una lista extensa de factores, pero la plausibilidad causal debe contrastarse con farmacología, anestesiología, cirugía y evidencia publicada.

Ejemplo breve 1. Producción de leche y temperatura

Tema: estrés térmico en bovinos lecheros.

Problema delimitado: se desconoce la forma y el punto de cambio de la relación entre el índice temperatura-humedad diario y la producción individual, considerando días en leche y repetición de medidas por vaca.

Método coherente: un modelo aditivo generalizado o un modelo mixto puede ser más informativo que un algoritmo de caja negra si la prioridad es estimar la curva y comunicar su incertidumbre.

Criterios para valorar un problema

- Pertinencia: el resultado puede modificar una decisión clínica, sanitaria, productiva o de bienestar.
- Originalidad situada: aporta evidencia nueva para una especie, población, contexto o mecanismo; no exige que nadie haya estudiado jamás el tema.
- Factibilidad: existen acceso, tiempo, muestra, instrumentos, competencias y presupuesto compatibles con el diseño.
- Rigor: el desenlace puede definirse y medirse con validez suficiente; los sesgos principales pueden controlarse.
- Ética: la pregunta no expone animales a daño evitable y contempla analgesia de rescate, consentimiento, confidencialidad y revisión ética.
- Utilidad: la magnitud de interés y el destinatario de la evidencia están definidos antes del análisis.

2.2. Formulación de preguntas científicas e hipótesis

La pregunta científica traduce el problema en una relación que puede examinarse mediante datos. En estudios de intervención resulta útil la estructura PICOT: población, intervención, comparación, desenlace y tiempo. Para exposiciones se emplea PECO; en diagnóstico puede añadirse la prueba índice y el estándar de referencia. Estas estructuras son ayudas, no moldes

rígidos. Una pregunta etológica cualitativa o una investigación de implementación puede requerir otros componentes.

Una pregunta robusta especifica el nivel de inferencia. "¿La radiografía identifica displasia?" es ambigua. Puede significar la exactitud de un veterinario, el desempeño de una red neuronal frente a un estándar, la concordancia entre lectores o el efecto de incorporar el algoritmo sobre decisiones y resultados. Cada versión demanda datos distintos. La IA generativa puede producir variantes, pero el investigador debe rechazar las que mezclan desenlaces, cambian la población o prometen causalidad con un diseño observacional.

Tipo de pregunta	**Forma orientativa**	**Diseño y análisis congruentes**
Intervención	En P, ¿I frente a C modifica Y durante T?	Ensayo aleatorizado; diferencia media, riesgo relativo, modelo mixto.
Etiología o exposición	En P, ¿E se asocia con Y respecto de C?	Cohorte o caso-control; regresión con confusores definidos a priori.
Diagnóstico	¿Qué exactitud tiene la prueba I frente al estándar R?	Estudio de exactitud; sensibilidad, especificidad, razones de verosimilitud.
Pronóstico	¿Qué variables disponibles en t0 predicen Y en T?	Cohorte; validación interna/externa, calibración y discriminación.
Prevalencia	¿Qué proporción de P presenta Y en un periodo?	Estudio transversal con muestreo representativo e intervalo de confianza.
Clasificación	¿Puede un modelo asignar correctamente clases definidas?	Conjunto etiquetado, partición por unidad independiente, matriz de confusión.
Causalidad	¿Cuál sería Y si la misma población recibiera I frente a C?	Ensayo o diseño observacional causal con supuestos explícitos.

Hipótesis científica, estadística y predictiva

La hipótesis científica propone una explicación o expectativa sustantiva. La hipótesis estadística formaliza un contraste sobre parámetros; por ejemplo, igualdad de medias o ausencia de interacción. Una hipótesis predictiva plantea que un conjunto de variables disponibles antes del desenlace permitirá predecirlo con rendimiento útil en datos nuevos. No todas las investigaciones necesitan hipótesis nula: un estudio de prevalencia puede centrarse en precisión, y un desarrollo predictivo debe preespecificar métricas y validación más que una prueba de significación.

En un ensayo analgésico, H1 podría establecer que añadir un fármaco al protocolo estándar reduce la trayectoria de dolor clínicamente relevante durante 24 horas. La hipótesis estadística podría expresarse como una diferencia distinta de cero en el efecto promedio de tratamiento o en la interacción tratamiento por tiempo. Para que sea interpretable, debe fijarse el desenlace primario, la dirección esperada, el periodo y una diferencia mínima de interés clínico. Elaborar la hipótesis después de conocer los resultados convierte una predicción en una explicación retrospectiva.

Ejemplo breve 2. Radiografías torácicas caninas

Pregunta de clasificación: ¿un modelo entrenado con radiografías etiquetadas por especialistas distingue cardiomegalia de ausencia de cardiomegalia en imágenes de hospitales no usados en el entrenamiento?

Unidad independiente: paciente, no imagen. Todas las proyecciones de un perro deben quedar en una sola partición.

Métricas: sensibilidad, especificidad, valores predictivos, AUC, calibración e intervalos de confianza. La exactitud aislada puede ser engañosa si las clases están desbalanceadas.

2.3. Objetivos de explicación, estimación, clasificación y predicción

El verbo del objetivo orienta el método. Explicar supone comprender mecanismos o relaciones causales; estimar significa cuantificar un parámetro con incertidumbre; clasificar asigna observaciones a categorías; predecir anticipa resultados en individuos o unidades nuevas. Un mismo conjunto de datos puede permitir varias tareas, pero no las vuelve equivalentes. Un predictor con alta importancia en random forest no es necesariamente una causa, y una asociación estadísticamente significativa puede añadir poco a la predicción.

Propósito	Producto científico	Pregunta clave	Criterios principales
Explicación	Mecanismo o efecto causal defendible	¿Qué cambia Y y por qué?	Diseño, control de confusión, temporalidad, supuestos causales.
Estimación	Parámetro e intervalo de confianza	¿Cuánto o con qué precisión?	Representatividad, sesgo, error estándar, intervalo de confianza.
Clasificación	Etiqueta o categoría	¿A qué clase pertenece?	Sensibilidad, especificidad, precisión, F1, errores por clase.

Predicción	Probabilidad o valor futuro	¿Qué ocurrirá en un caso nuevo?	Calibración, discriminación, error, validación externa y utilidad.
Descripción	Distribución o patrón observado	¿Qué ocurrió en la muestra?	Calidad del registro, denominadores y visualización adecuada.

Coherencia entre objetivo, datos y método

Si el objetivo es estimar prevalencia de resistencia antimicrobiana, la prioridad es un muestreo que represente a la población y una definición de caso consistente. Un clasificador complejo no soluciona el sesgo de remisión a un laboratorio. Si el objetivo es predecir hospitalización felina, las variables deben estar disponibles en el momento real de la decisión; incorporar resultados tomados después del ingreso produce fuga de información. Si se pretende estimar el efecto de una intervención nutricional, el reparto de tratamientos, la adherencia y la confusión requieren más atención que la AUC.

TRIPOD+AI amplía las recomendaciones de reporte para modelos de predicción construidos mediante regresión o aprendizaje automático y enfatiza transparencia en datos, desarrollo, evaluación y uso previsto (Collins et al., 2024). Para veterinaria, estas exigencias deben complementarse con información sobre especie, raza, centro, equipo de medición, prevalencia del desenlace y estructura jerárquica. Un modelo evaluado en imágenes del mismo escáner y hospital puede aprender señales locales que no persisten en otro entorno.

2.4. Uso de IA en la delimitación del problema

La IA puede actuar como herramienta de exploración controlada. Un modelo de lenguaje puede descomponer un tema en poblaciones, desenlaces, mecanismos y diseños; agrupar notas de campo; sugerir sinónimos; transformar una observación clínica en varias preguntas candidatas; o señalar variables que conviene discutir con especialistas. Su mejor uso ocurre cuando el investigador aporta contexto suficiente y exige que la salida distinga hechos, inferencias y elementos por verificar.

El proceso debería conservar un registro de las instrucciones, versión o fecha de la herramienta, texto de salida y decisiones humanas. La trazabilidad permite identificar cuándo una sugerencia modificó el protocolo. Chu (2024) describe usos de modelos generativos en ámbitos clínicos,

educativos y de investigación veterinaria, pero también subraya el riesgo de alucinación. Por ello, la salida no debe citarse como autoridad ni sustituir la consulta de fuentes originales.

Secuencia de delimitación asistida

- Redactar la observación inicial sin IA y separar hechos de interpretaciones.
- Proporcionar a la herramienta especie, contexto, población, decisión, variables disponibles y restricciones éticas.
- Solicitar alternativas de preguntas clasificadas por propósito científico, no una "pregunta perfecta".
- Contrastar cada alternativa con literatura primaria y con expertos clínicos, estadísticos y de bienestar animal.
- Elegir la pregunta por pertinencia y factibilidad; documentar las alternativas descartadas y la razón.
- Congelar la versión final antes de explorar los datos que se usarán para probarla.

Prompt científico orientativo

"Actúa como asistente metodológico. A partir de esta observación clínica [describir], genera cuatro preguntas diferentes: una de estimación, una causal, una predictiva y una de implementación. Para cada una especifica población, unidad experimental, desenlace, momento de medición, diseño compatible, principales sesgos y datos adicionales necesarios. No cites referencias ni supongas hechos no proporcionados. Marca todo elemento que requiera verificación por un veterinario especialista."

Por qué está bien formulado: define función, contexto, productos, límites y verificación humana; además, evita que la herramienta mezcle propósitos científicos.

2.5. Búsqueda, selección y evaluación crítica de literatura

La búsqueda bibliográfica rigurosa comienza con conceptos y no con frases completas. Para dolor posoperatorio canino pueden distinguirse especie, procedimiento, intervención, comparador, escala y periodo. Cada concepto se traduce a vocabulario controlado y términos libres, incluyendo variantes ortográficas. PubMed/MEDLINE, CAB Abstracts, Scopus y Web of Science cubren dominios diferentes; limitarse a una sola fuente puede omitir estudios veterinarios, producción animal o congresos pertinentes.

La IA generativa puede proponer sinónimos, pero no debe inventar descriptores ni asegurar que una búsqueda es exhaustiva. La estrategia final debe ejecutarse y documentarse en las bases elegidas. PRISMA 2020 promueve un reporte transparente de identificación, selección, evaluación y síntesis de estudios (Page et al., 2021); PRISMA-S profundiza en el registro reproducible de las búsquedas (Rethlefsen et al., 2021). Incluso en una revisión narrativa para marco teórico, guardar cadenas, fechas y criterios reduce la selección oportunista.

IA en el cribado: apoyo, no decisión opaca

El aprendizaje activo puede priorizar registros probablemente relevantes y reducir carga de cribado. El marco ASReview muestra cómo un sistema abierto puede aprender de decisiones sucesivas del revisor y ordenar los artículos pendientes (van de Schoot et al., 2021). Esto no significa que la máquina deba excluir definitivamente estudios sin supervisión. El equipo necesita un conjunto inicial de ejemplos, criterios explícitos, control de errores y una regla de detención justificable. En preguntas con pocos estudios relevantes o terminología inestable, un falso negativo puede cambiar toda la conclusión.

Fase	**Aporte posible de IA**	**Control humano indispensable**
Construcción de términos	Sinónimos, traducciones y combinaciones iniciales.	Verificar descriptores, alcance y sintaxis en cada base.
Deduplicación	Detectar coincidencias aproximadas entre registros.	Revisar títulos similares, preprints y versiones finales.
Cribado	Priorizar títulos/resúmenes mediante aprendizaje activo.	Doble revisión o auditoría; registrar exclusiones y regla de detención.
Extracción	Proponer campos y localizar fragmentos relevantes.	Cotejar cada dato con tabla, método o suplemento original.
Evaluación	Organizar respuestas a una lista de comprobación.	Juzgar sesgo y aplicabilidad; resolver discrepancias por consenso.
Síntesis	Agrupar temas y generar tablas preliminares.	Recalcular resultados y escribir interpretación propia.

Evaluación crítica

Leer no es acumular resultados. La evaluación crítica pregunta si el diseño puede responder a la pregunta, si la unidad experimental fue respetada, si la muestra representa la población, si hubo

aleatorización y enmascaramiento, cómo se manejaron pérdidas y datos faltantes, si el desenlace es válido, si se informó tamaño del efecto con incertidumbre y si los resultados son transferibles. En estudios predictivos se añaden separación correcta de datos, prevención de fuga, calibración, validación externa y comparación con una referencia clínica razonable.

Para investigación in vivo, ARRIVE 2.0 organiza requisitos esenciales de reporte, como diseño, tamaño muestral, criterios de inclusión y exclusión, aleatorización, enmascaramiento, desenlaces y métodos estadísticos (Percie du Sert et al., 2020). Estas pautas son útiles desde la planificación, no solo al redactar. En un trabajo veterinario clínico también deben consultarse guías específicas del tipo de estudio y requisitos del comité de ética institucional.

2.6. Construcción del marco teórico y estado del arte

El marco teórico explica los conceptos y relaciones que sostienen la pregunta. El estado del arte describe qué se ha investigado, con qué métodos, en qué poblaciones y con qué resultados o controversias. No son una sucesión de resúmenes. Una construcción sólida compara definiciones, identifica mecanismos plausibles y muestra cómo la evidencia conduce al problema específico.

En un estudio de dolor posoperatorio, el marco puede integrar nocicepción, sensibilización, comportamiento, analgesia multimodal, propiedades de la escala elegida y factores de variación. El estado del arte comparará protocolos, tiempos de observación, rescate analgésico, procedimientos quirúrgicos y limitaciones. Las guías AAHA y WSAVA recomiendan valoración sistemática y estrategias multimodales, con énfasis en reconocimiento y seguimiento del dolor (Gruen et al., 2022; Monteiro et al., 2023). Esas recomendaciones orientan la justificación, pero no reemplazan la evaluación de ensayos primarios.

Matriz de evidencia

Una matriz evita que la escritura dependa de memoria o de fragmentos generados. Por cada fuente se registran diseño, especie, población, intervención o exposición, comparador, desenlaces, estimadores, limitaciones, aplicabilidad y aporte al argumento. La IA puede extraer candidatos, pero cada celda debe incluir página, tabla o sección de procedencia. Si un dato no puede localizarse en el original, se elimina o se marca para verificación.

Campo	Contenido esperado
Referencia	Identificador verificable, DOI o registro bibliográfico.
Pregunta y diseño	Qué intentó responder y mediante qué arquitectura.
Población	Especie, raza, edad, condición, centro y criterios.
Unidad y muestra	Unidad experimental, unidad observacional y tamaño analizado.
Variables	Definiciones, instrumentos, tiempos y posibles errores.
Resultado	Estimador, intervalo de confianza y no solo valor p.
Sesgo/aplicabilidad	Limitaciones que afectan credibilidad y transferencia.
Uso en el capítulo	Afirmación concreta que la fuente respalda.

IA generativa y redacción del marco

Los modelos de lenguaje producen texto plausible a partir de patrones, no una lectura garantizada de la evidencia. Pueden fusionar estudios, cambiar denominadores o fabricar referencias. Experimentos con resúmenes científicos han mostrado que textos generados pueden parecer convincentes y resultar difíciles de distinguir, lo que refuerza la necesidad de verificación documental (Gao et al., 2023). Van Dis et al. (2023) proponen mantener control humano, transparencia, calidad y responsabilidad en el uso científico de sistemas conversacionales.

La práctica más segura consiste en escribir primero un esquema argumental basado en la matriz de evidencia. Después puede pedirse a la IA que detecte saltos lógicos, términos no definidos o redundancias, sin entregarle datos confidenciales. La versión final debe atribuir cada afirmación empírica a una fuente consultada, y el autor conserva responsabilidad por exactitud, originalidad y coherencia.

2.7. Identificación de vacíos de conocimiento

Un vacío no es la simple ausencia de publicaciones. Puede ser una inconsistencia, una población subrepresentada, un desenlace mal medido, falta de validación externa, insuficiente duración del seguimiento, heterogeneidad metodológica o ausencia de análisis de utilidad clínica. También puede existir mucha literatura y, sin embargo, poca evidencia capaz de responder a la decisión local.

Tipo de vacío	Ejemplo veterinario	Respuesta metodológica

Poblacional	Modelos de imagen entrenados casi solo con razas grandes.	Muestreo multicéntrico y evaluación por raza/tamaño.
Geográfico	Predicción de mastitis sin datos de sistemas tropicales.	Validación externa en hatos con clima y manejo distintos.
De medición	Dolor registrado con escalas no validadas.	Usar instrumento validado y entrenar evaluadores.
De diseño	Asociaciones observacionales interpretadas como efecto.	Ensayo o estrategia causal con confusores predefinidos.
De validación	AUC alta solo en validación interna.	Validación temporal o geográfica y análisis de calibración.
De implementación	Buen desempeño sin estudiar cambios en decisiones.	Estudio prospectivo de impacto, seguridad y carga de trabajo.
De equidad	Rendimiento no informado por subgrupos.	Evaluación estratificada y análisis de errores relevantes.

Procedimiento para justificar un vacío

- Definir el cuerpo de evidencia pertinente y su periodo, sin afirmar que "no existen estudios" tras una búsqueda superficial.
- Comparar poblaciones, diseños, mediciones, estimadores y validaciones, no solo conclusiones.
- Diferenciar evidencia ausente de evidencia inconsistente o imprecisa.
- Explicar por qué el vacío afecta una decisión veterinaria concreta.
- Proponer un estudio capaz de reducir esa incertidumbre; repetir el mismo diseño débil no llena el vacío.
- Reconocer cuando la contribución será local o exploratoria y evitar promesas universales.

Ejemplo breve 3. Mastitis subclínica bovina

Vacío aparente: "faltan modelos de IA para mastitis". Formulación insuficiente, porque ya existen múltiples enfoques.

Vacío defendible: falta validación prospectiva y calibración en hatos pequeños de altura que usan equipos y frecuencias de medición diferentes a los conjuntos de entrenamiento publicados.

Contribución coherente: evaluar un modelo existente o parsimonioso en ese dominio antes de desarrollar otro algoritmo.

2.8. Elaboración de protocolos de investigación

El protocolo convierte la pregunta en decisiones verificables antes de observar los resultados. Debe permitir que otra persona comprenda quién será incluido, qué se hará, qué se medirá, cuándo, cómo se analizará y cómo se protegerá a los animales y a las personas. La preespecificación reduce la flexibilidad analítica que favorece resultados selectivos.

En estudios que incorporan IA, el protocolo debe describir además origen y gobernanza de datos, variables disponibles en el momento de uso, partición por unidades independientes, preprocesamiento aprendido solo con datos de entrenamiento, manejo de clases desbalanceadas, ajuste de hiperparámetros, métricas, calibración, análisis de subgrupos y validación externa. La declaración TRIPOD+AI es una guía de reporte útil para anticipar estos componentes (Collins et al., 2024).

Sección del protocolo	Preguntas mínimas
Justificación	¿Qué evidencia y qué vacío sustentan el estudio?
Pregunta y objetivos	¿Cuál es el propósito: estimar, explicar, clasificar o predecir?
Diseño	¿Qué arquitectura responde a la pregunta y cuál es la unidad experimental?
Población y muestra	¿Quién es elegible, cómo se recluta y cómo se calculó el tamaño?
Intervenciones/exposiciones	¿Cómo se asignan, administran y registran?
Desenlaces	¿Cuál es el primario, con qué instrumento y en qué momentos?
Control de sesgo	¿Cómo se aleatoriza, bloquea, enmascara y controla confusión?
Plan estadístico	¿Qué estimando, modelo, contraste, intervalos y sensibilidad se usarán?
Plan de IA	¿Qué tarea, partición, validación, métricas, auditoría y versión?
Ética y bienestar	¿Qué riesgo, rescate, consentimiento, privacidad y monitoreo existen?
Reproducibilidad	¿Cómo se gestionarán datos, código, diccionario y cambios?
Difusión	¿Cómo se reportarán resultados favorables, nulos y adversos?

Plan de análisis antes de recolectar

El plan no debe limitarse a nombrar una prueba. Debe definir el estimando: por ejemplo, diferencia media entre grupos en el área bajo la curva de dolor de 2 a 24 horas, bajo una estrategia por intención de tratar. También especificará covariables, estructura de correlación, manejo de datos faltantes, supuestos, transformaciones, multiplicidad, análisis de sensibilidad y representación

gráfica. Si se planea un modelo predictivo, deben fijarse horizonte, población de uso, momento de predicción, resultado, referencia, métricas y umbral o regla de decisión.

Una IA generativa puede convertir notas en una lista de apartados o revisar consistencia terminológica. No debe elegir automáticamente el tamaño muestral ni una prueba sin conocer distribución, estructura jerárquica, efecto mínimo, tasa de eventos y pérdidas esperadas. Los cálculos generados deben reproducirse en software estadístico y ser revisados por una persona competente.

2.9. IA en las fases del ciclo científico

El ciclo científico no es una línea recta, sino una secuencia iterativa con puntos de control. Una hipótesis puede refinarse al revisar evidencia; un piloto puede revelar fallas de medición; una validación externa puede obligar a reconsiderar el dominio de uso. La IA aporta más cuando cada uso está vinculado a una tarea definida, una entrada documentada y una comprobación independiente.

Fase	Uso razonable	Evidencia de control
Observación y problema	Organizar notas, generar preguntas alternativas.	Versión humana inicial y registro de decisiones.
Literatura	Expandir términos, priorizar cribado, extraer candidatos.	Estrategias guardadas y cotejo con textos originales.
Diseño	Simular escenarios, revisar checklist, detectar omisiones.	Aprobación metodológica, ética y clínica.
Captura de datos	Reconocimiento de imagen o texto, alertas de calidad.	Muestra auditada, error por subgrupo y trazabilidad.
Análisis	Código preliminar, comparación de modelos, visualización.	Pruebas, revisión de código, partición correcta y reproducibilidad.
Interpretación	Explorar explicaciones alternativas.	Contraste con estimadores, incertidumbre y plausibilidad biológica.
Redacción	Editar claridad, resumir borradores propios, traducir.	Verificación de citas, declaración de uso y autoría humana.
Publicación/transferencia	Preparar materiales y formatos accesibles.	Revisión final, privacidad y comunicación de límites.

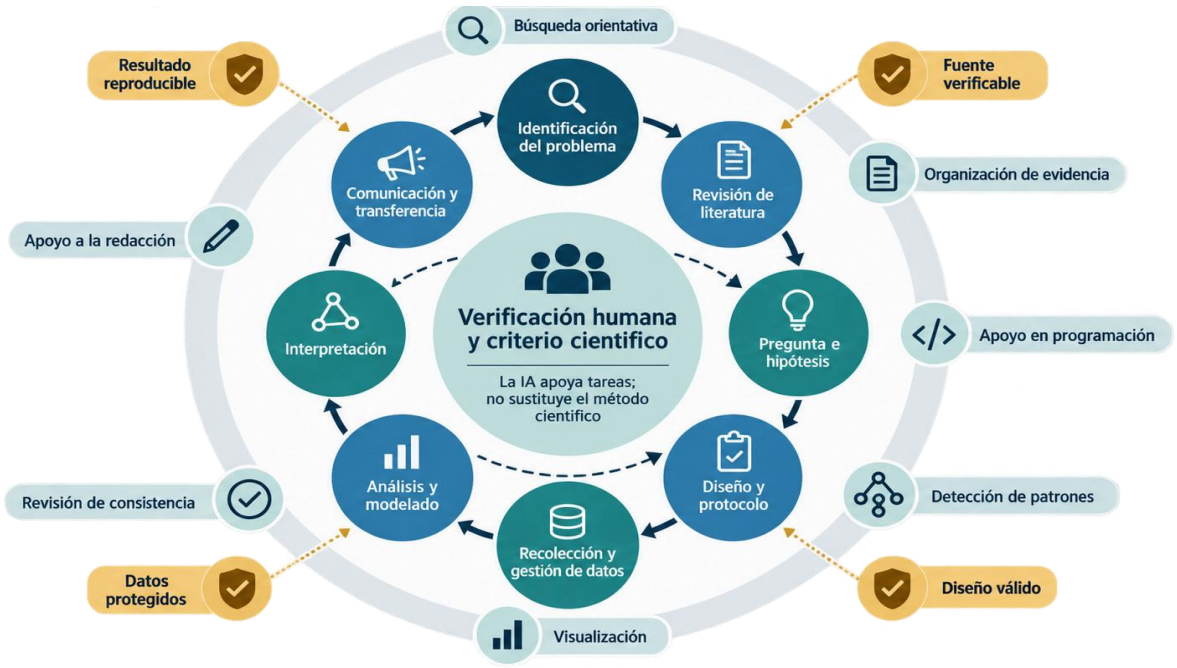
Gobernanza del uso de IA

La Organización Mundial de la Salud propone principios de autonomía, bienestar, seguridad, transparencia, responsabilidad, inclusión y sostenibilidad para IA en salud (World Health Organization, 2021). Su guía sobre modelos multimodales añade recomendaciones para investigación, desarrollo y despliegue responsable (World Health Organization, 2024). Aunque estos documentos se centran en salud humana, sus principios son pertinentes para información veterinaria que involucra tutores, productores, trabajadores, zoonosis y decisiones clínicas.

La gobernanza mínima incluye: clasificación de datos según sensibilidad; autorización para usos secundarios; anonimización o seudonimización; control de acceso; registro de herramientas y versiones; prohibición de cargar historias identificables en servicios no aprobados; auditoría de errores; conservación de código y semillas; y una declaración transparente del uso de IA. El equipo investigador, no el proveedor tecnológico, responde por el protocolo y el manuscrito.

Figura 2.2.

La inteligencia artificial en el ciclo científico



2.10. Riesgos de automatizar decisiones metodológicas

La automatización puede producir eficiencia aparente y desplazar el juicio sin que el equipo lo note. El sesgo de automatización ocurre cuando una recomendación computacional recibe más confianza de la que merece. Un texto fluido, una gráfica limpia o una métrica alta pueden ocultar premisas falsas. El riesgo aumenta cuando el usuario no dispone de conocimientos para formular preguntas de comprobación.

Riesgo	Manifestación	Medida preventiva
Alucinación	Referencia, resultado o instrumento inexistente.	Abrir la fuente original y verificar autor, año, DOI y afirmación.
Sesgo de automatización	Aceptar el diseño sugerido sin alternativas.	Revisión independiente y justificación escrita de decisiones.
Fuga de información	Variables posteriores al desenlace entran al modelo.	Definir el momento de predicción y auditar la canalización.
Pseudorreplicación	Imágenes o mediciones se tratan como animales independientes.	Partir y analizar por unidad experimental; modelar correlación.
Sobreajuste	Excelente rendimiento interno y fracaso externo.	Regularización, validación anidada y evaluación externa.
Sesgo de espectro	Muestra de casos extremos no representa la clínica.	Incluir el continuo de presentación y describir elegibilidad.
Confidencialidad	Carga de historias, imágenes o ubicación identificable.	Minimizar datos y usar entornos autorizados.
Deriva	Cambian población, protocolo, equipo o prevalencia.	Monitoreo temporal, recalibración y criterios de retirada.
Opacidad	No se conoce versión, prompt o preprocesamiento.	Bitácora, control de versiones y documentación reproducible.
Sustitución del objetivo	Se optimiza AUC y se olvida utilidad clínica.	Definir decisión, costos de error y beneficio antes de modelar.

Principio de proporcionalidad metodológica

La complejidad debe responder al problema y al volumen de información. Con 44 perros en un ensayo de dos grupos, un modelo mixto preespecificado puede estimar mejor el efecto que una batería de algoritmos. Con cientos de miles de imágenes, una red neuronal puede ser apropiada,

pero exige independencia entre particiones, etiquetas fiables y validación multicéntrica. El criterio no es modernidad, sino adecuación.

Tampoco debe automatizarse la decisión de excluir observaciones, imputar datos o transformar variables sin inspección. Cada regla afecta el estimando. Una herramienta puede señalar valores improbables, pero el investigador debe distinguir error de captura, valor biológicamente posible y evento adverso. Eliminar un valor extremo verdadero porque perjudica el ajuste empobrece la evidencia.

Lista de detención antes de aceptar una recomendación de IA

¿La recomendación responde al objetivo declarado? ¿Respeta la unidad experimental? ¿Usa información disponible en el momento previsto? ¿Sus supuestos pueden defenderse? ¿La elección sería la misma sin conocer el resultado? ¿Existe una alternativa más simple? ¿Puede reproducirse? ¿Quién revisó la decisión y con qué evidencia?

2.11. Caso didáctico: formulación de un estudio sobre dolor posoperatorio canino

Naturaleza del caso

Caso didáctico simulado. Los datos, tamaños y resultados esperados se incluyen con fines formativos; no corresponden a un ensayo real ni constituyen una recomendación terapéutica.

Todo protocolo clínico debe ajustarse a la legislación local, a la disponibilidad de medicamentos, al comité de ética y al juicio de un profesional veterinario responsable.

2.11.1. Observación y delimitación

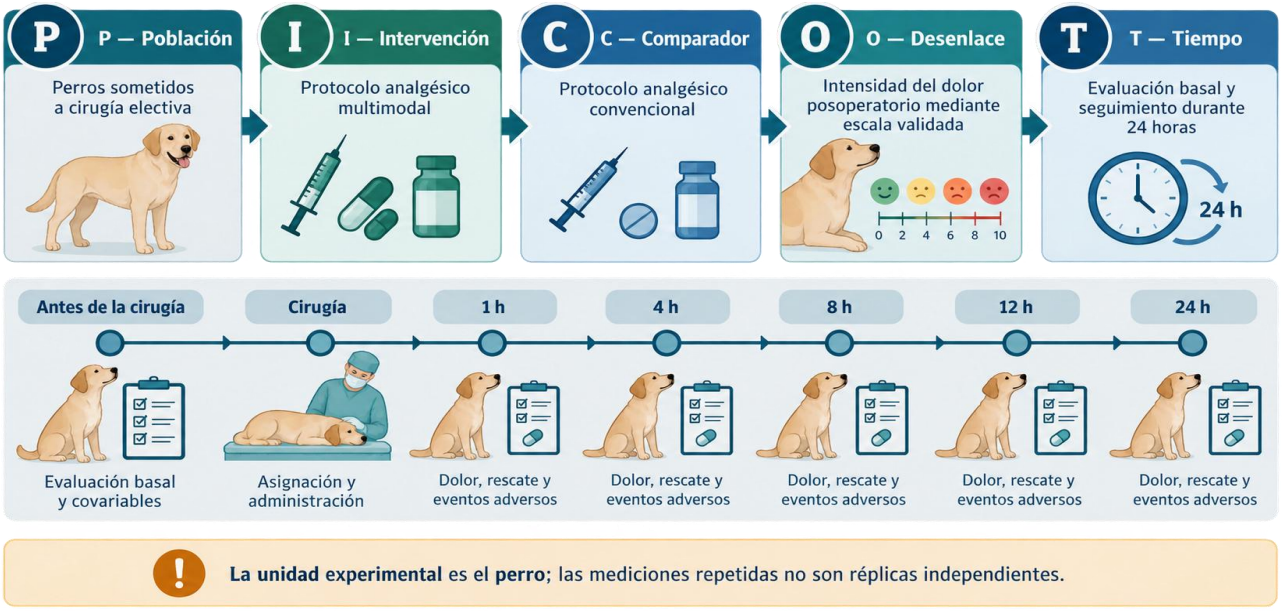
En dos clínicas docentes se ha observado que algunas perras sometidas a ovariohisterectomía electiva reciben analgesia de rescate durante las primeras 24 horas pese a un protocolo preventivo estándar. Los registros históricos no permiten comparar tratamientos porque la escala, los horarios y los criterios de rescate variaron. El problema no es demostrar que existe dolor, sino estimar si una estrategia multimodal adicional reduce su intensidad sin aumentar eventos adversos en una población y procedimiento definidos.

2.11.2. Pregunta PICOT

En perras adultas clínicamente sanas sometidas a ovariohisterectomía electiva en dos clínicas docentes (P), ¿la adición de un analgésico coadyuvante al protocolo estándar aprobado (I), comparada con el protocolo estándar más placebo o control equivalente (C), reduce el área bajo la curva de la puntuación CMPS-SF durante las primeras 24 horas (O, T), manteniendo criterios predefinidos de rescate y vigilancia de eventos adversos?

Figura 2.3.

Estructura PICOT del estudio sobre el dolor postoperatorio canino



2.11.3. Objetivos e hipótesis

- Objetivo principal: estimar la diferencia entre grupos en el área bajo la curva de dolor CMPS-SF desde 2 hasta 24 horas después de la extubación.
- Objetivos secundarios: comparar la proporción y el tiempo hasta analgesia de rescate, describir eventos adversos y estimar la recuperación funcional temprana.
- Objetivo exploratorio: evaluar si variables basales predicen rescate analgésico. Este objetivo no justificará un modelo complejo si el número de eventos es insuficiente.

- Hipótesis científica: la estrategia adicional produce una reducción clínicamente relevante de la trayectoria de dolor sin una frecuencia inaceptable de eventos adversos.
- Hipótesis estadística principal: la diferencia media ajustada en el desenlace primario entre grupos es igual a cero frente a la alternativa bilateral de diferencia distinta de cero.

2.11.4. Diseño y unidades

Se propone un ensayo clínico aleatorizado, controlado, con evaluación enmascarada y dos grupos paralelos. La unidad experimental es cada perra, porque el tratamiento se asigna al animal. Las puntuaciones en distintos horarios son unidades observacionales repetidas, no réplicas independientes. Las clínicas constituyen contextos o centros; si el número de centros es solo dos, su efecto debe interpretarse con prudencia.

La secuencia de aleatorización se generará antes del reclutamiento, con bloques de tamaño variable y estratificación por centro. La ocultación de la asignación se mantendrá mediante sobres opacos numerados o un sistema electrónico independiente. El evaluador del dolor, el responsable del análisis y, cuando sea factible, el equipo clínico permanecerá enmascarados. Si las intervenciones no pueden igualarse visualmente, se documentará qué roles conservaron el cegamiento.

2.11.5. Población, criterios y muestra

Elemento	**Especificación simulada**
Población objetivo	Perras adultas sanas candidatas a ovariohisterectomía electiva.
Inclusión	Estado físico anestésico compatible, consentimiento del tutor y evaluación basal completa.
Exclusión	Dolor preexistente, enfermedad sistémica relevante, medicación analgésica reciente o conversión quirúrgica mayor.
Retiro	Decisión clínica por seguridad; el retiro del tratamiento no implica eliminar datos ya autorizados.
Tamaño ilustrativo	Si se espera una diferencia de 1,2 puntos y DE de 1,3, una aproximación bilateral con α=0,05 y potencia 80% requiere cerca de 19 animales por grupo; con 10% adicional, 22 por grupo.
Advertencia	El cálculo definitivo debe basarse en el desenlace primario exacto, correlación repetida, diferencia clínica y pérdidas justificadas; no en una cifra propuesta por IA.

2.11.6. Variables y medición

El desenlace principal será el área bajo la curva de CMPS-SF entre 2, 4, 6, 8, 12 y 24 horas. La Glasgow Composite Measure Pain Scale-Short Form es una escala conductual empleada en perros; el protocolo definirá entrenamiento, condiciones de observación, orden de examen y umbral de rescate. Las guías recientes de AAHA y WSAVA sostienen la necesidad de evaluación repetida y tratamiento individualizado del dolor (Gruen et al., 2022; Monteiro et al., 2023).

Las variables basales incluirán edad, peso, condición corporal, temperamento, puntuación de dolor basal y centro. Se documentarán duración de anestesia y cirugía, técnica, experiencia del cirujano, fármacos perioperatorios, eventos adversos, sedación y hora de rescate. Sedación y dolor pueden confundirse conductualmente; por ello, una escala de sedación y reglas de interpretación se incorporarán a priori.

2.11.7. Bienestar, rescate y consentimiento

Ningún objetivo metodológico justifica mantener dolor no tratado. El protocolo fijará un umbral de rescate, permitirá rescate por juicio clínico aun cuando la puntuación no lo alcance y registrará fármaco, dosis y tiempo. Un monitor independiente revisará eventos graves. El tutor recibirá información sobre aleatorización, beneficios inciertos, riesgos, confidencialidad y derecho a retirar al animal sin afectar su atención.

2.11.8. Plan estadístico

- Análisis principal por intención de tratar, preservando la asignación inicial.
- Estimación de la diferencia entre grupos en el área bajo la curva, con intervalo de confianza del 95% y tamaño del efecto.
- Modelo mixto complementario para puntuaciones repetidas, con efectos de tratamiento, tiempo e interacción, e intercepto por animal.
- Proporción de rescate mediante riesgo relativo o diferencia de riesgos; tiempo al rescate con curvas de Kaplan-Meier y modelo apropiado si se cumplen supuestos.
- Descripción separada de eventos adversos; no inferir “seguridad equivalente” a partir de ausencia de significación en una muestra pequeña.
- Análisis de sensibilidad para datos faltantes y desviaciones; razones de pérdida reportadas por grupo.

- Interpretación conjunta de intervalo, diferencia clínica, bienestar y consistencia temporal, no únicamente de $p < 0,05$.

2.11.9. Papel limitado y auditable de la IA

Tarea	Uso permitido	Validación
Literatura	Sugerir términos y priorizar registros.	Revisión en bases, deduplicación y cribado humano.
Protocolo	Detectar campos ausentes frente a listas de comprobación.	Decisión clínica, estadística y ética documentada.
Captura	Señalar datos fuera de rango.	Verificación con fuente y prohibición de borrado automático.
Código	Proponer funciones para limpieza y modelos.	Pruebas con datos simulados, revisión y control de versiones.
Predicción exploratoria	Comparar un modelo parsimonioso con alternativas.	Validación anidada; suspender si eventos/muestra son insuficientes.
Redacción	Mejorar claridad de un borrador propio.	Comprobación de cada número y referencia; declaración de uso.

2.11.10. Resultado interpretativo simulado

Supóngase que la diferencia estimada en el área bajo la curva favorece al grupo coadyuvante, pero el intervalo de confianza incluye desde un beneficio pequeño hasta uno clínicamente importante. La conclusión correcta sería que los datos son compatibles con beneficio, aunque la precisión no permite fijar su magnitud con seguridad. No correspondería afirmar eficacia definitiva solo porque un valor p cruzó 0,05. Si, además, la proporción de rescate disminuye de forma coherente y no surge una señal adversa, el conjunto de resultados respaldaría un ensayo mayor o una adopción cautelosa con vigilancia, según el contexto.

Si un clasificador exploratorio predice rescate con AUC alta dentro de la misma muestra, su resultado no debe presentarse como herramienta clínica. Con pocos eventos, la incertidumbre será amplia y el sobreajuste probable. El hallazgo útil puede ser simplemente que la predicción necesita más datos multicéntricos. Saber cuándo no desplegar IA también es una conclusión científica.

2.12. Síntesis, autoevaluación y actividad práctica

Síntesis

La IA se integra de manera legítima al método científico cuando amplía tareas concretas sin desplazar la responsabilidad del investigador. La calidad del estudio sigue dependiendo de un problema relevante, una pregunta precisa, un objetivo coherente, un diseño capaz de sostener la inferencia y una medición válida. Explicación, estimación, clasificación y predicción producen conocimientos distintos; mezclarlos lleva a interpretar asociaciones como causas o rendimiento interno como utilidad clínica.

En la revisión de literatura, la IA puede expandir términos, deduplicar y priorizar el cribado, pero los criterios, las decisiones de inclusión y la extracción deben ser auditables. En el protocolo, puede señalar omisiones y asistir con código, mientras que unidad experimental, tamaño muestral, desenlace, rescate y análisis requieren juicio especializado. La transparencia acerca de herramientas, datos, prompts, versiones y verificaciones forma parte de la reproducibilidad.

Autoevaluación

1. ¿Cuál es la diferencia entre un tema y un problema de investigación veterinaria?
2. ¿Por qué una base de datos disponible no basta para justificar una pregunta científica?
3. Formule una pregunta PECO sobre parasitismo gastrointestinal en ovinos.
4. ¿Qué diferencia existe entre una hipótesis científica y una hipótesis predictiva?
5. ¿Por qué la importancia de una variable en random forest no demuestra causalidad?
6. ¿Qué riesgo se produce al dividir imágenes del mismo animal entre entrenamiento y prueba?
7. Mencione cuatro elementos que deben registrarse al usar IA para cribado bibliográfico.
8. ¿Cómo se distingue un vacío poblacional de un vacío de validación?
9. ¿Qué debe contener un plan de análisis además del nombre de una prueba?
10. ¿Por qué la ausencia de eventos adversos significativos no demuestra equivalencia de seguridad?
11. En el caso didáctico, ¿cuál es la unidad experimental y cuáles son las unidades observacionales?
12. ¿En qué circunstancias sería metodológicamente preferible no entrenar un modelo de aprendizaje automático?

Orientaciones de respuesta

1. El tema delimita un campo; el problema articula una incertidumbre específica, relevante y comprobable dentro de una población y contexto.

2. Porque los datos pueden no representar a la población, no medir el constructo o no sostener el objetivo causal o predictivo.

3. Debe incluir población ovina, exposición, comparador y desenlace medible, además del periodo y contexto cuando sean relevantes.

4. La científica propone una relación sustantiva; la predictiva anticipa desempeño en casos nuevos y requiere validación fuera del ajuste.

5. Porque la importancia mide contribución al desempeño condicionada por los datos y el algoritmo, no un efecto de intervención libre de confusión.

6. Fuga de información y estimación optimista, pues el modelo reconoce patrones específicos del paciente ya visto.

7. Herramienta/versión, criterios, decisiones de entrenamiento, regla de detención y auditoría de registros excluidos.

8. El primero afecta la representación de grupos; el segundo aparece cuando el desempeño no se ha probado en datos independientes.

9. Estimando, covariables, supuestos, datos faltantes, multiplicidad, sensibilidad, intervalos y visualización.

10. Porque una muestra pequeña puede carecer de potencia para detectar daños y “no significativo” no equivale a ausencia ni equivalencia.

11. La perra es la unidad experimental; las puntuaciones repetidas son observaciones correlacionadas dentro del animal.

12. Cuando la muestra es pequeña, el objetivo es una estimación interpretable, faltan eventos o un modelo simple responde mejor a la pregunta.

Actividad práctica integradora

Seleccione uno de estos contextos: mastitis subclínica bovina, clasificación radiográfica canina, riesgo de mortalidad neonatal en lechones, resistencia antiparasitaria en ovinos o recuperación posquirúrgica felina. Prepare una ficha metodológica de cuatro páginas siguiendo los pasos siguientes.

1. Describa la observación inicial y delimite el problema con población, contexto, periodo, consecuencia, conocimiento y vacío.

2. Formule cuatro preguntas: estimación, explicación/causalidad, clasificación y predicción. Elija una y justifique por qué.

3. Defina objetivo general, tres objetivos específicos, hipótesis cuando corresponda y desenlace primario.

4. Diseñe una búsqueda en dos bases. Registre conceptos, términos libres, descriptores, operadores, fecha y resultados.

5. Construya una matriz con cinco fuentes verificadas publicadas preferentemente desde 2021.

6. Especifique diseño, unidad experimental, muestra, variables, control de sesgo, análisis y limitaciones.

7. Identifique dos tareas en las que usaría IA y dos decisiones que no delegaría. Incluya el prompt, la salida resumida y la comprobación realizada.

8. Prepare una declaración breve de uso de IA y un plan de confidencialidad / reproducibilidad.

Rúbrica sugerida

Criterio	Puntaje	Desempeño esperado
Problema y relevancia	20	Delimitación específica, sustentada y vinculada a una decisión veterinaria.
Coherencia metodológica	25	Pregunta, objetivo, diseño, unidad, variables y análisis son congruentes.

Literatura y vacío	20	Búsqueda reproducible, fuentes verificadas y vacío defendible.
Uso responsable de IA	20	Tareas, límites, prompts, verificación, privacidad y trazabilidad explícitos.
Argumentación y APA	15	Escritura propia, crítica, precisa y referencias consistentes.
Total	100	Producto metodológicamente justificable y reproducible.

CAPÍTULO 3

ÉTICA INTEGRIDAD CIENTÍFICA Y GOBERNANZA DE DATOS VETERINARIOS

Introducción

La investigación veterinaria asistida por inteligencia artificial no se vuelve éticamente aceptable por el solo hecho de trabajar con datos ya existentes. Una radiografía archivada, una grabación de conducta, una secuencia genómica, un registro de producción o una historia clínica electrónica conservan relaciones con animales, personas, establecimientos y comunidades. De esas relaciones nacen obligaciones: evitar sufrimiento innecesario, respetar las decisiones del tutor, proteger identidades, justificar la finalidad del análisis, documentar el origen de los datos y responder por los efectos de un modelo cuando se incorpora a una decisión clínica o sanitaria.

La gobernanza ética comienza antes del entrenamiento del algoritmo. Se concreta al formular la pregunta, elegir la población, definir qué datos son indispensables, obtener autorizaciones, establecer controles de acceso y anticipar usos secundarios. Continúa durante el análisis mediante trazabilidad, evaluación de sesgos, validación por subgrupos y documentación de cambios. No termina con la publicación: los conjuntos de datos, el código, las versiones del modelo, las incidencias y los criterios de retiro deben conservarse durante el periodo definido por la institución y la normativa aplicable.

Este capítulo adopta un enfoque de riesgo. No todas las aplicaciones exigen el mismo nivel de control. Un modelo exploratorio para clasificar imágenes en una práctica de aula no equivale a un sistema que prioriza pacientes con disnea, recomienda eutanasia, selecciona animales para sacrificio sanitario o asigna recursos durante un brote. A mayor posibilidad de daño, irreversibilidad, escala y opacidad, mayor debe ser la evidencia, la supervisión humana y la capacidad de auditoría. La responsabilidad profesional no se transfiere al proveedor del software ni a la herramienta generativa.

Objetivos de aprendizaje

- Aplicar principios éticos y las 3R a estudios experimentales, observacionales y computacionales en Ciencias Veterinarias.
- Distinguir consentimiento para la atención clínica, autorización para investigación y permiso para usos secundarios de datos o muestras.
- Diseñar un esquema de gobernanza que defina finalidad, roles, acceso, seguridad, conservación, reutilización y eliminación de datos veterinarios.

- Reconocer y medir sesgos algorítmicos entre especies, razas, sistemas productivos, regiones y centros de atención.
- Documentar el desarrollo, validación, uso y monitoreo de un sistema de IA con criterios de transparencia y rendición de cuentas.
- Diferenciar apoyo automatizado, contribución intelectual humana y autoría; declarar de forma verificable el uso de herramientas de IA.
- Evaluar riesgos de la IA generativa para confidencialidad, exactitud, referencias, código, imágenes y reproducibilidad científica.

3.1 Principios éticos aplicados a la investigación con animales

La ética de la investigación veterinaria tiene al menos tres sujetos de consideración: el animal, cuya integridad y bienestar pueden verse afectados; la persona o comunidad vinculada con él, que aporta decisiones, información o recursos; y la sociedad, que espera conocimiento válido, uso prudente de fondos y protección frente a riesgos sanitarios. A ellos se suman los profesionales responsables y las instituciones custodias. Esta configuración impide trasladar sin ajustes un esquema de ética clínica humana o reducir el análisis a una autorización administrativa. La literatura veterinaria reciente sitúa precisamente la responsabilidad, el bienestar, la privacidad y la justicia entre los problemas centrales de la adopción de IA (Appleby & Basran, 2022; Coghlan & Quinn, 2023).

Una evaluación sólida combina valor científico, validez metodológica, proporcionalidad entre daños y beneficios, selección justa, cuidado competente, transparencia y revisión independiente. El valor científico exige que la pregunta sea relevante para la salud, el bienestar, la producción sostenible o la salud pública veterinaria. La validez obliga a utilizar un diseño capaz de responderla. Un estudio subpotenciado, sin aleatorización cuando corresponde o con pseudorreplicación puede exponer animales y consumir muestras sin producir inferencias confiables; por ello, el defecto metodológico también es un problema ético.

La proporcionalidad no se resuelve afirmando que el beneficio potencial es importante. Deben describirse la naturaleza, intensidad y duración de los daños previsibles, las medidas de refinamiento, los puntos finales humanitarios y la incertidumbre del beneficio. En un ensayo de analgesia posoperatoria canina, por ejemplo, no sería aceptable retirar un tratamiento de rescate

para ampliar la separación estadística entre grupos. La medicación de rescate debe definirse de antemano, registrarse como desenlace y activarse con criterios clínicos objetivos.

La revisión ética independiente examina conflictos que el equipo puede subestimar por familiaridad con el procedimiento o interés en el resultado. La WOAH recomienda sistemas institucionales de supervisión, registros de intervenciones y responsabilidades compartidas entre científicos, veterinarios y personal de cuidado (World Organisation for Animal Health [WOAH], 2024). La aprobación no es una licencia general: cualquier desviación sustantiva, evento adverso inesperado o cambio de finalidad requiere el procedimiento de notificación o enmienda establecido por la institución.

Principio	Pregunta de control	Ejemplo veterinario
Valor científico	¿El conocimiento esperado justifica el uso de animales o datos?	Validar un marcador temprano de mastitis antes de ampliar el muestreo.
Validez	¿El diseño puede responder la pregunta sin pseudorreplicación?	Considerar al hato, y no cada lectura del mismo sensor, cuando el tratamiento se asigna por finca.
Proporcionalidad	¿Se minimizaron dolor, estrés, privación y riesgo?	Usar analgesia de rescate y puntos finales humanitarios en un ensayo posquirúrgico.
Justicia	¿La selección distribuye cargas y beneficios de modo defendible?	No reclutar solo animales de refugios por conveniencia si la aplicación será comercial.
Responsabilidad	¿Hay una persona competente que pueda detener el procedimiento?	El veterinario responsable suspende la intervención ante deterioro clínico.

3.2 Bienestar animal y principio de las 3R

Las 3R —reemplazo, reducción y refinamiento— constituyen un marco mínimo para planificar investigación con animales. Reemplazar significa evitar o sustituir el uso de animales vivos cuando una pregunta pueda resolverse mediante cultivos, organoides, simulación, bases de datos, muestras archivadas o métodos no animales. Reducir no significa usar el menor número posible a cualquier costo, sino obtener información válida con el número mínimo compatible con precisión, potencia y variabilidad biológica. Refinar consiste en modificar procedimientos, alojamiento,

manipulación, anestesia, analgesia, monitoreo y puntos finales para disminuir dolor, miedo, estrés o daño duradero.

La IA puede contribuir a las 3R, pero su aporte debe demostrarse. Un modelo que extrae fenotipos de historias clínicas puede reemplazar una nueva cohorte prospectiva; visión computacional puede detectar alteraciones de marcha sin instrumentación invasiva; un diseño adaptativo puede reducir asignaciones a opciones inferiores. También puede ocurrir lo contrario: el entusiasmo por generar grandes conjuntos de imágenes puede multiplicar procedimientos, y un sistema de monitoreo continuo puede normalizar condiciones de alojamiento deficientes en lugar de corregirlas.

La reducción depende del diseño estadístico. El cálculo de tamaño muestral debe declarar desenlace primario, diferencia biológicamente relevante, variabilidad esperada, alfa, potencia y pérdidas. En modelos predictivos, el tamaño no se justifica solo por una regla universal de "eventos por variable"; deben considerarse prevalencia del desenlace, complejidad, sobreajuste, calibración y disponibilidad de una validación externa. Reutilizar miles de registros no elimina el sesgo si todos proceden de una única clínica de referencia.

Figura 3.1.

Aplicación del principio de las 3R en proyectos veterinarios con IA



PREPARE orienta la planificación y el diálogo entre investigadores y personal de las instalaciones, mientras ARRIVE 2.0 mejora el reporte de estudios in vivo (Smith et al., 2018; Percie du Sert et al., 2020). Son instrumentos complementarios: uno no sustituye la evaluación ética ni el otro corrige un protocolo deficiente. Deben integrarse desde la planificación, no añadirse como lista de comprobación al redactar el artículo.

R	Aplicación adecuada de IA	Riesgo de uso superficial
Reemplazo	Entrenar con imágenes diagnósticas obtenidas durante atención indicada clínicamente.	Provocar nuevas lesiones o repetir estudios solo para aumentar el conjunto de entrenamiento.
Reducción	Simular escenarios y usar diseños eficientes antes de la fase animal.	Afirmar que “big data” justifica cualquier tamaño sin evaluar independencia y calidad.
Refinamiento	Detectar tempranamente dolor o cambios conductuales con sensores validados.	Usar alertas no validadas que retrasen la observación clínica o generen fatiga de alarmas.

Ejemplo aplicado. En un estudio de cicatrización posquirúrgica tras ovariohisterectomía, la cámara no debe imponer manipulaciones adicionales solo para estandarizar imágenes. Puede definirse una distancia fija, iluminación difusa y una única toma durante el control clínico programado. La unidad experimental sigue siendo la paciente; múltiples fotografías por herida son mediciones repetidas y no réplicas independientes. El refinamiento incluye analgesia, criterios de infección, retiro y tratamiento de rescate.

3.3 Consentimiento informado de tutores y propietarios

En animales con tutor, el consentimiento es un proceso de comunicación y decisión, no una firma aislada. La persona debe comprender el objetivo, los procedimientos adicionales a la atención habitual, la asignación de tratamientos, los riesgos previsibles, los beneficios directos inciertos, las alternativas clínicas, los costos, la compensación, la atención de eventos adversos, el uso de datos y muestras, y el derecho a retirarse sin perder la atención ordinaria. Frederick (2024) subraya la necesidad de documentos comprensibles y tiempo suficiente para decidir en ensayos clínicos veterinarios.

Conviene separar tres autorizaciones que suelen confundirse: consentimiento para diagnosticar o tratar; participación del animal en investigación; y reutilización futura de datos, imágenes o

muestras. Un tutor puede aceptar la cirugía y rechazar que las fotografías de la herida entrenen un algoritmo. También puede aceptar el proyecto actual y no un repositorio abierto. Las opciones deben presentarse sin casillas premarcadas ni lenguaje que haga parecer obligatoria la investigación.

La capacidad de decisión del propietario puede verse afectada por urgencia, dependencia del servicio, barreras lingüísticas, baja alfabetización o expectativa de tratamiento gratuito. El equipo debe evitar influencia indebida, explicar incertidumbre y distinguir claramente investigación de atención individual. Cuando la intervención experimental ofrece un posible beneficio pero existe tratamiento estándar, ambas vías deben describirse con equilibrio. La gratuidad no elimina el riesgo de coerción percibida.

Para datos retrospectivos, un comité puede autorizar una dispensa de consentimiento si la normativa y la política institucional lo permiten, el riesgo es mínimo, contactar a todos resulta impracticable y existen salvaguardas robustas. “Retrospectivo” no equivale automáticamente a “exento”. La base jurídica, la decisión del comité y los límites de reutilización deben quedar documentados.

Elemento	Contenido mínimo
Finalidad	Pregunta del estudio y diferencia frente a la atención clínica.
Procedimientos	Qué se hará, frecuencia, duración, aleatorización y uso de IA.
Riesgos y rescate	Dolor, estrés, falsos positivos, incidentes de privacidad y atención disponible.
Datos y muestras	Qué se recoge, quién accede, cuánto se conserva y si habrá reutilización.
Voluntariedad	Posibilidad de rechazar o retirarse y consecuencias reales del retiro.
Contacto	Responsable clínico, investigador y canal para dudas o reclamos.

3.4 Privacidad y confidencialidad de historias clínicas

Aunque el paciente animal no sea titular de derechos de protección de datos personales en muchos ordenamientos, la historia veterinaria contiene nombres, teléfonos, direcciones, localización de fincas, información económica, fotografías del hogar, matrículas, conversaciones y datos de trabajadores. Esa información puede identificar directa o indirectamente a una persona. En Ecuador, la Ley Orgánica de Protección de Datos Personales regula el tratamiento de información asociada a personas naturales identificadas o identificables; su Reglamento General fue expedido

mediante Decreto Ejecutivo 904 de 2023 (Asamblea Nacional del Ecuador, 2021; Presidencia de la República del Ecuador, 2023). La institución debe verificar además normas sectoriales, contratos y políticas vigentes.

Confidencialidad describe el deber de no revelar información a personas no autorizadas. Privacidad se refiere al control legítimo sobre el tratamiento de información personal. Seguridad comprende medidas técnicas y organizativas para preservar confidencialidad, integridad y disponibilidad. Anonimización pretende impedir razonablemente la reidentificación; seudonimización sustituye identificadores por códigos, pero conserva una clave o posibilidad de enlace y, por tanto, no convierte automáticamente los datos en anónimos.

Eliminar el nombre no basta. Una combinación de especie rara, diagnóstico, fecha exacta, barrio, fotografía y número de microchip puede revelar la identidad del tutor o del establecimiento. En producción animal, coordenadas, tamaño del hato y brotes pueden identificar una finca y producir pérdidas comerciales o estigma sanitario. Antes de compartir debe evaluarse el riesgo de singularidad, enlace con fuentes públicas y divulgación por texto libre o metadatos de imágenes.

Nunca deben copiarse historias clínicas completas en un chatbot público o cuenta personal sin autorización, acuerdo de tratamiento, evaluación de seguridad y controles de retención. La interfaz puede registrar indicaciones, archivos y respuestas; la ausencia visible de un nombre no elimina otros identificadores. Para análisis con modelos generativos se prefieren entornos institucionales aprobados, minimización de campos, seudonimización local y contratos que impidan usos incompatibles.

Capa	Control recomendado	Evidencia de cumplimiento
Datos	Minimización, seudonimización, generalización de fechas y revisión de texto libre.	Diccionario de variables y registro de transformaciones.
Acceso	Mínimo privilegio, autenticación multifactor y revisión periódica.	Matriz de roles y bitácoras.

Infraestructura	Cifrado en tránsito y reposo, copias verificadas y gestión de vulnerabilidades.	Configuración, prueba de restauración e informe de incidentes.
Personas	Acuerdos de confidencialidad y capacitación basada en casos.	Registros de formación y compromisos firmados.
Ciclo de vida	Plazos de conservación y eliminación verificable.	Calendario de retención y actas de eliminación.

3.5 Propiedad custodia y reutilización de datos

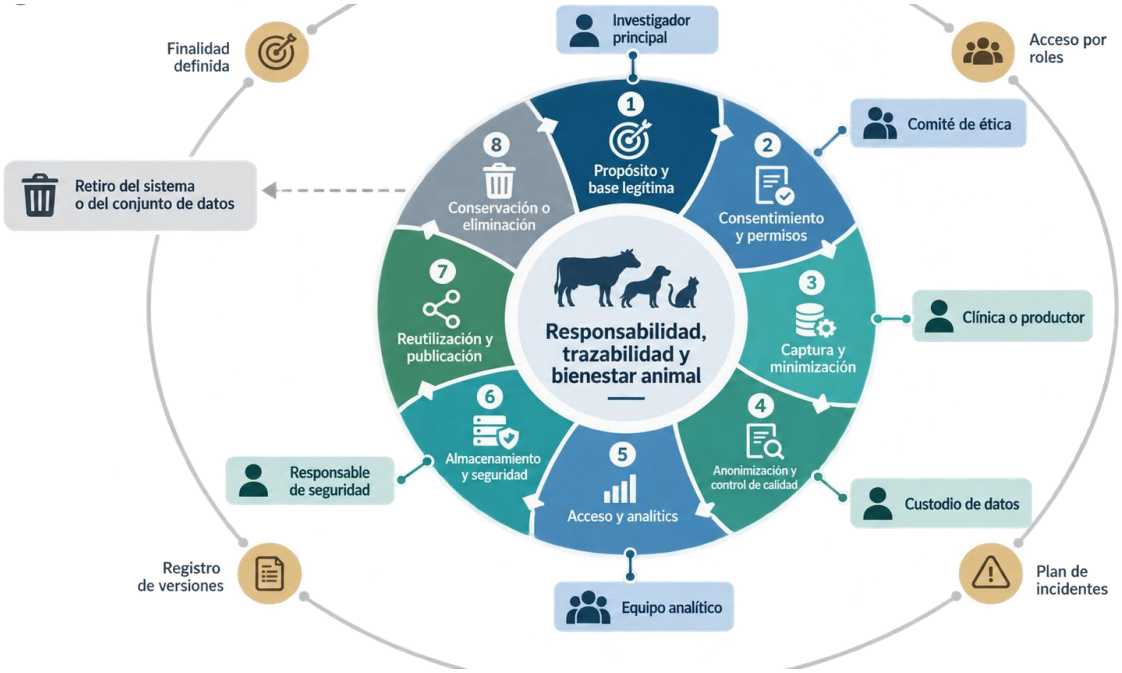
La pregunta "¿de quién son los datos?" rara vez admite una respuesta única. Pueden coexistir derechos y obligaciones del tutor, la clínica que generó el expediente, la universidad, el financiador, el laboratorio, el proveedor del sistema, el investigador y la comunidad de origen. Además, la propiedad material del archivo no equivale a libertad para reutilizar su contenido. Un contrato puede reconocer custodia institucional y, al mismo tiempo, limitar finalidades, transferencias, sublicencias o explotación comercial.

La gobernanza debe asignar roles explícitos. El responsable o controlador decide finalidades y medios esenciales; el encargado o procesador trata datos por cuenta del responsable; el custodio administra conservación, integridad y acceso; el investigador autorizado utiliza solo lo aprobado. Las denominaciones jurídicas varían, pero la separación funcional evita que una cuenta con acceso técnico se interprete como permiso científico irrestricto.

Los principios FAIR proponen datos localizables, accesibles, interoperables y reutilizables (Wilkinson et al., 2016). "Accesibles" no significa necesariamente abiertos. Un conjunto sensible puede ser FAIR mediante metadatos públicos y acceso controlado bajo criterios transparentes. En veterinaria, deben considerarse vocabularios de especie, raza, anatomía, diagnóstico, unidad de medida, sistema productivo y contexto geográfico. Sin metadatos, un código de "0" puede significar ausencia, desconocido, no evaluado o resultado negativo.

Figura 3.2.

Ciclo de gobernanza de datos veterinaries



La reutilización requiere compatibilidad de finalidad, base jurídica o consentimiento aplicable, evaluación ética, minimización y acuerdo de uso. Debe prohibirse la reidentificación, fijarse un periodo, describirse resultados permitidos, resolver publicaciones y propiedad intelectual, y exigir devolución o destrucción. Si un proveedor entrena su producto con los expedientes de la clínica, esa finalidad debe constar expresamente; no puede presumirse por el simple uso del software.

Rol	**Decisión o deber principal**	**Documento**
Responsable del tratamiento	Define finalidad, fundamento y destinatarios.	Protocolo y registro de actividades.
Custodio de datos	Preserva integridad, versiones, respaldos y acceso.	Plan de gestión de datos.
Investigador principal	Garantiza uso conforme al protocolo y reporte de incidentes.	Compromiso del equipo y bitácora.
Comité ético o instancia equivalente	Evalúa proporcionalidad, consentimiento y salvaguardas.	Dictamen y enmiendas.
Proveedor tecnológico	Opera bajo condiciones contractuales verificables.	Acuerdo de tratamiento, seguridad y nivel de servicio.

3.6 Sesgos algorítmicos y desigualdad entre poblaciones animales

Un algoritmo puede presentar buena discriminación promedio y perjudicar a subpoblaciones. El sesgo aparece cuando la selección, medición, etiquetado, modelado o despliegue produce errores sistemáticos que afectan de manera desigual. En veterinaria, las diferencias relevantes no se limitan a sexo y edad: incluyen especie, raza, conformación, color de pelaje, tamaño, estado reproductivo, sistema de manejo, altitud, clima, región, capacidad económica del tutor, tipo de clínica y disponibilidad diagnóstica.

Los datos de hospitales de referencia sobrerrepresentan casos complejos, propietarios con capacidad de pago y pruebas avanzadas. Un modelo de hospitalización entrenado allí puede sobreestimar riesgo en atención primaria o no reconocer patrones rurales. Las razas frecuentes dominan bancos de radiografías y fotografías; una CNN puede aprender marcadores del equipo, posición o institución en lugar de la lesión. Norori et al. (2021) relacionan la reducción de sesgos con transparencia de datos y ciencia abierta, mientras la literatura veterinaria advierte que la validación y la regulación siguen rezagadas frente a la oferta comercial (Bellamy, 2023; Heinlein, 2026).

La evaluación exige más que dividir al azar registros del mismo hospital. Deben evitarse fugas entre imágenes o visitas del mismo animal y separar por paciente, finca o centro según la unidad de dependencia. La validación externa temporal y geográfica revela cambios de prevalencia, equipos y prácticas. Deben informarse sensibilidad, especificidad, valores predictivos y calibración por subgrupo con intervalos de confianza. Una diferencia aparente en grupos pequeños puede ser imprecisa; por eso también se reporta el denominador y no solo el porcentaje.

Figura 3.3.

Cómo se origina y amplifica el sesgo algorítmico



No existe una métrica universal de equidad que resuelva conflictos normativos. Igualar sensibilidad entre razas puede modificar falsos positivos; igualar valores predictivos puede ser imposible si la prevalencia difiere. La decisión debe partir del daño clínico. En triaje de urgencias se puede priorizar sensibilidad para no omitir pacientes graves; en sacrificio sanitario un falso positivo tiene consecuencias irreversibles y exige confirmación independiente. PROBAST+AI y TRIPOD+AI ofrecen preguntas y pautas útiles para evaluar sesgo y transparencia de modelos predictivos, aunque su origen sea salud humana y deban adaptarse al contexto veterinario (Collins et al., 2024; Moons et al., 2025).

Fuente de sesgo	Manifestación veterinaria	Control
Selección	Solo hospitales urbanos o animales asegurados.	Muestreo multicéntrico y descripción del flujo de inclusión.
Medición	Equipos radiográficos y protocolos distintos.	Armonización, análisis por dispositivo y validación externa.
Etiqueta	Diagnóstico basado en opinión sin referencia uniforme.	Criterios explícitos, adjudicación y acuerdo entre evaluadores.
Fuga de datos	Imágenes del mismo animal en entrenamiento y prueba.	Separación por animal, finca o episodio antes de particionar.
Cambio de dominio	Modelo de razas pequeñas aplicado a animales de producción.	Restricción de uso, recalibración y nueva validación.
Despliegue	Usuarios confían en la alerta aunque contradiga el examen.	Capacitación, registro de discrepancias y autoridad para anular.

3.7 Transparencia y responsabilidad profesional

La transparencia útil permite comprender para qué fue construido el sistema, con qué población, qué entrada requiere, qué salida produce, cómo se validó y cuándo no debe usarse. No exige revelar cada secreto industrial ni convertir redes profundas en reglas simples. Sí exige información suficiente para que una clínica, comité o investigador evalúe pertinencia, límites y riesgos.

La documentación mínima incluye versión del modelo, fecha, procedencia y criterios de inclusión de los datos, definición del desenlace, manejo de faltantes, partición, métricas con intervalos, desempeño por subgrupos, calibración, validación externa, umbral, condiciones fuera de alcance, ciberseguridad, monitoreo y procedimiento de actualización. Basran y Appleby (2024) proponen una caja de herramientas para despliegue seguro que integra equipo multidisciplinario, calidad de datos, implementación, formación y evaluación continua. En imagenología, ACVR y ECVDI sostienen la necesidad de desarrollo y uso ético, transparente y clínicamente seguro (Appleby et al., 2025).

Responsabilidad significa poder atribuir decisiones y acciones. El desarrollador responde por el proceso de diseño y la evidencia proporcionada; la institución por selección, contratación, integración y vigilancia; el profesional por interpretar la salida dentro del caso concreto; y el investigador por el uso de datos, análisis y comunicación. Esta distribución no diluye responsabilidades. Debe existir un propietario operativo del sistema, un canal de incidentes y una autoridad con capacidad de suspenderlo.

La supervisión humana debe ser sustantiva. Mostrar una probabilidad junto a un botón de “aceptar” no constituye supervisión si el usuario carece de tiempo, información o facultad para contradecirla. Deben analizarse sesgo de automatización, fatiga de alertas y posible pérdida de habilidades. Cohen y Gordon (2022) recuerdan que, en radiología y oncología radioterápica veterinaria, el deber de no dañar exige considerar errores, opacidad, responsabilidad y comunicación con el cliente.

Momento	Pregunta de rendición de cuentas	Acción

Antes del uso	¿La población de validación se parece a la atendida?	Revisión clínica y técnica independiente.
Durante el uso	¿Quién confirma o anula la recomendación?	Registrar decisión humana y motivo.
Monitoreo	¿Cambió prevalencia, equipo o flujo de trabajo?	Vigilar calibración, errores y subgrupos.
Incidente	¿Hubo daño o casi incidente?	Contener, notificar, investigar y corregir.
Actualización	¿La nueva versión conserva desempeño?	Validar, versionar y comunicar cambios.
Retiro	¿El beneficio dejó de superar el riesgo?	Suspender y preservar evidencia de auditoría.

3.8 Autoría plagio e integridad científica

La integridad científica comprende honestidad en la generación, análisis y comunicación de evidencia; trazabilidad de decisiones; reconocimiento justo de contribuciones; conservación de datos; manejo de conflictos; y corrección del registro cuando aparece un error. Fabricar datos, falsificar imágenes o resultados y plagiar son faltas graves, pero la zona de riesgo es más amplia: selección oportunista de desenlaces, exclusiones no declaradas, hiperparámetros ajustados sobre el conjunto de prueba, duplicación de animales como observaciones independientes, supresión de resultados negativos y autoría por jerarquía también distorsionan la evidencia.

La autoría combina crédito y responsabilidad. Los cuatro criterios del ICMJE exigen contribución sustancial, participación en redacción o revisión intelectual, aprobación final y responsabilidad por la integridad del trabajo (International Committee of Medical Journal Editors [ICMJE], 2026). Quien solo consigue financiamiento, facilita acceso a historias o realiza una corrección lingüística debe reconocerse según su contribución, pero no necesariamente como autor. La taxonomía CRediT puede complementar el acuerdo del equipo al describir conceptualización, curación de datos, software, análisis, supervisión o redacción.

Una herramienta de IA no puede aprobar la versión final, responder a cuestionamientos ni asumir responsabilidad; por ello no reúne criterios de autoría. Tampoco “posee” intención científica. Los seres humanos deben verificar cada cifra, referencia, afirmación y fragmento de código. Copiar una salida sin atribución o incorporarla sin revisión puede producir plagio, citas inexistentes y errores encubiertos por una prosa plausible.

La integridad computacional requiere preservar datos sin procesar cuando sea legítimo, scripts, semillas aleatorias, versiones de paquetes, arquitectura, hiperparámetros, divisiones de datos y registro de exclusiones. El conjunto de prueba no debe convertirse en un recurso de ajuste iterativo. Si se prueban muchas configuraciones y se publica solo la mejor, debe declararse el proceso de selección y usarse validación anidada o un conjunto final verdaderamente independiente.

Práctica problemática	**Por qué compromete la evidencia**	**Prevención**
Autoría honoraria	Atribuye crédito sin responsabilidad.	Acuerdo temprano y declaración CRediT.
Salami slicing	Fragmenta resultados y oculta la unidad del estudio.	Plan de publicaciones y referencias cruzadas.
HARKing	Presenta hipótesis formuladas después como si fueran previas.	Prerregistro y distinción confirmatorio-exploratorio.
Ajuste al conjunto de prueba	Infla el desempeño esperado.	Congelar prueba y documentar iteraciones.
Manipulación de imágenes	Cambia evidencia visual o elimina hallazgos.	Conservar originales y registrar procesamiento.
Citas generadas no verificadas	Introduce fuentes falsas o incongruentes.	Verificar título, autor, revista, año y DOI en la fuente primaria.

3.9 Declaración del uso de herramientas de IA

La declaración debe permitir reconstruir qué herramienta intervino, para qué, con qué datos y bajo qué control humano. El ICMJE indica que el uso de tecnologías asistidas por IA debe declararse al enviar el manuscrito; si se usaron en recolección, análisis o figuras, corresponde describirlas en métodos, y si se limitaron a asistencia de escritura puede corresponder un reconocimiento, según la política editorial (ICMJE, 2026). Las reglas de la revista de destino prevalecen y deben consultarse en su versión vigente.

Una frase genérica como “se utilizó IA para mejorar el texto” es insuficiente cuando la herramienta generó código, extrajo variables o transformó imágenes. Deben consignarse nombre y versión accesible, proveedor, fecha o periodo de acceso, finalidad, entradas relevantes sin revelar datos sensibles, parámetros reproducibles cuando existan, procedimientos de verificación,

modificaciones humanas y responsabilidad final. Para interfaces que cambian sin control del usuario, conservar fecha, exportación o identificador de versión mejora la trazabilidad.

La transparencia no legitima un uso impropio. Declarar que se cargaron expedientes identificables en una plataforma no autorizada documenta el problema, pero no lo resuelve. Primero se requiere base legítima, seguridad y aprobación; después, reporte. Tampoco es necesario sobredimensionar usos triviales si la política editorial los exceptúa. La decisión debe ser coherente y verificable.

Modelo de declaración adaptable

Durante la preparación del manuscrito se utilizó [herramienta, proveedor, versión o fecha] para [finalidad específica]. Se ingresaron únicamente [tipo de contenido; indicar que estaba anonimizado o era sintético cuando corresponda]. Las salidas fueron verificadas por [rol de las personas] mediante [procedimiento: comparación con datos fuente, ejecución independiente del código, comprobación de referencias]. La herramienta no tomó decisiones clínicas ni determinó la inclusión de resultados. Los autores revisaron y aprobaron la versión final y asumen responsabilidad por su exactitud, originalidad e integridad.

Ejemplos de ubicación

- Métodos: uso de un modelo de lenguaje para extraer signos clínicos, con esquema de etiquetas, versión, validación frente a doble revisión humana y métricas.
- Métodos de análisis: asistencia para redactar código, seguida de pruebas unitarias, revisión por un estadístico y ejecución reproducible sobre el conjunto autorizado.
- Reconocimientos: apoyo lingüístico sin generación ni interpretación de resultados, si así lo permite la revista.
- Leyenda de figura: imagen sintética o componente generado, cuando la política editorial lo autoriza y no representa evidencia primaria.

3.10 Regulación de sistemas de inteligencia artificial

La regulación de IA es fragmentaria y depende del territorio, finalidad, producto y nivel de riesgo. Un mismo algoritmo puede ser una herramienta de investigación, un componente de software clínico, un servicio comercial o parte de una decisión administrativa. Antes de usarlo debe determinarse qué normas se aplican a protección de datos, ejercicio profesional, bienestar animal,

dispositivos o productos veterinarios, propiedad intelectual, consumo, contratación y ciberseguridad.

El Reglamento europeo de IA, Reglamento (UE) 2024/1689, adopta obligaciones diferenciadas por riesgo y entró en vigor el 1 de agosto de 2024; la mayoría de sus disposiciones son aplicables desde el 2 de agosto de 2026, con excepciones y periodos específicos (European Parliament & Council of the European Union, 2024; European Commission, 2026). Su alcance no convierte automáticamente a todo software veterinario en sistema de alto riesgo. La clasificación depende de uso previsto y anexos aplicables. Sin embargo, sus ideas de gestión de riesgos, documentación, calidad de datos, supervisión humana, transparencia y vigilancia ofrecen una referencia útil.

En Ecuador, la primera pregunta para expedientes veterinarios es si se tratan datos personales del tutor, trabajador o productor. La LOPDP y su Reglamento exigen identificar fundamento, finalidad, proporcionalidad, medidas de seguridad y derechos de titulares. La gobernanza de un algoritmo debe integrarse a estas obligaciones. Para un proyecto multinacional pueden concurrir reglas de varios países y transferencias internacionales; no basta almacenar el archivo fuera del país para dejar de aplicar la normativa de origen.

NIST publicó un marco voluntario de gestión de riesgos de IA organizado alrededor de gobernar, mapear, medir y gestionar (Tabassi, 2023). UNESCO aprobó en 2021 una recomendación global sobre ética de IA que enfatiza derechos humanos, proporcionalidad, no discriminación, transparencia, responsabilidad y supervisión humana (UNESCO, 2021). Estos instrumentos no sustituyen la ley, pero ayudan a convertir principios en controles institucionales.

En veterinaria persisten vacíos específicos, como advirtió Bellamy (2023). La ausencia de una autorización sectorial expresa no equivale a ausencia de deber profesional. Una clínica que adopta un detector radiográfico comercial debe exigir población de entrenamiento, validación externa, limitaciones, actualización, seguridad, gestión de incidentes y condiciones de uso de sus datos. Si el proveedor no ofrece evidencia suficiente, el sistema no debería influir en decisiones de alto impacto.

Marco	Naturaleza	Aplicación práctica al proyecto veterinario

LOPDP Ecuador 2021 y Reglamento 2023	Normativa de datos personales.	Base de tratamiento, información al titular, seguridad, contratos y derechos.
Reglamento UE 2024/1689	Norma horizontal de IA basada en riesgos.	Clasificar rol y uso previsto; revisar obligaciones de transparencia y gestión.
WOAH Código Terrestre capítulo 7.8	Estándar internacional de bienestar.	Supervisión institucional, 3R, registros y responsabilidades.
NIST AI RMF 1.0	Marco voluntario de gestión.	Gobernar, mapear, medir y gestionar riesgos durante el ciclo de vida.
TRIPOD+AI y PROBAST+AI	Guías metodológicas.	Reporte completo y evaluación de sesgo y aplicabilidad de modelos.

3.11 Riesgos de la IA generativa en investigación

Los modelos generativos producen secuencias plausibles, no verificaciones automáticas de verdad. Pueden inventar referencias, mezclar resultados de estudios, omitir condiciones, presentar código que se ejecuta pero responde otra pregunta o redactar interpretaciones causales a partir de asociaciones. En un experimento veterinario, una salida fluida puede ocultar que la unidad experimental fue mal definida o que se aplicó una t de Student a medidas repetidas sin considerar dependencia.

La alucinación bibliográfica merece un control separado. Cada cita debe verificarse en la publicación original: autores, título, revista, año, volumen, páginas o número de artículo y DOI. Una herramienta generativa no es fuente científica. ICMJE responsabiliza a los autores por exactitud, originalidad, ausencia de plagio y atribución completa, incluso cuando intervino IA (ICMJE, 2026). Estudios experimentales han mostrado que textos científicos sintéticos pueden resultar convincentes para revisores humanos, lo que refuerza la necesidad de comprobar evidencia y no juzgar por estilo (Gao et al., 2023).

El riesgo de confidencialidad aparece al introducir expedientes, imágenes, bases inéditas, manuscritos bajo revisión o resultados con propiedad intelectual en servicios externos. También existe riesgo de memorización, retención contractual, acceso indebido y transferencia

internacional. La regla práctica es sencilla: ningún dato no público debe enviarse a una herramienta sin aprobación institucional, base jurídica y revisión de sus condiciones técnicas y contractuales.

En programación, la IA puede acelerar borradores, pero el equipo debe entender y probar el código. Son errores frecuentes la recodificación invertida del desenlace, imputación antes de separar datos, fuga de información, cálculo de métricas sobre entrenamiento, uso de accuracy en clases desbalanceadas y selección del mejor modelo usando repetidamente la prueba. La verificación incluye datos sintéticos con resultados conocidos, pruebas unitarias, revisión de objetos intermedios y comparación con un método de referencia.

En imágenes, la generación o edición puede crear estructuras anatómicas inexistentes, borrar artefactos o modificar una lesión. Nunca debe emplearse para "mejorar" evidencia primaria de manera que cambie su contenido. Los ajustes permitidos deben aplicarse de forma uniforme, conservar originales y declararse. Una ilustración sintética educativa debe rotularse como tal y no confundirse con una radiografía o histología real.

Riesgo	**Señal de alerta**	**Verificación humana**
Referencia falsa	DOI no resuelve o título no coincide.	Comprobar en editorial, PubMed o registro DOI.
Error estadístico	Código ejecuta sin revisar supuestos o unidad.	Auditar diseño, datos simulados y resultados intermedios.
Fuga de confidencialidad	Prompt contiene nombres, fechas o ubicaciones.	Minimizar localmente y usar entorno aprobado.
Plagio o atribución deficiente	Prosa muy cercana a fuente sin cita.	Comparar, reescribir desde comprensión y citar el original.
Imagen alterada	Textura o anatomía inconsistente.	Conservar archivo original y registro de procesamiento.
Dependencia cognitiva	Se acepta la salida porque suena segura.	Solicitar evidencia, buscar contraejemplos y revisión experta.

3.12 Caso didáctico uso secundario de expedientes clínicos veterinarios

Caso didáctico simulado. Una clínica universitaria dispone de 18 400 expedientes de perros atendidos entre 2021 y 2025. Un equipo propone desarrollar un modelo que estime, al ingreso, el riesgo de hospitalización en las siguientes 24 horas. Los registros contienen datos demográficos

del animal, signos vitales, motivo de consulta, texto libre, pruebas solicitadas, diagnóstico, dirección y teléfono del tutor, profesional tratante, pagos y desenlace. No hubo un consentimiento específico para entrenamiento de IA cuando se recogieron los datos.

Planteamiento y valor científico

La pregunta es si la información disponible al momento de la admisión permite estimar un riesgo suficientemente calibrado para apoyar priorización clínica. El objetivo no es sustituir el triaje ni decidir altas. La utilidad potencial consiste en identificar pacientes que requieren evaluación más rápida y planificar recursos. Antes de extraer datos, el equipo revisa literatura, define el uso previsto y compara el modelo con una regla clínica simple. Si la regla ofrece desempeño similar, la mayor complejidad no estaría justificada.

Población unidad y desenlace

La población objetivo son perros atendidos en servicios de primera opinión con características semejantes a la clínica. La unidad de análisis es el episodio clínico, pero un mismo animal puede aportar varios episodios; para evitar fuga, todas sus visitas deben permanecer en una sola partición. El desenlace “hospitalización en 24 horas” se define mediante hora de admisión, orden y hora de ingreso. Se excluyen hospitalizaciones programadas porque responden a un proceso distinto. Las exclusiones se fijan antes de observar el desempeño.

Evaluación ética y base de tratamiento

El protocolo se somete al comité institucional o instancia competente. Se solicita dispensa de consentimiento individual por riesgo mínimo e impracticabilidad de contactar a miles de tutores, pero la dispensa no se da por supuesta. La institución documenta su base jurídica bajo la LOPDP, evalúa compatibilidad de la finalidad, informa cuando corresponda mediante aviso de privacidad y limita el conjunto a variables necesarias. Dirección, teléfono, nombres y pagos se eliminan antes de que los analistas reciban los datos.

Flujo de datos y seguridad

Un custodio autorizado extrae los registros dentro del entorno institucional. Genera un identificador aleatorio de investigación y conserva la tabla de correspondencia cifrada en una ubicación separada. Las fechas exactas se transforman cuando no son indispensables; el texto libre se procesa con reglas y revisión de identificadores. Los archivos no se envían a chatbots públicos. El acceso se concede por rol, vence al terminar el proyecto y queda registrado. Se establece respuesta a incidentes y un plazo de conservación.

Diseño analítico y sesgo

Se reserva el año 2025 como validación temporal. Los años 2021-2024 se separan por animal en entrenamiento y validación interna. Se compara regresión logística penalizada con gradient boosting. Las variables posteriores a la decisión de hospitalizar, como tratamientos administrados o costo final, se excluyen por fuga temporal. Los datos faltantes se describen por variable y periodo; la imputación se ajusta solo con entrenamiento.

Además de área bajo la curva ROC, el equipo informa sensibilidad, especificidad, valores predictivos, Brier y calibración en el umbral clínico. Examina desempeño por sexo, grupos de edad, tamaño corporal, condición reproductiva, razas con suficiente muestra, procedencia urbana/rural y turno. Para grupos pequeños se presentan intervalos amplios y no se concluye equivalencia por falta de significación. Si la calibración falla en animales geriátricos, el sistema se restringe o recalibra antes de uso.

Gobernanza del despliegue

El modelo se implementa primero en modo silencioso: calcula riesgo sin mostrarlo al clínico. Esta fase permite evaluar rendimiento prospectivo, latencia y datos faltantes. Solo después de aprobación se muestra como apoyo. El profesional mantiene autoridad, documenta discrepancias y no comunica la probabilidad al tutor sin contexto. Se monitorean hospitalizaciones omitidas, falsas alertas, calibración y desempeño por subgrupo. Un comité revisa mensualmente incidentes y puede suspender el sistema.

Decisión razonada del comité

Aprobación condicionada. El proyecto tiene valor potencial y usa datos existentes, pero debe cumplir: minimización y seudonimización antes del acceso analítico; justificación documentada de la dispensa; prohibición contractual de uso por proveedores para entrenamiento propio; separación por animal; validación temporal; auditoría por subgrupos; plan de incidentes; revisión antes de cada actualización; y publicación transparente de limitaciones. Cualquier expansión a otras especies, centros o decisiones de alta requiere una nueva evaluación.

Riesgo del caso	**Probabilidad**	**Impacto**	**Salvaguarda principal**
Reidentificación desde texto libre	Media	Alta	Desidentificación combinada y auditoría de muestra.
Fuga de datos entre visitas	Alta sin control	Alta	Partición por identificador de animal.
Sesgo por capacidad de pago	Media	Alta	Excluir pagos; evaluar proxies y subgrupos.
Automatización del triaje	Media	Alta	Uso de apoyo, modo silencioso y supervisión real.
Deriva temporal	Alta	Media-alta	Monitoreo de calibración y criterio de retiro.
Uso comercial no autorizado	Baja-media	Alta	Contrato, finalidad limitada y control de exportaciones.

3.13 Síntesis autoevaluación y actividad práctica

Síntesis

- La calidad metodológica es una obligación ética: un diseño incapaz de responder la pregunta desperdicia animales, datos y confianza.
- Las 3R se aplican desde la planificación; la IA puede favorecerlas o aumentar procedimientos si no se controla su finalidad.
- El consentimiento clínico, la participación en investigación y la reutilización de datos son decisiones distintas.
- Los expedientes veterinarios pueden contener datos personales del tutor, trabajadores y productores, aun cuando el paciente sea un animal.
- FAIR no significa abierto sin restricciones; los datos sensibles pueden ser localizables y reutilizables mediante acceso controlado.
- La exactitud global no demuestra equidad. La validación debe respetar la unidad, evitar fuga y examinar desempeño y calibración por subgrupos.
- La supervisión humana requiere competencia, tiempo, información y poder real para contradecir o suspender el sistema.

- Las herramientas de IA no son autoras ni fuentes científicas; su uso se declara y toda salida se verifica.
- La regulación depende del uso previsto y la jurisdicción. Los vacíos sectoriales no eliminan el deber de diligencia profesional.
- La gobernanza cubre todo el ciclo: propósito, acceso, análisis, despliegue, monitoreo, incidentes, actualización y retiro.

Autoevaluación

1. ¿Por qué un estudio con pseudorreplicación constituye también un problema ético?

Respuesta orientativa: Porque trata observaciones dependientes como réplicas, subestima la incertidumbre y puede producir conclusiones inválidas tras usar animales o datos.

2. Diferencie reemplazo, reducción y refinamiento con un ejemplo veterinario para cada principio.

Respuesta orientativa: Reemplazo evita animales vivos; reducción optimiza el número compatible con validez; refinamiento disminuye daño y mejora bienestar.

3. ¿Por qué "retrospectivo" no significa automáticamente "exento de consentimiento"?

Respuesta orientativa: Porque persisten riesgos, finalidades no previstas, datos personales y requisitos de revisión; una dispensa debe estar fundamentada.

4. Explique la diferencia entre anonimización y seudonimización.

Respuesta orientativa: La anonimización busca impedir razonablemente el vínculo; la seudonimización conserva la posibilidad de enlace mediante clave o información adicional.

5. ¿Cómo puede una historia clínica animal contener datos personales protegidos?

Respuesta orientativa: Incluye identificadores y circunstancias del tutor, trabajadores, finca, pagos, imágenes o geolocalización.

6. ¿Por qué FAIR no equivale a acceso público irrestricto?

Respuesta orientativa: La accesibilidad puede ser controlada; FAIR exige condiciones claras y metadatos, no divulgación de datos sensibles.

7. Describa dos formas de fuga de datos en un modelo basado en expedientes veterinarios.

Respuesta orientativa: Visitas del mismo animal en entrenamiento y prueba; variables recogidas después del desenlace disponibles durante el entrenamiento.

8. ¿Qué métricas revisaría por subgrupo para evaluar un clasificador de urgencias?

Respuesta orientativa: Sensibilidad, especificidad, valores predictivos, tasa de falsos negativos y calibración, con intervalos y denominadores.

9. ¿Por qué no siempre es posible satisfacer simultáneamente todas las definiciones de equidad?

Respuesta orientativa: Con prevalencias diferentes, igualar una métrica puede empeorar otra; debe priorizarse según el daño clínico.

10. Enumere cinco elementos mínimos de transparencia antes de adquirir un sistema de IA veterinaria.

Respuesta orientativa: Uso previsto, población, datos, definición de desenlace, validación, métricas, subgrupos, limitaciones, versión y monitoreo.

11. ¿Qué convierte una supervisión humana en meramente simbólica?

Respuesta orientativa: La falta de competencia, tiempo, información o facultad para contradecir la **salida.**

12. Aplique los cuatro criterios de autoría del ICMJE a un colaborador que solo aportó datos.

Respuesta orientativa: No basta aportar datos; debe participar intelectualmente, aprobar la versión y asumir responsabilidad, o figurar como contribuidor.

13. ¿Dónde declararía una IA usada para extraer variables clínicas y qué información incluiría?

Respuesta orientativa: En métodos, con herramienta y versión, finalidad, entradas, parámetros, validación y revisión humana.

14. ¿Qué riesgos introduce un chatbot cuando se le entrega texto libre de historias clínicas?

Respuesta orientativa: Divulgación o retención de identificadores, transferencia no autorizada, reidentificación y uso secundario por el proveedor.

15. Proponga un criterio de suspensión para un modelo de predicción de hospitalización.

Respuesta orientativa: Por ejemplo, suspender si la sensibilidad cae bajo el umbral predefinido o aparece daño grave atribuible.

Actividad práctica auditoría ética y de gobernanza

Propósito. Elaborar una evaluación previa de un proyecto que utiliza IA con datos veterinarios. La actividad puede realizarse individualmente o en equipos de tres a cinco integrantes. El producto final será una matriz de riesgos, un flujo de datos y una decisión motivada de aprobar, condicionar o no aprobar.

Escenario

Una red de cuatro clínicas desea entrenar un modelo para detectar riesgo de enfermedad renal crónica felina a partir de edad, peso, creatinina, densidad urinaria, notas clínicas y fotografías. Dos centros son urbanos, uno rural y otro de referencia. El proveedor solicita copiar todos los expedientes a su nube y conservarlos para mejorar futuros productos. No ofrece resultados por centro ni raza; promete 94 % de accuracy en datos internos.

Tareas

- Defina pregunta, uso previsto, población objetivo, unidad de análisis y desenlace. Identifique qué variables no son necesarias.

- Dibuje el flujo desde la clínica hasta el análisis y marque puntos de identificación, transferencia, acceso y eliminación.
- Determine qué consentimientos, avisos, bases jurídicas y revisiones institucionales deben comprobarse.
- Construya un registro con al menos ocho riesgos: bienestar, privacidad, seguridad, sesgo, metodología, automatización, integridad y contratación.
- Proponga una partición que evite fuga por animal y centro, y un plan de validación temporal o externa.
- Defina métricas clínicas y subgrupos. Explique qué daño producirían falsos positivos y falsos negativos.
- Redacte cinco condiciones contractuales para el proveedor, incluyendo prohibición o autorización expresa de reutilización.
- Redacte la declaración del uso de IA que aparecería en métodos.
- Emita una decisión motivada con condiciones, responsable, plazo y criterios de suspensión.

Rúbrica de evaluación

Criterio	Excelente	Aceptable	Insuficiente
Problema y diseño	Uso previsto, unidad, tiempo y desenlace son coherentes.	La mayoría está definida, con una ambigüedad menor.	Confunde objetivo, unidad o desenlace.
Ética y datos	Distingue consentimiento, fundamento, minimización y 3R.	Reconoce controles principales sin justificar todos.	Supone que los datos retrospectivos son libres.
Sesgo y validación	Evita fuga, valida externamente y analiza subgrupos.	Separa datos correctamente pero limita análisis de equidad.	Usa partición aleatoria con dependencia o solo accuracy.
Gobernanza	Asigna responsables, incidentes, actualización y retiro.	Incluye roles y monitoreo básicos.	No define rendición de cuentas.
Argumentación	Decisión proporcional, documentada y viable.	Decisión razonable con evidencia parcial.	Decisión basada en promesas del proveedor.

Lista de verificación para proyectos veterinarios con IA

- La pregunta y el beneficio esperado están definidos antes de acceder a datos.
- La unidad experimental u observacional y las dependencias están identificadas.
- Se justificaron 3R, riesgos, rescate y puntos finales cuando participan animales.

- Consentimiento, dispensa o fundamento de reutilización fueron revisados por la instancia competente.
- Solo se recogen variables necesarias y se controlan identificadores directos e indirectos.
- Roles, permisos, transferencias, conservación y eliminación están documentados.
- La partición evita fuga por animal, finca, centro, dispositivo y tiempo.
- Se evalúan discriminación, calibración, utilidad clínica y desempeño por subgrupo.
- Versión, datos, código, parámetros y decisiones se conservan para auditoría.
- La supervisión humana tiene autoridad efectiva y existe un canal de incidentes.
- El uso de IA generativa se limita a entornos y datos autorizados y se declara.
- Hay criterios de actualización, recalibración, suspensión y retiro.

CAPÍTULO 4

DISEÑO DE INVESTIGACIONES VETERINARIAS ASISTIDAS POR INTELIGENCIA ARTIFICIAL

Introducción

La calidad de una investigación veterinaria no depende de la sofisticación del algoritmo empleado al final del proceso, sino de la correspondencia entre la pregunta científica, la estructura de asignación de los tratamientos, el modo de obtener las observaciones y el análisis previsto. Un modelo de aprendizaje automático puede reconocer patrones en miles de registros o imágenes; sin embargo, no puede reconstruir un grupo control inexistente, separar un efecto terapéutico de un sesgo de selección ni convertir mediciones repetidas del mismo animal en réplicas biológicas independientes. La inteligencia artificial amplía la capacidad de planificación, simulación y auditoría, pero conserva los límites impuestos por el diseño.

Las ciencias veterinarias presentan estructuras de datos particularmente jerárquicas. Los animales pertenecen a camadas, corrales, hatos, clínicas o regiones; una vaca aporta varios ordeños y cuartos mamarios; un perro puede generar múltiples radiografías; un ave se observa muchas veces dentro de un galpón; una herida produce mediciones sucesivas del área. Estas dependencias deben reconocerse antes de calcular el tamaño muestral o dividir una base para entrenar un modelo predictivo. Ignorarlas produce intervalos de confianza demasiado estrechos, valores p artificialmente pequeños y métricas de clasificación exageradas.

El capítulo desarrolla los fundamentos de los estudios observacionales y experimentales, distingue las unidades experimental y observacional, examina la aleatorización, el bloqueo, el enmascaramiento, las réplicas y la pseudorreplicación, y conecta estos conceptos con diseños factoriales, medidas repetidas, ensayos clínicos y estudios longitudinales. También se explican dos lógicas distintas de tamaño muestral: la inferencial, centrada en estimar o contrastar efectos, y la predictiva, orientada a obtener modelos estables, calibrados y transportables. La última parte integra estos principios en un caso didáctico simulado de cicatrización posquirúrgica canina.

Objetivos de aprendizaje

- Diferenciar estudios observacionales, experimentos controlados y estudios de desarrollo o validación de modelos predictivos.

- Identificar correctamente la unidad experimental, la unidad observacional y los niveles de agrupamiento de los datos veterinarios.
- Seleccionar controles, comparadores, procedimientos de aleatorización, bloqueo y enmascaramiento compatibles con la pregunta científica.
- Distinguir réplicas biológicas de réplicas técnicas y reconocer situaciones de pseudorreplicación.
- Relacionar diseños completamente aleatorizados, en bloques, factoriales, longitudinales y de medidas repetidas con modelos estadísticos apropiados.
- Justificar el tamaño muestral para objetivos inferenciales y predictivos sin recurrir a reglas universales.
- Prevenir confusión, sesgo de selección y fuga de información en proyectos que incorporan aprendizaje automático.
- Utilizar herramientas de IA como apoyo auditable a la planificación, sin delegar decisiones científicas, éticas o clínicas.

Principio rector

La unidad que recibe independientemente un tratamiento o exposición define la replicación efectiva. La unidad que produce una medición define la resolución observacional. Confundir ambas suele inflar el tamaño muestral aparente.

Primero se diseña la comparación científica; después se elige el análisis y, si procede, el algoritmo. La secuencia inversa favorece preguntas moldeadas por el software y resultados difíciles de interpretar.

4.1. Estudios observacionales y experimentales

Un estudio observacional registra exposiciones y desenlaces sin que el investigador asigne deliberadamente la exposición principal. Son observacionales una cohorte de perros con enfermedad valvular seguidos durante dos años, un estudio transversal de seroprevalencia de brucelosis en hatos lecheros o un diseño de casos y controles sobre factores asociados con mastitis

clínica. La ausencia de asignación no implica ausencia de diseño: exige definir población fuente, criterios de inclusión, momento cero, estrategia de muestreo, temporalidad entre exposición y resultado, variables de confusión y mecanismos de pérdida de seguimiento, como sistematiza la guía STROBE-Vet y su documento explicativo (Sargeant et al., 2016; O'Connor et al., 2016).

El estudio experimental introduce una intervención bajo control del investigador. En un ensayo terapéutico, los animales pueden asignarse a un analgésico estándar o a una combinación farmacológica; en producción animal, lotes o corrales pueden recibir dietas diferentes; en un estudio de laboratorio, cultivos celulares se exponen a concentraciones crecientes de un compuesto. La asignación permite fortalecer la interpretación causal, especialmente cuando es aleatoria y se preservan la adherencia, el seguimiento y el enmascaramiento. No obstante, la sola etiqueta "experimental" no garantiza validez: si todos los animales tratados se alojan en una sala y todos los controles en otra, tratamiento y sala quedan completamente confundidos.

Los estudios de predicción constituyen otra dimensión. Pueden utilizar datos observacionales o experimentales, pero su objetivo no es necesariamente estimar el efecto de una exposición. Un modelo que predice mastitis subclínica a partir de conductividad eléctrica, producción, días en leche y temperatura ambiental busca anticipar un estado futuro. Para evaluarlo se requieren separación temporal o geográfica, métricas de discriminación y calibración, análisis de utilidad clínica y validación externa. Preguntar si un factor "predice" no equivale a preguntar si "causa"; una variable puede ser un buen marcador predictivo sin ser modificable ni causal.

Tabla 4.1.

Relación entre pregunta, diseño y conclusión admisible

Pregunta principal	Diseño preferente	Unidad frecuente	Conclusión defendible	Riesgo dominante
¿Cuál es la prevalencia de parasitosis?	Transversal con muestreo probabilístico	Animal; posible agrupamiento por finca	Frecuencia y distribución en la población definida	Sesgo de cobertura o selección
¿La exposición antecede al desenlace?	Cohorte prospectiva o retrospectiva bien definida	Animal, hato o ambos	Asociación temporal; causalidad	Confusión y pérdidas

			condicionada al control de sesgos	
¿El tratamiento mejora el resultado?	Ensayo aleatorizado controlado	Animal o conglomerado asignado	Efecto causal promedio bajo los supuestos del ensayo	Incumplimiento, desenmascaramiento
¿Quién desarrollará mastitis?	Cohorte para desarrollo y validación predictiva	Vaca/hato con fecha índice	Riesgo individual estimado en un horizonte definido	Sobreajuste y fuga de información
¿Cómo interactúan dieta y ambiente?	Factorial, idealmente bloqueado	Corral, animal o unidad productiva	Efectos principales e interacción	Baja potencia para la interacción

Ejemplo aplicado: radiografías caninas

Pregunta diagnóstica: clasificar radiografías como compatibles o no con enfermedad torácica. Si cada perro aporta cuatro proyecciones y las imágenes se dividen al azar, una imagen del mismo paciente puede quedar en entrenamiento y otra en prueba. El resultado es una evaluación contaminada.

Decisión de diseño: todas las imágenes de un perro deben permanecer en una única partición. Si el objetivo es generalizar a otros hospitales, la validación externa debe provenir de una clínica no utilizada durante el entrenamiento ni el ajuste.

4.2. Unidad experimental y unidad observacional

La unidad experimental es la entidad mínima que puede recibir de manera independiente una condición experimental. La unidad observacional es la entidad sobre la cual se registra una medición. Pueden coincidir, como cuando cada perro se asigna individualmente a un tratamiento y se mide una vez su puntuación de dolor. También pueden diferir: si cuatro dietas se asignan a corrales de pollos, el corral es la unidad experimental, aunque se pese individualmente a cada ave.

Los pesos mejoran la descripción de la variabilidad dentro del corral, pero no convierten a las aves en asignaciones independientes.

Una investigación puede contener más de un nivel de observación. En mastitis bovina, el hato condiciona manejo y bioseguridad; la vaca aporta edad, paridad y producción; el cuarto mamario aporta conductividad y cultivo; el ordeño aporta una serie temporal. El análisis debe representar esa jerarquía mediante efectos aleatorios, correlaciones residuales u otras estructuras adecuadas. El número de filas de una base no es sinónimo de número de unidades experimentales ni de tamaño muestral efectivo.

La definición correcta depende de cómo se aplica la intervención. Si un aerosol se administra a toda una nave porcina, la nave es la unidad experimental. Si dentro de cada nave los cerdos reciben individualmente un medicamento oral aleatorizado, el cerdo puede ser la unidad experimental para ese factor, aunque persista correlación por nave. En diseños con intervenciones a distintos niveles —por ejemplo, ventilación asignada a galpones y suplemento asignado a aves— existe una estructura de parcela dividida que debe anticiparse desde el protocolo.

Tabla 4.2.

Identificación de unidades en escenarios veterinarios

Escenario	**Asignación**	**Unidad experimental**	**Unidad observacional**	**Análisis orientativo**
Dieta aplicada a 12 corrales de cerdos	Una dieta por corral	Corral	Cerdo y corral	Modelo mixto o análisis por corral
Analgésico asignado a cada perro	Individual	Perro	Perro en varios tiempos	Modelo mixto de medidas repetidas
Desinfectante aplicado a salas de hospital	Una condición por sala/periodo	Sala-periodo	Superficie o muestra microbiológica	Diseño cruzado por periodos; efecto de sala
Cultivo celular con tres pozos por muestra	Tratamiento aplicado a pozos	Pozo, si la aplicación es independiente; donante para inferencia biológica	Lectura por pozo	Modelo jerárquico donante/pozo

Vacuna aplicada por lote avícola	Lote completo	Lote	Ave, muestra o mortalidad diaria	Modelo de conglomerados

4.3. Tratamientos, controles y comparadores

Un tratamiento es una condición deliberadamente aplicada; un comparador es la referencia frente a la cual se interpreta su efecto. Puede ser placebo, atención habitual, tratamiento activo, dosis alternativa, control histórico o ausencia de intervención. La selección no es una formalidad: define la pregunta. Comparar un apósito nuevo con "ningún cuidado" responde si es mejor que la inacción; compararlo con el estándar quirúrgico responde si aporta beneficio sobre la práctica vigente.

El control negativo estima el comportamiento sin el componente activo; el control positivo verifica que el sistema detecta una respuesta conocida; el control vehicular separa el efecto del principio activo del efecto de su medio de administración. En microbiología y ensayos moleculares, los controles de contaminación y amplificación pertenecen al control técnico, no sustituyen controles biológicos. En estudios de IA, un modelo simple —por ejemplo, regresión logística o la prevalencia basal— debe funcionar como comparador. Una red neuronal no demuestra valor por superar una clasificación al azar; debe competir con una referencia clínica plausible y bajo idéntico esquema de validación.

Cuando el tratamiento estándar es eficaz, retirar terapia para construir un placebo puede ser éticamente improcedente. En ese caso se emplea un comparador activo y se decide si la pregunta es de superioridad, no inferioridad o equivalencia. Los márgenes de no inferioridad deben justificarse clínicamente antes de observar los datos. En animales de compañía, el consentimiento del tutor no reemplaza la revisión ética ni legitima un comparador que exponga al paciente a dolor evitable.

4.4. Aleatorización, bloqueo y enmascaramiento

La aleatorización crea un mecanismo conocido de asignación y reduce la posibilidad de que la gravedad, la edad, el sexo, el operador o la preferencia del investigador determinen el tratamiento. No consiste en alternar tratamientos, usar la fecha de ingreso ni aceptar el orden de llegada: esos

procedimientos son previsibles. Una secuencia aleatoria debe generarse con un método reproducible, conservarse con ocultamiento de la asignación y aplicarse en el nivel de la unidad experimental.

El bloqueo agrupa unidades similares antes de aleatorizar, con el propósito de separar variación conocida de la variación residual. En un ensayo de cicatrización canina, pueden formarse bloques por centro clínico o por categoría de tamaño corporal; dentro de cada bloque se asignan los tratamientos. En producción lechera, hato, paridad o etapa de lactancia pueden ser factores de bloqueo. El bloqueo mejora precisión cuando el factor está relacionado con el desenlace; un exceso de estratos en muestras pequeñas, en cambio, produce celdas vacías y procedimientos difíciles de ejecutar.

El enmascaramiento protege diferentes etapas. El personal que administra el tratamiento puede desconocer la asignación; el cirujano puede no estar cegado por razones operativas, pero el evaluador de la herida sí; las fotografías pueden codificarse antes del análisis; el estadístico puede trabajar inicialmente con grupos A y B. Debe declararse quién estuvo enmascarado, cómo se mantuvo el cegamiento y qué situaciones permitían romperlo. “Doble ciego” es insuficiente si no se identifican las partes cegadas.

La IA puede ayudar a producir secuencias estratificadas, verificar desequilibrios y detectar desviaciones de protocolo. No debe recibir datos identificables en plataformas no autorizadas ni decidir por sí sola exclusiones posteriores. Una herramienta generativa puede escribir código aparentemente correcto que, por un error de semilla, estrato o nivel de asignación, genere una secuencia inapropiada; toda salida exige revisión estadística y prueba con datos ficticios.

Ejemplo aplicado: dolor posoperatorio canino

Diseño: perros sometidos a un procedimiento quirúrgico comparable se asignan a meloxicam o meloxicam más un analgésico de rescate planificado. La aleatorización se estratifica por centro y categoría de intensidad quirúrgica.

Enmascaramiento: la medicación se prepara por un profesional no implicado en la valoración. Un evaluador entrenado aplica una escala de dolor validada a tiempos predefinidos. El rescate se activa por umbral clínico y se registra como desenlace relevante, no como "dato perdido".

Análisis: un modelo mixto puede representar mediciones repetidas dentro del perro; la proporción que requiere rescate puede analizarse por separado. La interpretación debe incluir diferencia estimada, intervalo de confianza y relevancia clínica.

4.5. Réplicas biológicas y técnicas

Una réplica biológica representa una unidad biológica independiente pertinente para la población de interés: animales distintos, hatos distintos o donantes celulares distintos, según la inferencia buscada. Una réplica técnica repite la medición o el procesamiento sobre la misma unidad para cuantificar variabilidad del instrumento, del ensayo o del operador. Tres lecturas de glucosa del mismo tubo no equivalen a tres animales; tres pozos derivados de la misma muestra no representan tres donantes.

Las réplicas técnicas son útiles para estimar precisión analítica, identificar fallas y promediar ruido, pero no aumentan directamente la diversidad biológica. Su tratamiento estadístico depende del objetivo. Puede promediarse la lectura técnica antes del análisis, incorporarse un nivel jerárquico o modelarse la variación de medición. Contar todas las lecturas como observaciones independientes subestima la incertidumbre.

En visión computacional ocurre una analogía: recortes, rotaciones o variaciones de una misma radiografía son aumentos de datos, no nuevos pacientes. El aumento puede mejorar la robustez del entrenamiento, pero el tamaño biológico continúa determinado por el número de animales

independientes. Las variantes derivadas deben permanecer en la misma partición que la imagen original.

4.6. Pseudorreplicación e independencia de observaciones

La pseudorreplicación aparece cuando observaciones no independientes se analizan como si fueran réplicas independientes del tratamiento. El error puede ser espacial, temporal, jerárquico o técnico. En un experimento con una sola finca tratada y una sola finca control, cientos de vacas no permiten separar el efecto del tratamiento del efecto de finca. En una jaula tratada y una jaula control, el tratamiento queda confundido con la jaula. En una herida fotografiada diariamente, las fotografías no son 21 animales distintos.

La independencia debe evaluarse mediante el mecanismo generador de los datos, no por el aspecto de la hoja de cálculo. Dos observaciones pueden tener identificadores distintos y compartir ambiente, genética, operador o episodio clínico. El remedio no siempre es agregar un efecto aleatorio después de recolectar los datos: cuando no existen réplicas del nivel asignado, ningún modelo identifica el efecto separado de ese nivel. El diseño debe crear replicación real antes del análisis.

En aprendizaje automático, la dependencia genera fuga entre particiones. Si registros del mismo animal aparecen en entrenamiento y prueba, el modelo puede reconocer características individuales. Si datos de 2025 se usan para imputar o normalizar los de 2024 antes de una validación temporal, el futuro informa al pasado. Las particiones agrupadas, la validación dejando un hato o centro fuera y los pipelines que ajustan el preprocesamiento solo en el conjunto de entrenamiento son controles esenciales. Kapoor y Narayanan (2023) documentaron que la fuga constituye un mecanismo importante de resultados excesivamente optimistas en ciencia basada en aprendizaje automático.

Figura 4.1.

Unidad experimental, unidad observacional y réplica técnica

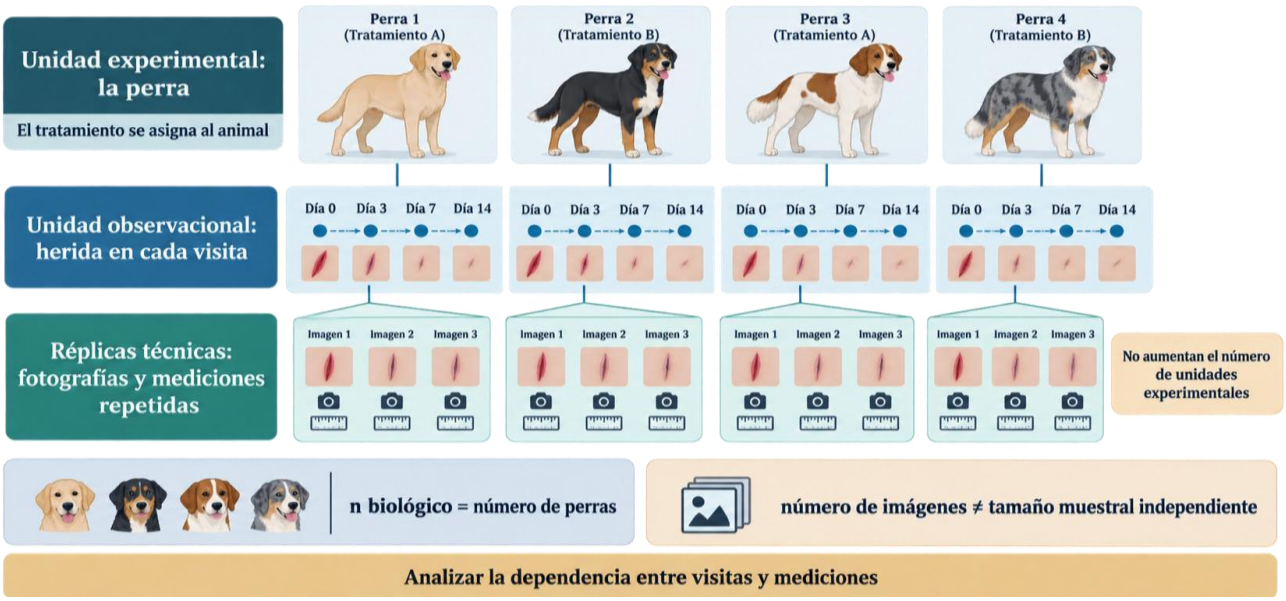


Tabla 4.3.

Pseudorreplicación y corrección de diseño

Situación	Error aparente	Problema real	Corrección
20 aves medidas en un único corral por dieta	n = 20 por grupo	n = 1 corral por dieta	Asignar varias unidades de corral por dieta
30 imágenes por perro divididas al azar	900 imágenes independientes	Dependencia por perro y fuga	Dividir por identificador de perro
Tres lecturas ELISA por muestra	Triplicar n	Réplicas técnicas	Promediar o modelar nivel técnico
Una granja tratada y una control	Muchas vacas por condición	Tratamiento confundido con granja	Aleatorizar múltiples granjas o usar periodos cruzados apropiados
Dolor a 1, 4, 8 y 24 h	Cuatro observaciones independientes	Correlación intraanimal	Modelo de medidas repetidas o mixto

4.7. Diseños completamente aleatorizados y en bloques

En un diseño completamente aleatorizado, todas las unidades experimentales elegibles tienen la misma estructura de asignación y los tratamientos se distribuyen sin bloques explícitos. Es adecuado cuando las unidades son relativamente homogéneas y las condiciones de manejo son comparables. Un ensayo con perros de características clínicas similares, atendidos en un único

centro y asignados individualmente a dos apósitos, podría adoptar esta estructura. El análisis de un desenlace continuo único puede realizarse mediante t de Student o un modelo lineal; con más de dos tratamientos, mediante ANOVA o regresión lineal con contrastes planificados.

El diseño en bloques completos aleatorizados controla una fuente importante de heterogeneidad. Cada bloque contiene todas las condiciones, asignadas aleatoriamente. En estudios de campo, el hato, la camada, el día quirúrgico o el lote de procesamiento pueden ser bloques. Festing (2020) subraya la utilidad general de los diseños completamente aleatorizados y en bloques para investigación preclínica, siempre que se definan unidades y estructura de error correctamente.

Bloquear no significa ajustar retrospectivamente por cualquier covariable. El bloqueo es una decisión prospectiva de asignación; el ajuste por covariables ocurre en el análisis. Ambos pueden combinarse. El factor de bloqueo debe incorporarse en el modelo para respetar el diseño y recuperar la precisión prevista. Si el número de niveles es pequeño y de interés específico, puede tratarse como efecto fijo; si representa una muestra de centros o lotes y se busca generalización, un efecto aleatorio puede ser razonable, siempre con suficientes niveles para estimarlo.

4.8. Diseños factoriales y medidas repetidas

Un diseño factorial estudia simultáneamente dos o más factores y permite estimar interacciones. En un factorial 2 × 2 sobre cicatrización podrían evaluarse tipo de apósito —convencional o con miel— y protocolo analgésico —estándar o multimodal—. El efecto principal del apósito resume su diferencia promedio sobre ambos protocolos analgésicos; la interacción pregunta si el efecto del apósito cambia según el protocolo. Interpretar solo efectos principales cuando existe una interacción relevante puede ocultar respuestas opuestas.

Los factoriales utilizan eficientemente las unidades cuando varios factores pueden aplicarse de manera compatible. No autorizan a tratar cada combinación como un estudio aislado ni a multiplicar comparaciones sin planificación. El tamaño muestral debe considerar la interacción si esta constituye el objetivo principal, pues las interacciones suelen requerir más información que un efecto principal. También debe verificarse el nivel de aplicación: si la dieta se asigna al corral y el fármaco al animal, el diseño es factorial con parcela dividida y tiene más de un término de error.

Las medidas repetidas se obtienen sobre la misma unidad a lo largo del tiempo, bajo diferentes condiciones o en distintos sitios anatómicos. La correlación intraanimal forma parte del diseño. Un ANOVA convencional que ignore esa correlación no es apropiado. Los modelos mixtos permiten incluir interceptos —y a veces pendientes— aleatorios por animal, efectos de tiempo, tratamiento e interacción tratamiento × tiempo. Los modelos de ecuaciones de estimación generalizadas son otra opción para efectos promedio poblacionales. La elección depende del tipo de desenlace, la estructura de datos faltantes y el objetivo inferencial.

El tiempo debe representarse con sentido biológico. Tratar los días 0, 3, 7, 14 y 21 como una tendencia lineal puede ser irreal en cicatrización, donde la velocidad cambia por fases. Puede modelarse como factor categórico, usar transformaciones, splines o modelos aditivos generalizados si la pregunta y el tamaño muestral lo permiten. Un método flexible no compensa intervalos de medición pobres ni ausencia de observaciones en la fase crítica.

Ejemplo aplicado: producción de leche y estrés térmico

Pregunta: ¿la relación entre temperatura-humedad y producción diaria difiere entre vacas primíparas y multíparas?

Estructura: días dentro de vacas, vacas dentro de hatos; estación y etapa de lactancia actúan como covariables. Un GAM mixto puede representar no linealidad y dependencia.

Validación predictiva: para pronosticar producción futura, se respeta el orden temporal. La división aleatoria de días permitiría que registros posteriores de la misma vaca ayuden a predecir registros anteriores.

4.9. Estudios longitudinales y ensayos clínicos veterinarios

Un estudio longitudinal sigue unidades a través del tiempo y permite describir incidencia, trayectorias y secuencia temporal. Debe fijarse un tiempo cero coherente: diagnóstico, cirugía, inicio de exposición o ingreso al hato. Sin una fecha índice uniforme, las ventanas de predicción y seguimiento se mezclan. La pérdida de seguimiento puede introducir sesgo cuando depende del

pronóstico; por ello deben registrarse razones y tiempos de salida, y considerarse análisis de supervivencia, modelos conjuntos o análisis de sensibilidad según el desenlace.

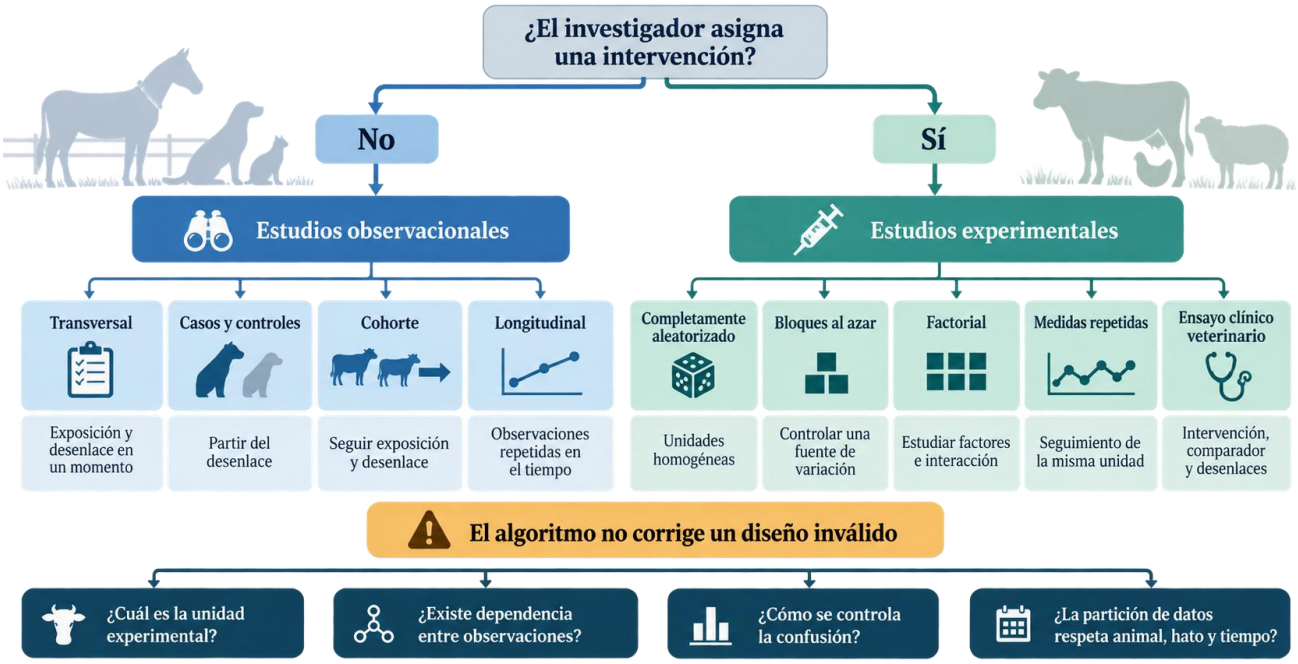
Los ensayos clínicos veterinarios con pacientes de propietarios requieren protocolo prospectivo, revisión ética, consentimiento informado, criterios de inclusión y exclusión, plan de rescate, monitorización de eventos adversos y reglas de retiro. VICH GL9 proporciona un marco de buenas prácticas para estudios de productos veterinarios en la especie objetivo (Food and Drug Administration, 2001). Las recomendaciones de Bertout et al. (2021) enfatizan una revisión ética adaptada a las particularidades de los pacientes veterinarios, sus tutores y los equipos clínicos.

La asignación por conglomerados puede ser necesaria cuando la intervención se implementa en hospitales, refugios, hatos o granjas completas. El cálculo muestral debe incorporar la correlación intraclase y el número de conglomerados, no solo el número total de animales. Con muy pocos centros, la estimación del efecto de centro y la validez externa son frágiles. En diseños cruzados, cada unidad recibe más de una condición en periodos sucesivos; se requiere estabilidad clínica, lavado suficiente y ausencia razonable de efectos residuales.

Si la intervención incluye un sistema de IA —por ejemplo, una alerta automática de deterioro— el protocolo debe describir la versión del modelo, entradas, salida, interacción humano-algoritmo, manejo de errores y actualizaciones. SPIRIT-AI y CONSORT-AI, aunque desarrolladas para salud humana, ofrecen principios transferibles de transparencia que deben adaptarse críticamente al contexto veterinario (Cruz Rivera et al., 2020; Liu et al., 2020).

Figura 4.2.

Principales diseños de investigación veterinaria



4.10. Muestreo y representatividad

El muestreo determina a quién representan los resultados. La población objetivo puede ser "perros sometidos a ovariohisterectomía electiva en clínicas urbanas del centro del Ecuador", mientras la población accesible se limita a pacientes de tres clínicas durante seis meses. La muestra analítica será todavía menor después de exclusiones y pérdidas. Cada transición debe documentarse; generalizar desde una muestra de conveniencia a todas las razas, edades, regiones y niveles de atención exige argumentos, no solo un tamaño grande.

El muestreo probabilístico —aleatorio simple, sistemático, estratificado o por conglomerados— permite vincular probabilidades de selección con estimadores poblacionales. En encuestas de sanidad animal, suele ser eficiente seleccionar primero comunidades o fincas y luego animales. El análisis debe utilizar ponderaciones y estructura de conglomerados cuando corresponda. En vigilancia clínica basada en historias electrónicas, el tamaño masivo no elimina el sesgo: quedan fuera animales sin acceso, pruebas no solicitadas y episodios tratados en otros servicios.

Para aprendizaje automático, representatividad significa que las condiciones de entrenamiento cubren la población y el flujo de uso previsto. Un modelo entrenado en radiografías de equipos digitales de un hospital universitario puede degradarse en equipos portátiles, clínicas rurales o

razas braquicéfalas poco representadas. Conviene planificar validación por centro, equipo, periodo, especie, raza, sexo o subgrupo clínico relevante. El desempeño global puede ocultar errores concentrados en grupos pequeños.

4.11. Tamaño muestral para inferencia y predicción

El tamaño muestral inferencial parte de un estimando y un desenlace principal. Para contrastar dos tratamientos deben especificarse el efecto mínimo de interés clínico, la variabilidad esperada, el nivel de significación, la potencia, la razón de asignación y la prueba o modelo previsto. Para estimar una prevalencia importan la precisión deseada y la frecuencia esperada; para supervivencia, el número de eventos; para diseños por conglomerados, el coeficiente de correlación intraclase; para medidas repetidas, la correlación dentro del animal. Piper et al. (2023) recomiendan que la justificación declare de forma transparente supuestos, software, pérdidas previstas y coherencia con el análisis primario.

Una fórmula básica para dos medias independientes con asignación 1:1 es n por grupo $\approx 2(Z\alpha/2 + Z\beta)^2\sigma^2/\Delta^2$, donde Δ es la diferencia mínima clínicamente relevante y σ la desviación estándar. Es una aproximación didáctica: debe ajustarse cuando hay bloques, covariables, no normalidad marcada, medidas repetidas o agrupamiento. Si se esperan pérdidas en proporción r, el total requerido se calcula como n/(1-r), redondeando hacia arriba; sumar simplemente “20 % de n” subestima el total cuando la tasa se aplica al número reclutado.

La predicción plantea otro problema. No basta con “10 eventos por variable” ni con reservar arbitrariamente 20 % de los datos. El tamaño depende de la frecuencia del desenlace, el número y complejidad de parámetros candidatos, la señal esperada, el sobreajuste admisible, la precisión de la calibración y la estabilidad de las predicciones individuales. Los trabajos de Riley et al. (2019, 2020) proponen criterios basados en sobreajuste, estimación de la varianza residual o riesgo global y precisión, en lugar de una regla única.

Separar datos en entrenamiento, validación y prueba reduce la información disponible en cada fase. Con bases pequeñas, la validación cruzada anidada puede aprovechar mejor los datos para desarrollo y ajuste, pero no sustituye una validación externa. En datos agrupados o longitudinales, el remuestreo debe respetar animal, hato, centro y tiempo. El conjunto de prueba final se mantiene

intacto hasta congelar preprocesamiento, variables, hiperparámetros y umbral de decisión. TRIPOD+AI exige describir con transparencia el origen de los datos, participantes, resultados, desarrollo, evaluación y limitaciones del modelo (Collins et al., 2024).

El principio ético no es utilizar el menor número posible de animales a cualquier costo, sino un número científicamente justificado que evite tanto el uso innecesario como un estudio incapaz de responder la pregunta. Un ensayo insuficiente expone animales sin generar evidencia concluyente; uno sobredimensionado puede detectar diferencias triviales y consumir recursos sin beneficio proporcional. ARRIVE 2.0 incorpora la justificación del tamaño muestral entre sus elementos esenciales de reporte (Percie du Sert et al., 2020a, 2020b).

Tabla 4.4.

Componentes mínimos de una justificación muestral

Objetivo	**Información indispensable**	**Indicador central**	**Advertencia**
Comparar medias	Δ clínico, σ, α, potencia, asignación	Diferencia e IC	No usar solo $p < 0,05$
Estimar prevalencia	Prevalencia esperada, precisión, confianza, diseño	IC de la prevalencia	Ajustar por conglomerados
Supervivencia	Hazard ratio o diferencia relevante, eventos, censura	Número de eventos e IC	n total no reemplaza eventos
Desarrollar predicción binaria	Prevalencia, parámetros, R^2 esperado, sobreajuste	Calibración y estabilidad	Evitar reglas EPV rígidas
Validar modelo	Prevalencia, precisión deseada de AUC/calibración	IC de AUC, pendiente y O:E	Requiere datos independientes
Ensayo por conglomerados	Tamaño y número de clústeres, ICC	Efecto con incertidumbre entre clústeres	Muchos animales en pocos clústeres no resuelven el problema

4.12. Confusión, sesgo de selección y fuga de información

La confusión aparece cuando una variable está asociada con la exposición y con el desenlace, sin ser una consecuencia de la exposición, y distorsiona la comparación. Si un tratamiento regenerativo se reserva para perros jóvenes con lesiones menos graves, la edad y gravedad pueden explicar parte de la mejor recuperación. La aleatorización previene confusión conocida y desconocida en expectativa; en estudios observacionales se recurre a restricción, estratificación,

emparejamiento, modelos multivariables, ponderación por puntaje de propensión o métodos causales más avanzados, siempre bajo supuestos explícitos.

No toda variable asociada debe ajustarse. Controlar un mediador puede bloquear parte del efecto que se desea estimar; controlar un colisionador puede crear una asociación espuria. Un diagrama causal dirigido ayuda a expresar supuestos antes del análisis, pero no decide por sí mismo si esos supuestos son verdaderos. Una herramienta de IA puede proponer variables, aunque su conocimiento sea incompleto o confunda asociaciones publicadas con relaciones causales. La decisión corresponde al equipo clínico, epidemiológico y estadístico.

El sesgo de selección surge cuando la inclusión, permanencia o disponibilidad de datos depende de factores relacionados con exposición y desenlace. Estudiar solo perros que regresan a control puede sobrestimar la cicatrización si los tutores con complicaciones acuden a otros centros o si quienes evolucionan bien no vuelven. Excluir expedientes incompletos puede cambiar la población analítica. Se deben comparar incluidos y excluidos, describir el flujo y explorar mecanismos de ausencia.

La fuga de información es el equivalente predictivo de permitir que el modelo acceda a datos que no estarán disponibles en el momento real de uso o a información del conjunto de evaluación. Ejemplos veterinarios: incluir “tratamiento de rescate” para predecir dolor antes de que se administre; normalizar toda la base antes de dividir; seleccionar variables usando el conjunto completo; imputar con estadísticas que incluyen prueba; mantener visitas del mismo animal en particiones distintas; usar diagnósticos registrados después del horizonte de predicción; ajustar hiperparámetros repetidamente mirando el test final.

Tabla 4.5.

Auditoría de fuga de información

Punto de control	**Pregunta de auditoría**	**Acción preventiva**
Tiempo índice	¿La variable existía al momento de predecir?	Definir fecha índice y ventanas de extracción
Identidad	¿El mismo animal/hato aparece en varios conjuntos?	Partición agrupada por la unidad independiente

Preprocesamiento	¿Imputación o escalado vieron datos de prueba?	Ajustar cada paso dentro del pipeline de entrenamiento
Selección de variables	¿Se usó el desenlace completo para seleccionar?	Realizar selección dentro de cada remuestreo
Ajuste	¿El test guio hiperparámetros o umbrales?	Usar validación interna; abrir el test una sola vez
Aumento de datos	¿Derivados de una imagen cruzan particiones?	Aumentar solo después de dividir y dentro de entrenamiento
Validación externa	¿Proviene de otro flujo clínico real?	Reservar centro, periodo o región verdaderamente independiente

4.13. Uso responsable de IA en la planificación experimental

La IA puede asistir en tareas bien delimitadas: convertir un objetivo clínico en una lista preliminar de variables, generar escenarios para simulaciones de potencia, revisar consistencia entre unidad experimental y modelo, producir código reproducible de aleatorización, identificar posibles fuentes de fuga o construir diccionarios de datos. Su valor depende de instrucciones completas, datos ficticios o anonimizados, verificación humana y registro de versiones. No es una autoridad metodológica ni una fuente bibliográfica.

El equipo debe conservar una cadena de responsabilidad. El investigador principal define la pregunta y el estimando; el clínico juzga relevancia y seguridad; el bioestadístico traduce el diseño a supuestos y análisis; el responsable de datos implementa controles; el comité de ética evalúa bienestar, riesgo y consentimiento. Una herramienta generativa puede apoyar, pero no firmar un protocolo, aprobar un comparador, fijar el efecto clínicamente importante ni asumir responsabilidad por un daño.

La información introducida requiere gobernanza. Historias clínicas, imágenes y datos de tutores no deben copiarse en servicios sin contrato, base jurídica y controles adecuados. Deben emplearse identificadores seudonimizados, acceso por roles, trazabilidad y minimización de datos. Para sistemas predictivos se conserva la versión de datos, código, modelo y dependencias; también se documentan modificaciones posteriores. Si el sistema aprende o cambia durante el estudio, el protocolo debe establecer cómo se controlará esa adaptación.

Una salida generativa puede contener referencias inexistentes, fórmulas incorrectas o recomendaciones incompatibles con el bienestar animal. La revisión mínima comprende: comprobar cada cita en una fuente primaria; reproducir cálculos con software estadístico validado; simular la aleatorización; revisar la unidad de asignación; ejecutar pruebas automatizadas del pipeline; y contrastar el plan con guías como PREPARE (Smith et al., 2018), ARRIVE 2.0, VICH GL9 y, cuando corresponda, TRIPOD+AI. La declaración del uso de IA debe especificar herramienta, versión o fecha, finalidad, alcance de la revisión humana y ausencia de delegación de decisiones clínicas.

Prompt científico bien formulado para apoyo metodológico

"Actúa como revisor metodológico. Evalúa el siguiente protocolo ficticio de cicatrización posquirúrgica canina. Identifica: a) unidad experimental y observacional; b) niveles de agrupamiento; c) comparador; d) riesgos de pseudorreplicación; e) posibles variables de bloqueo; f) fuente de confusión; g) estrategia de partición si se desarrolla un modelo predictivo. No calcules tamaño muestral sin que se proporcionen efecto clínicamente relevante, variabilidad, nivel α, potencia y pérdidas. Presenta supuestos y preguntas pendientes. No inventes referencias."

Por qué es adecuado: delimita el rol, usa un escenario sin datos sensibles, exige reconocer incertidumbre, impide una cifra muestral automática y obliga a separar inferencia de predicción.

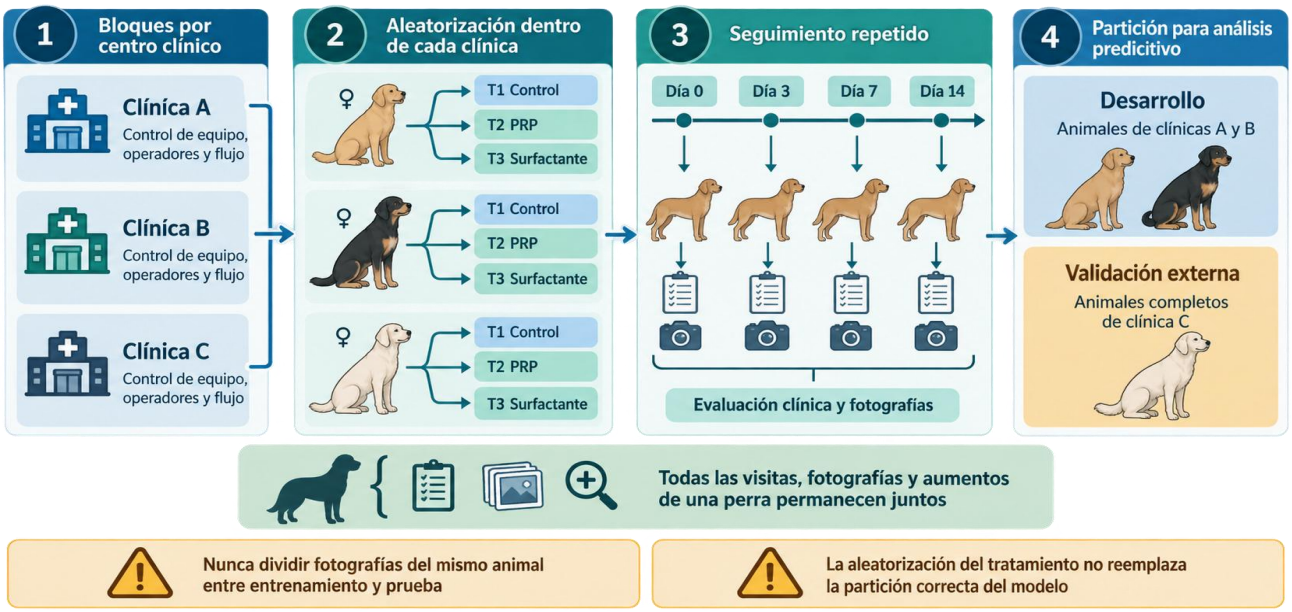
4.14. Caso didáctico: comparación de tratamientos para cicatrización posquirúrgica

Caso didáctico simulado

Todos los datos, tamaños, efectos y resultados de esta sección son ficticios y se presentan exclusivamente con fines de enseñanza. No constituyen evidencia clínica ni recomendación terapéutica.

Figura 4.3.

Aleatorización, bloqueo y partición del caso de cicatrización posquirúrgica



4.14.1. Problema y pregunta científica

Tres clínicas veterinarias desean comparar el cuidado convencional de la herida con un apósito a base de miel médica estandarizada después de ovariohisterectomía electiva en perras clínicamente sanas. La pregunta primaria es: en perras sometidas a un protocolo quirúrgico y analgésico estandarizado, ¿el apósito con miel reduce el tiempo hasta cicatrización clínica completa frente al cuidado convencional? Una pregunta secundaria es si modifica la trayectoria del área de la herida y la incidencia de eventos adversos locales.

La hipótesis confirmatoria se formula sobre un único desenlace primario. H0: la diferencia entre grupos en el tiempo medio hasta cicatrización es cero. H1: el apósito con miel reduce el tiempo en una magnitud clínicamente relevante. El equipo fija prospectivamente como diferencia mínima de interés dos días, no porque garantice significación estadística, sino porque puede reducir visitas, manipulación y carga para tutor y paciente.

4.14.2. Población, unidades y variables

- Población objetivo: perras adultas sometidas a ovariohisterectomía electiva sin enfermedad sistémica relevante.
- Criterios de inclusión: evaluación preoperatoria completa, consentimiento del tutor, posibilidad de controles hasta el día 21.

- Exclusiones preasignación: infección cutánea activa, inmunosupresión, gestación, cirugía concurrente mayor o medicamento que altere cicatrización.
- Unidad experimental: cada perra, porque el tratamiento se asigna individualmente.
- Unidad observacional: la perra/herida en cada visita; las fotografías y tres mediciones digitales repetidas son réplicas técnicas o lecturas derivadas, no unidades independientes.
- Desenlace primario: días desde cirugía hasta cicatrización clínica completa, definida mediante criterios prespecificados y evaluador enmascarado.
- Desenlaces secundarios: porcentaje de reducción del área, puntuación de exudado, dehiscencia, infección, dolor, tratamiento de rescate y evento adverso.

4.14.3. Diseño, asignación y enmascaramiento

Se propone un ensayo multicéntrico, aleatorizado, controlado, de grupos paralelos y superioridad, con asignación 1:1. La aleatorización se estratifica por clínica y utiliza bloques de tamaño variable no revelado. Una persona independiente conserva la secuencia. El cirujano recibe el material después del cierre, pero el evaluador de fotografías desconoce la asignación. Las imágenes se codifican, se toman con distancia, escala, iluminación y orientación estandarizadas, y se evalúan en orden aleatorio.

La clínica se considera factor de diseño porque puede capturar diferencias de operador, ambiente y seguimiento. No se asigna un tratamiento completo a una clínica: hacerlo dejaría tratamiento y clínica confundidos. Se estandarizan técnica de cierre, sutura, analgesia, collar isabelino y recomendaciones domiciliarias. La adherencia al cuidado se registra, pero el análisis primario conserva a los animales según asignación —principio de intención de tratar— con un análisis complementario claramente identificado.

4.14.4. Justificación muestral ilustrativa

Supóngase, solo para el ejercicio, una desviación estándar de 2,5 días, una diferencia mínima de dos días, α bilateral de 0,05 y potencia de 80 %. La aproximación para dos medias independientes produce $n \approx 2(1{,}96 + 0{,}84)^2(2{,}5^2)/(2^2) = 24{,}5$, es decir, 25 perras por grupo. Si se anticipa 15 % de pérdida, se reclutan 25/(1-0,15) = 29,4, redondeado a 30 por grupo: 60 animales en total. En un protocolo real, los supuestos deben proceder de evidencia o datos piloto apropiados y el cálculo debe alinearse con el análisis del desenlace temporal; esta simplificación se utiliza para enseñar los componentes.

El tamaño también debe revisarse respecto a la precisión de los eventos adversos. Con 60 animales, los eventos poco frecuentes se estimarán con gran incertidumbre, por lo que ausencia de eventos observados no demostrará seguridad. Si la intención adicional fuese entrenar un algoritmo de segmentación o predecir retraso de cicatrización, 60 animales probablemente serían insuficientes para un modelo complejo y para una evaluación externa independiente.

4.14.5. Plan de análisis inferencial

El análisis primario estimará la diferencia o razón de tiempos hasta cicatrización con intervalo de confianza del 95 %, utilizando un método compatible con la distribución y la posible censura. Si todas las pacientes completan seguimiento y el desenlace se aproxima a una variable continua sin censura, puede emplearse un modelo lineal ajustado por clínica y valor basal relevante. Si existe censura, Kaplan-Meier y un modelo de Cox pueden ser más apropiados, verificando proporcionalidad. La elección exacta debe fijarse antes de abrir los códigos de grupo.

La trayectoria del área se analizará con un modelo mixto que incluya tratamiento, tiempo, interacción tratamiento × tiempo, clínica y un intercepto aleatorio por perra. La transformación o distribución se elegirá mediante razonamiento sobre la variable y diagnóstico de residuos. No se realizan t de Student separadas en cada día, porque multiplican pruebas e ignoran correlación. Los eventos binarios se presentarán con riesgos absolutos, diferencia de riesgos o razones de riesgo e intervalos de confianza.

La interpretación distingue significación, magnitud y relevancia. Una reducción estimada de 0,6 días con p = 0,03 podría ser estadísticamente compatible con un efecto distinto de cero y, al mismo tiempo, clínicamente modesta. Una reducción de 1,8 días con intervalo de −0,2 a 3,8 sería imprecisa: no autoriza a concluir equivalencia ni ausencia de beneficio. Se reportarán protocolos de rescate, desviaciones, pérdidas y eventos adversos por grupo.

4.14.6. Componente de IA para medición y predicción

Como apoyo a la medición, se plantea un algoritmo de segmentación del área de la herida en fotografías. El algoritmo no reemplaza el desenlace clínico principal hasta demostrar validez. Un conjunto de imágenes se anota por dos evaluadores; se cuantifica acuerdo y se resuelven

discrepancias. La segmentación se evalúa por animal, con métricas como Dice o intersección sobre unión, error absoluto de área y análisis de fallas. Un buen solapamiento promedio puede ocultar errores sistemáticos en pelajes oscuros, iluminación deficiente o heridas pequeñas.

La partición se realiza por perra y por clínica. Todas las fotografías y aumentos derivados de una perra permanecen juntos. Una clínica puede reservarse como validación externa si el número de centros y animales lo permite; con solo tres clínicas, esa validación será preliminar. El preprocesamiento, selección de arquitectura y ajuste de umbrales se realizan sin consultar el conjunto final. Cada versión del algoritmo queda congelada antes de evaluar el test.

Un segundo objetivo exploratorio podría predecir al día 3 qué pacientes tardarán más de 14 días en cicatrizar. Solo se usarían variables disponibles hasta el día 3. Incluir el diagnóstico posterior de infección, el número total de controles o el tratamiento de rescate administrado en el día 7 produciría fuga temporal. Debido al pequeño número de animales, el análisis se limitaría a modelos parsimoniosos y a estimaciones con incertidumbre; no se anunciaría una herramienta clínica lista para uso.

4.14.7. Matriz de riesgos y salvaguardas

Tabla 4.6.

Riesgos metodológicos del caso y medidas de control

Riesgo	**Consecuencia**	**Salvaguarda prospectiva**	**Evidencia de cumplimiento**
Evaluador conoce tratamiento	Sesgo de medición	Codificación y evaluación ciega	Registro de códigos y prueba de cegamiento
Fotos del mismo animal en train/test	Desempeño inflado	Partición por perra antes de aumentar datos	Lista de IDs por partición
Clínica confundida con tratamiento	Efecto no identificable	Aleatorización dentro de clínica	Secuencia estratificada
Pérdidas diferenciales	Sesgo de selección	Seguimiento protocolizado; razones documentadas	Diagrama de flujo

Múltiples desenlaces sin jerarquía	Falsos positivos	Desenlace primario único y plan previo	Protocolo/prerregistro
Algoritmo actualizado durante evaluación	Test contaminado	Congelar versión antes del test	Hash o registro de versión
Diferencia estadística trivial	Sobreinterpretación	Umbral clínico e IC	Discusión clínica explícita

4.14.8. Resultados ficticios para interpretación

Supóngase que se analizan 58 perras: 29 por grupo. El grupo con miel presenta 9,8 días hasta cicatrización y el convencional 11,1; diferencia ajustada −1,3 días, IC 95 %: −2,4 a −0,2. El resultado es compatible con una reducción promedio, pero el intervalo incluye efectos menores que el umbral clínico predefinido de dos días. La conclusión prudente sería que existe evidencia de una reducción modesta, cuya importancia clínica y costo-beneficio requieren valoración adicional.

El modelo mixto indica una interacción tratamiento × tiempo compatible con reducción más rápida del área entre los días 3 y 10, con convergencia posterior. Dos eventos de irritación ocurren en el grupo con miel y uno en el control; la muestra no permite establecer diferencias de seguridad. El segmentador obtiene Dice 0,89 en validación interna y 0,76 en la clínica externa, con peor rendimiento en imágenes con reflejo intenso. La caída externa es más informativa que el valor interno y obliga a mejorar la estandarización antes de utilizar la medición automática como desenlace principal.

4.15. Síntesis, autoevaluación y actividad práctica

4.15.1. Síntesis del capítulo

- La pregunta científica determina el diseño; el algoritmo se selecciona después de fijar estimando, unidades, variables y análisis.
- La unidad experimental recibe independientemente la intervención. Las mediciones repetidas, réplicas técnicas e imágenes derivadas no aumentan el número de unidades biológicas.
- Aleatorización, ocultamiento, bloqueo y enmascaramiento controlan mecanismos de sesgo diferentes y deben describirse operativamente.

- Los diseños factoriales permiten estudiar interacción; las medidas repetidas requieren modelar correlación intraanimal.
- El muestreo define la población representada. Una base grande de conveniencia puede conservar sesgos graves.
- El tamaño muestral inferencial se basa en efecto relevante y precisión o potencia; el predictivo se basa en estabilidad, prevalencia, complejidad, calibración y validación.
- La fuga de información puede ocurrir por identidad, tiempo, preprocesamiento, selección de variables o uso reiterado del conjunto de prueba.
- La IA es útil para simular, auditar y documentar, pero no corrige pseudorreplicación, confusión estructural ni decisiones éticas deficientes.

4.15.2. Autoevaluación

1. En un ensayo de cuatro dietas asignadas a ocho corrales, con 25 cerdos por corral, ¿cuál es la unidad experimental y cuál es el tamaño de replicación por dieta?
2. Explique por qué alternar pacientes entre tratamientos A y B no equivale a aleatorizar.
3. ¿Qué diferencia existe entre ocultamiento de la asignación y enmascaramiento del evaluador?
4. Un laboratorio realiza tres PCR por muestra de 12 gatos. ¿Cuántas réplicas biológicas y técnicas existen?
5. ¿Por qué un único hato tratado y un único hato control no permiten estimar separadamente el efecto del tratamiento?
6. Proponga un factor de bloqueo razonable para un ensayo multicéntrico de analgesia posoperatoria canina y justifique su elección.
7. En un factorial 2 × 2, ¿qué pregunta responde la interacción?
8. ¿Cuándo un conjunto de mediciones en diferentes días corresponde a medidas repetidas y cuándo a grupos independientes?
9. Mencione cuatro componentes necesarios para justificar el tamaño muestral de una comparación confirmatoria de dos medias.
10. ¿Por qué la regla de diez eventos por predictor no debe utilizarse como criterio universal para un modelo predictivo?
11. Dé un ejemplo veterinario de fuga temporal y otro de fuga por identidad.
12. Diferencie confusor, mediador y colisionador en términos de la decisión de ajuste.
13. ¿Qué evidencia solicitaría antes de aceptar un algoritmo de segmentación de heridas como instrumento de medición?
14. Explique por qué una AUC alta no demuestra utilidad clínica ni ausencia de sesgo.
15. Enumere cinco tareas en las que la IA puede apoyar responsablemente la planificación y tres decisiones que no deben delegarse.

4.15.3. Guía breve de respuestas

1. El corral es la unidad experimental; hay dos corrales por dieta. Los 25 cerdos son subunidades observacionales dentro de cada corral.
2. La alternancia es previsible y permite manipulación consciente o inconsciente; no produce una secuencia probabilística con ocultamiento.
3. El ocultamiento evita conocer la próxima asignación antes de incluir al animal; el enmascaramiento evita conocer el tratamiento durante aplicación, evaluación o análisis.
4. Doce réplicas biológicas; tres lecturas técnicas por muestra, 36 lecturas en total.
5. Tratamiento y hato están completamente confundidos; cualquier diferencia puede atribuirse a manejo, genética, ambiente o intervención.
6. El centro clínico, porque resume diferencias de equipo, operadores y flujo; la aleatorización debe realizarse dentro de cada centro.
7. Si el efecto de un factor cambia según el nivel del otro; por ejemplo, si un apósito funciona de forma distinta bajo dos protocolos analgésicos.
8. Son repetidas si pertenecen al mismo animal; son independientes si el procedimiento terminal obliga a usar animales distintos en cada tiempo.
9. Efecto mínimo relevante, variabilidad, α, potencia, asignación y pérdidas; se aceptan cuatro correctamente explicados.
10. Porque ignora prevalencia, complejidad total, señal, sobreajuste, calibración y precisión de predicciones; puede ser insuficiente o innecesariamente restrictiva.
11. Temporal: usar infección diagnosticada al día 7 para predecir al día 3. Identidad: imágenes del mismo perro en entrenamiento y prueba.
12. El confusor puede requerir control; el mediador está en la vía causal y ajustarlo cambia el estimando; el colisionador no debe condicionarse sin justificación.
13. Anotación de referencia, acuerdo entre evaluadores, validación por animal y externa, error de área, análisis de fallas y desempeño por subgrupos.
14. La AUC mide discriminación relativa; no informa por sí sola calibración, valor predictivo en otra prevalencia, consecuencias de errores ni beneficio neto.
15. Apoyo: simulación, código de aleatorización, auditoría de consistencia, diccionario de datos, detección preliminar de fuga. No delegar: decisión ética, comparador clínico, efecto relevante, aprobación o responsabilidad profesional.

4.15.4. Actividad práctica integradora

Título: Auditoría y rediseño de un protocolo veterinario asistido por IA.

Situación inicial: una clínica dispone de 1 200 fotografías de heridas correspondientes a 80 perros, obtenidas durante 15 días. Desea comparar dos apósitos y entrenar una red neuronal para predecir cicatrización tardía. El tratamiento A se aplicó durante los primeros tres meses y el B durante los tres meses siguientes. Las fotografías se dividieron al azar en 70 % entrenamiento y 30 % prueba; la normalización se calculó con todas las imágenes. No se registraron cirujano, gravedad basal ni tratamiento analgésico.

Producto esperado: un informe de 1 500 a 2 000 palabras y un diagrama del flujo de datos que contenga los siguientes elementos:

- Pregunta inferencial y pregunta predictiva redactadas por separado.
- Población objetivo, población accesible, unidad experimental, unidad observacional y niveles de agrupamiento.
- Identificación de al menos seis amenazas: confusión por periodo, dependencia por perro, fuga por partición, fuga por preprocesamiento, variables omitidas y posible deriva temporal.
- Rediseño prospectivo con comparador, aleatorización, bloqueo, enmascaramiento, calendario de visitas y desenlace primario.
- Plan de tamaño muestral: lista de supuestos necesarios para inferencia y para predicción, sin inventar valores.
- Esquema de partición agrupada y temporal; definición de entrenamiento, validación interna y prueba externa.
- Métricas clínicas y predictivas, incluyendo calibración y consecuencias de falsos negativos.
- Registro del uso de IA: prompt, salida, errores encontrados, correcciones humanas y referencias verificadas.

Criterios de evaluación: coherencia entre pregunta y diseño (25 %); identificación de unidades y dependencias (20 %); control de sesgo y fuga (20 %); justificación estadística y predictiva (20 %); ética, reproducibilidad y claridad (15 %).

Lista de verificación para planificar el estudio

Tabla 4.7.

Lista de verificación operativa

Componente	Pregunta de control	Estado
Pregunta	¿Se distingue estimación causal, asociación y predicción?	☐
Estimando	¿Está definido el efecto o parámetro principal?	☐
Unidades	¿La asignación y la observación están identificadas por nivel?	☐
Comparador	¿Responde a la decisión clínica y es ético?	☐
Asignación	¿La secuencia es aleatoria, reproducible y oculta?	☐
Bloqueo	¿Controla una fuente importante sin crear estratos inviables?	☐
Enmascaramiento	¿Se especifica quién, cómo y cuándo?	☐
Muestreo	¿La población y la probabilidad de selección son claras?	☐
Tamaño	¿Se documentan supuestos, precisión/potencia y pérdidas?	☐
Dependencia	¿Se consideran animal, camada, corral, hato, centro y tiempo?	☐
Partición ML	¿Se divide por unidad independiente y se respeta el tiempo?	☐
Pipeline	¿Preprocesamiento y selección ocurren dentro del entrenamiento?	☐
Validación	¿Existe evaluación interna honesta y externa cuando procede?	☐
Ética	¿Hay bienestar, rescate, consentimiento y gobernanza de datos?	☐
Reproducibilidad	¿Protocolo, código, semillas, versiones y desviaciones quedan registrados?	☐

CAPÍTULO 5

BIOESTADÍSTICA INFERENCIA Y CAUSALIDAD EN LA ERA DE LA INTELIGENCIA ARTIFICIAL

Introducción

La bioestadística convierte observaciones clínicas, productivas y epidemiológicas en argumentos cuantitativos que pueden ser examinados, refutados y reproducidos. Su cometido no se limita a calcular valores p. Incluye definir la población de interés, representar la incertidumbre, estimar magnitudes, comprobar la compatibilidad de los datos con modelos explícitos y decidir hasta dónde permiten llegar el diseño y la calidad de la información. En Ciencias Veterinarias, estas decisiones se complican por la diversidad de especies, la dependencia entre animales de un mismo hato o camada, las mediciones repetidas, la censura en estudios de supervivencia y la variación entre clínicas, fincas y laboratorios.

La inteligencia artificial amplía la capacidad para reconocer patrones, combinar muchas variables y construir predicciones. No elimina las exigencias del razonamiento estadístico. Un algoritmo con excelente discriminación puede estar mal calibrado, aprender señales producidas por el hospital en lugar de la enfermedad o fallar al aplicarse en otra población. Tampoco convierte por sí mismo una asociación en causalidad. Para sostener que una intervención produce un cambio se requiere una pregunta causal bien definida, un diseño que controle las rutas de confusión y supuestos defendibles. La calidad del análisis depende primero de cómo se generaron los datos.

Este capítulo conecta las herramientas estadísticas que suelen emplearse en investigación veterinaria con preguntas concretas: comparar tratamientos, estimar riesgos, modelar desenlaces no normales, analizar observaciones agrupadas, estudiar tiempos hasta un evento, describir comunidades microbianas y representar relaciones no lineales. El propósito no es formar operadores de software, sino investigadores capaces de escoger un método por su correspondencia con el problema biológico y de interpretar los resultados mediante estimaciones, intervalos de confianza, tamaños del efecto y relevancia clínica. La recomendación reciente para medicina veterinaria es explícita: la lectura científica debe desplazarse desde la dicotomía significativo/no significativo hacia la magnitud y precisión de los efectos (Weng & Messam, 2025).

Objetivos de aprendizaje

- Distinguir descripción, inferencia, predicción y razonamiento causal en problemas veterinarios.

- Seleccionar resúmenes y gráficos según la escala de medición, la distribución y la estructura jerárquica de los datos.
- Interpretar estimaciones, intervalos de confianza, valores p y tamaños del efecto sin reducir la evidencia a un umbral fijo.
- Elegir entre pruebas t, ANOVA, ANCOVA, métodos no paramétricos, regresiones, modelos generalizados, modelos mixtos y análisis de supervivencia.
- Reconocer cuándo PCA, PERMANOVA y GAM responden mejor a una pregunta multivariada o no lineal.
- Evaluar críticamente modelos de IA según su objetivo predictivo, validación, calibración, transportabilidad y relación con una pregunta causal.

Mapa metodológico del capítulo

Tabla 5.1

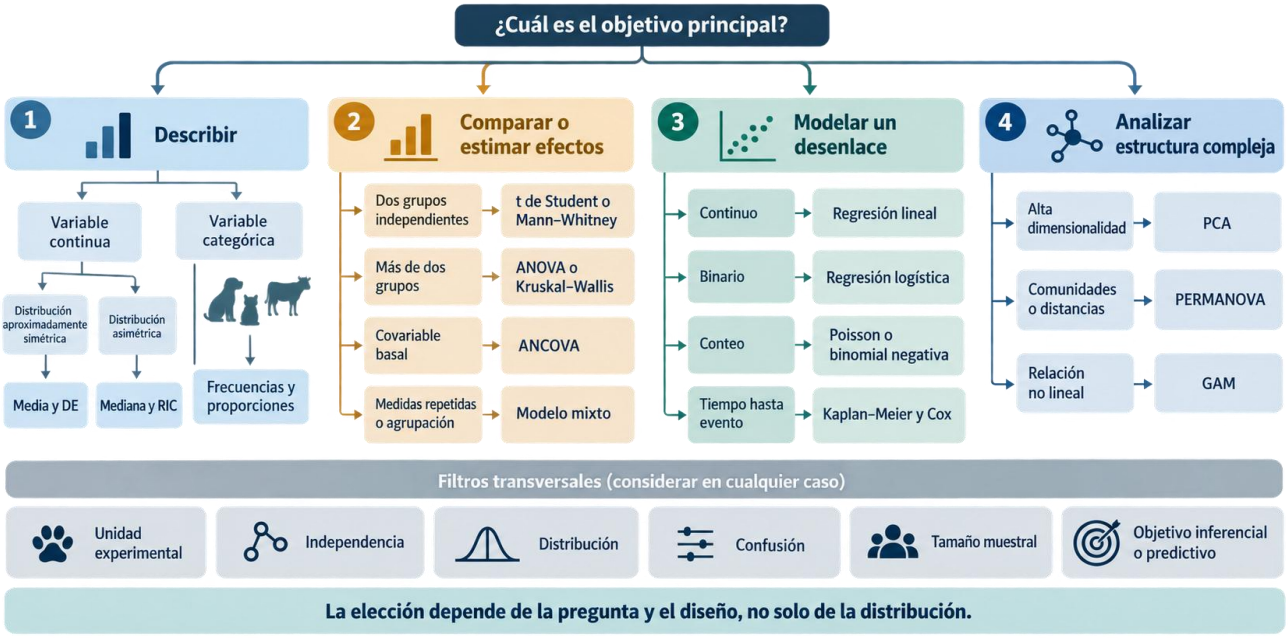
Correspondencia entre pregunta científica y familia de métodos

Pregunta	Desenlace y estructura	Método inicial	Resultado central
¿Difieren dos tratamientos en una medida continua?	Continuo, grupos independientes	t de Student o Welch	Diferencia de medias e IC
¿Difieren tres o más tratamientos?	Continuo, un factor	ANOVA y comparaciones planificadas	Contrastes, IC y tamaño del efecto
¿Difieren grupos ajustando una línea basal?	Continuo con covariable pretratamiento	ANCOVA	Diferencia ajustada
¿Cuál es la probabilidad de enfermedad?	Binario	Regresión logística	Odds ratio y probabilidades
¿Cuántos episodios se esperan?	Conteos; posible sobredispersión	Poisson o binomial negativa	Razón de tasas
¿Cómo cambia un animal a través del tiempo?	Medidas repetidas o agrupadas	Modelo mixto	Efectos fijos y varianza entre unidades
¿Cuánto tiempo transcurre hasta recaída o muerte?	Tiempo con censura	Kaplan-Meier y Cox	Supervivencia, mediana y hazard ratio
¿Difiere una comunidad microbiana?	Matriz multivariada de abundancias	PCA/ordenación y PERMANOVA	Patrones, pseudo-F y R^2
¿La relación es suave y no lineal?	Continuo o conteo con curvatura	GAM	Función suavizada e incertidumbre

¿Se busca pronosticar un nuevo paciente?	Predictores múltiples; validación	Regresión penalizada o aprendizaje automático	Discriminación, calibración y utilidad

Figura 5.1.

Mapa para seleccionar un método estadístico veterinario



5.1 Descripción y exploración de datos veterinarios

La exploración comienza antes de cualquier prueba. El investigador debe verificar qué representa cada fila, cuál es la unidad experimental, qué variable identifica al animal, al tutor, al corral, a la finca o a la clínica, y cómo se codificaron los valores ausentes. Una base con 600 registros no equivale necesariamente a 600 animales independientes: puede contener seis visitas de 100 perros, muestras duplicadas o registros diarios de vacas pertenecientes a pocos hatos. Confundir número de filas con tamaño muestral produce errores estándar demasiado pequeños y conclusiones artificialmente precisas.

Las variables nominales, como especie, sexo, diagnóstico o tratamiento, se resumen mediante frecuencias y proporciones. Las ordinales, como una escala de condición corporal, conservan un orden pero no garantizan distancias iguales entre categorías. Las continuas, como concentración sérica de creatinina, ganancia diaria de peso o diámetro tumoral, admiten medidas de localización

y dispersión. La media y la desviación estándar son informativas cuando la distribución es aproximadamente simétrica y no está dominada por valores extremos. La mediana y el rango intercuartílico describen mejor tiempos de hospitalización, conteos parasitarios o concentraciones con cola derecha.

El gráfico debe mostrar la estructura real de los datos. Un diagrama de cajas aislado oculta el tamaño de la muestra; conviene superponer puntos individuales. Para medidas repetidas se necesitan trayectorias por animal, no barras de medias desconectadas. En estudios multicéntricos, los gráficos estratificados por clínica permiten detectar diferencias de medición o selección. La matriz de datos también debe examinarse por fechas, rangos imposibles, unidades mezcladas y duplicados. Una glucosa registrada como 7 puede representar mmol/L, mientras otra de 126 puede estar en mg/dL. El algoritmo no puede resolver esa ambigüedad sin metadatos.

Ejemplo veterinario. En un estudio de dolor posoperatorio canino se registran puntuaciones Glasgow CMPS-SF antes de la cirugía y a 2, 6, 12 y 24 horas. La exploración adecuada muestra cada trayectoria, el número de perros observados en cada tiempo, la distribución por tratamiento y el patrón de rescate analgésico. Si los perros con dolor intenso reciben rescate y después dejan de aportar puntuaciones comparables, el valor ausente depende del estado clínico. Imputar esos datos como si faltaran al azar puede sesgar la comparación.

La exploración no debe transformarse en una búsqueda ilimitada de resultados favorables. Las decisiones tomadas después de observar los datos se documentan como exploratorias. Las hipótesis, desenlaces principales y contrastes confirmatorios deberían estar definidos en el protocolo. ARRIVE 2.0 insiste en informar diseño, tamaño muestral, criterios de inclusión, aleatorización y análisis con suficiente detalle para evaluar la fiabilidad de la investigación animal (Percie du Sert et al., 2020).

Tabla 5.2

Resúmenes según tipo de variable

Variable veterinaria	Escala	Resumen recomendado	Visualización
Diagnóstico	Nominal	n y porcentaje	Barras con denominador

Condición corporal 1 a 9	Ordinal	Frecuencias; mediana e IQR	Puntos o barras ordenadas
Peso corporal	Continua	Media y DE o mediana e IQR	Histograma y puntos
Huevos por gramo	Conteo	Mediana, IQR, ceros y rango	Puntos en escala apropiada
Tiempo hasta recaída	Tiempo censurado	Kaplan-Meier; mediana si se alcanza	Curva de supervivencia

5.2 Distribuciones variabilidad e incertidumbre

Una distribución describe los valores posibles de una variable y la frecuencia o probabilidad con que aparecen. La distribución observada en una muestra no es idéntica a la distribución de la población; contiene variación biológica, error de medición y azar de muestreo. Separar estas fuentes es esencial. La variación entre bovinos puede reflejar genética, lactancia, dieta y ambiente. La variación dentro de una misma muestra analizada dos veces corresponde principalmente al proceso técnico. Aumentar réplicas técnicas mejora la precisión del laboratorio, pero no sustituye nuevas unidades biológicas.

La distribución normal es útil para errores aproximadamente simétricos y para medias de muchas variables, pero no constituye un requisito universal de los datos crudos. En una regresión lineal o ANOVA, el supuesto concierne a los residuos condicionales, no a cada variable examinada de forma aislada. Los conteos de huevos por gramo suelen mostrar asimetría, exceso de ceros y sobredispersión; una distribución binomial negativa puede ser más coherente que transformar el conteo y aplicar ANOVA. Las proporciones con numerador y denominador conocidos pueden modelarse mediante una distribución binomial. Los tiempos hasta un evento exigen métodos que incorporen censura.

La desviación estándar cuantifica dispersión entre observaciones. El error estándar cuantifica la incertidumbre de un estimador, por ejemplo, la media. Son conceptos distintos: una población heterogénea puede tener gran desviación estándar, mientras una muestra grande permite estimar su media con error estándar pequeño. Presentar media ± error estándar como si describiera variabilidad animal reduce visualmente la heterogeneidad y puede inducir una lectura equivocada.

La incertidumbre no desaparece porque un modelo sea complejo. En aprendizaje automático aparece en la selección de la muestra, la medición de predictores, la partición de datos, el ajuste de hiperparámetros y la población futura. Una única división entrenamiento-prueba puede dar una estimación inestable. La validación cruzada repetida ayuda a medir esa variación durante el desarrollo, pero la evidencia más exigente proviene de una validación externa temporal o geográfica. TRIPOD+AI solicita describir con transparencia los datos, el desarrollo, la evaluación y las limitaciones de modelos que utilizan regresión o aprendizaje automático (Collins et al., 2024).

5.3 Estimación e intervalos de confianza

La estimación responde cuánto. Puede referirse a la diferencia media de dolor entre tratamientos, la razón de riesgos de infección, la razón de tasas de mastitis o el cambio esperado en producción por cada grado de temperatura. Una estimación puntual resume la mejor aproximación obtenida con la muestra. El intervalo de confianza añade el rango de valores compatibles con los datos y el modelo bajo repeticiones hipotéticas del procedimiento.

Un intervalo de confianza del 95 % no significa que exista 95 % de probabilidad de que el parámetro fijo esté dentro del intervalo ya calculado. La interpretación frecuentista es procedimental: si se repitiera el estudio y se construyeran intervalos del mismo modo, aproximadamente 95 % cubrirían el verdadero parámetro cuando se cumplen los supuestos. En la práctica científica, el intervalo permite valorar precisión y compatibilidad. Un intervalo ancho revela que los datos no distinguen con claridad entre beneficios pequeños, efectos clínicamente relevantes o incluso daño.

Para una diferencia de medias se puede expresar la estructura general como:

$$\textit{estimación} \pm \textit{valor crítico} \times \textit{error estándar}$$

La fórmula exacta cambia según el diseño y la distribución. En grupos independientes con varianzas desiguales, el intervalo de Welch es preferible al intervalo que impone varianzas iguales. En datos agrupados, el error estándar debe derivarse de un modelo que reconozca la correlación. En una proporción pequeña o cercana a uno, los intervalos de Wald pueden comportarse mal; son preferibles procedimientos binomiales con mejor cobertura.

Ejemplo. Una dieta renal reduce el fósforo sérico medio en gatos en 0,42 mg/dL frente al control, con IC 95 % de −0,78 a −0,06. Los datos son compatibles con una reducción pequeña o moderada, pero no establecen automáticamente que el cambio mejore supervivencia o calidad de vida. Si el umbral mínimo clínicamente relevante se definió en 0,50 mg/dL, el intervalo cruza ese umbral: persiste incertidumbre sobre la importancia clínica aunque no cruce cero. Weng y Messam (2025) sostienen precisamente que las estimaciones y sus intervalos permiten una interpretación veterinaria más útil que la clasificación aislada por significación estadística.

5.4 Pruebas de hipótesis y valor p

Una prueba de hipótesis examina la compatibilidad de los datos con un modelo nulo. El valor p es la probabilidad, calculada bajo la hipótesis nula y los supuestos del análisis, de obtener un resultado al menos tan extremo como el observado. No es la probabilidad de que la hipótesis nula sea verdadera, no mide la magnitud del efecto y no informa por sí solo sobre la calidad del estudio. La declaración de la American Statistical Association advirtió que las conclusiones científicas no deben basarse únicamente en que un valor p cruce un umbral (Wasserstein & Lazar, 2016).

El umbral de 0,05 no crea dos realidades. Los resultados $p = 0,049$ y $p = 0,051$ aportan evidencia casi idéntica, aunque una práctica dicotómica los rotule de manera opuesta. La interpretación debe integrar diseño, estimación, intervalo de confianza, plausibilidad biológica, multiplicidad, calidad de medición y evidencia previa. La expresión $p < 0,001$ es preferible a $p = 0,000$; una probabilidad calculada no es exactamente cero.

La potencia se refiere a la probabilidad de detectar un efecto especificado cuando existe, bajo un diseño y nivel de significación dados. No debe usarse la potencia observada después del análisis para explicar un resultado no significativo, porque es una transformación del propio valor p. Es más informativo examinar el intervalo de confianza: ¿descarta efectos relevantes o sigue siendo demasiado amplio? La revisión de Wang et al. (2025) encontró que solo 24 de 98 estudios clínicos veterinarios evaluados informaron con detalle el cálculo del tamaño muestral, señal de una debilidad todavía común.

La multiplicidad surge cuando se prueban muchos desenlaces, momentos, subgrupos o variables. Si se ejecutan veinte pruebas nulas independientes con $\alpha = 0,05$, la probabilidad de al menos un

falso positivo excede 0,60. El protocolo debe identificar el desenlace principal, limitar comparaciones confirmatorias y definir el ajuste apropiado. Bonferroni controla con severidad el error familiar; Holm suele ser menos conservador; el control de la tasa de falsos descubrimientos puede ser razonable en estudios ómicos exploratorios. Ningún ajuste reemplaza una pregunta bien delimitada.

5.5 Tamaño del efecto y relevancia clínica

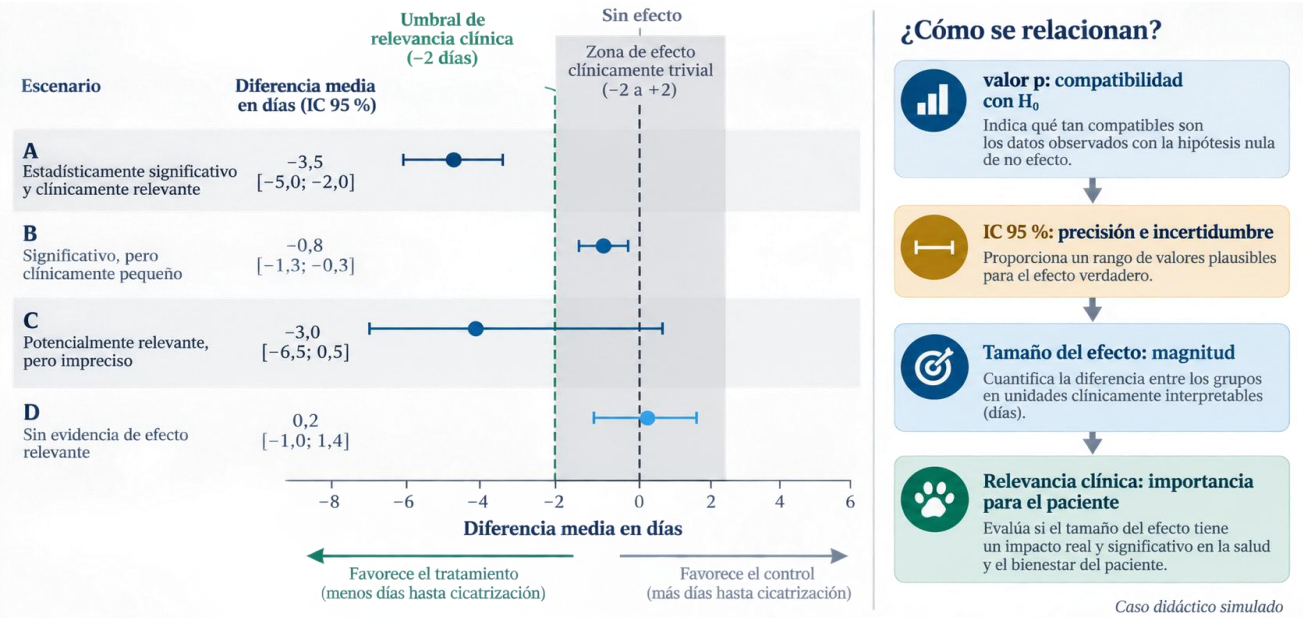
El tamaño del efecto expresa magnitud. La elección depende de la pregunta: diferencia de medias en unidades originales, diferencia estandarizada, razón de riesgos, odds ratio, razón de tasas, hazard ratio, R^2 parcial o medidas de asociación. Siempre que sea posible, conviene priorizar unidades clínicas. Decir que un tratamiento reduce dos días la cicatrización es más interpretable que informar únicamente d = 0,65.

La diferencia estandarizada de medias facilita comparaciones entre escalas, pero depende de la variabilidad de la muestra. Dos estudios con la misma diferencia clínica pueden producir d distintos si sus desviaciones estándar difieren. Para desenlaces binarios, el riesgo relativo suele comunicarse con más facilidad que el odds ratio. Cuando el evento es frecuente, el odds ratio se aleja del riesgo relativo y puede parecer más grande. La diferencia absoluta de riesgos permite calcular el número necesario a tratar cuando el contexto y el horizonte temporal están claramente definidos.

La relevancia clínica exige un umbral formulado antes de conocer los resultados. En un ensayo analgésico, podría ser la reducción mínima en una escala validada o la menor proporción de animales que necesita rescate. En producción, una diferencia estadísticamente detectable de 0,15 litros de leche por día puede carecer de valor económico después de considerar costos. En epidemiología, un efecto pequeño puede ser relevante si la exposición es común y el desenlace grave.

Figura 5.2.

Relación entre valor p, intervalo de confianza, tamaño del efecto y relevancia clínica



La importancia biológica tampoco equivale a significación. Un intervalo que incluye tanto ausencia de efecto como un beneficio clínicamente relevante indica inconclusión, no equivalencia. Para afirmar equivalencia o no inferioridad se requiere un margen justificado y un diseño específico. Las guías veterinarias recientes recomiendan informar estimaciones, precisión y supuestos de manera suficiente para reproducir el análisis (Villarino, 2025).

Tabla 5.3

Medidas de efecto y lectura veterinaria

Medida	Pregunta	Lectura recomendada	Advertencia
Diferencia de medias	¿Cuánto cambia una medida continua?	Unidades originales e IC	No confundir con d estandarizada
Riesgo relativo	¿Cuántas veces cambia el riesgo?	RR y riesgo basal	Necesita horizonte temporal
Diferencia de riesgos	¿Cuántos eventos cambian por 100 animales?	Puntos porcentuales e IC	Depende del riesgo basal
Odds ratio	¿Cómo cambian las odds?	OR y probabilidades predichas	No es RR si el evento es frecuente
Hazard ratio	¿Cómo cambia la tasa instantánea?	HR e IC; curvas	No es razón de medianas

5.6 Pruebas t y análisis de varianza

La prueba t para muestras independientes compara dos medias cuando las observaciones son independientes y los residuos son aproximadamente normales. La versión de Welch no requiere igualdad de varianzas y suele ser una opción inicial segura. La prueba t pareada se aplica cuando dos mediciones pertenecen a la misma unidad o a pares emparejados; analiza las diferencias dentro de cada par. Aplicar una t independiente a medidas antes y después desperdicia el emparejamiento y estima mal la incertidumbre.

El ANOVA de un factor extiende la comparación a tres o más grupos. Su prueba F evalúa si la variación entre medias es grande en relación con la variación residual. Un resultado global no identifica cuáles tratamientos difieren. Las comparaciones deben corresponder a hipótesis definidas: contrastes planificados si interesan preguntas específicas o procedimientos como Tukey para todas las comparaciones por pares. Ejecutar múltiples pruebas t aumenta el error de tipo I.

Los supuestos se evalúan sobre el modelo. La independencia depende del diseño y no puede diagnosticarse con un gráfico. La normalidad de residuos merece atención especial en muestras pequeñas; con muestras moderadas, el ANOVA puede tolerar desviaciones leves. La homogeneidad de varianzas puede examinarse gráficamente y mediante conocimiento del proceso. Si las varianzas difieren, Welch-ANOVA es una alternativa. Una transformación puede facilitar el modelado, pero cambia la escala y requiere interpretación cuidadosa.

Ejemplo simulado. En 36 perros sometidos a ovariohisterectomía se comparan tres protocolos de cuidado de herida, con 12 animales por grupo. El desenlace es el número de días hasta epitelización. El ANOVA global responde si alguna media difiere; Tukey permite comparaciones por pares. Deben informarse medias, desviaciones estándar, diferencias con IC, diagnóstico de residuos y el criterio clínico. Si la fecha de cirugía se organiza por jornadas y los tratamientos se aplican dentro de cada jornada, esta última puede actuar como bloque. Ignorarla pierde precisión y puede confundir el efecto del tratamiento.

Tabla 5.4

Selección entre pruebas de comparación

Diseño	Método	Estimando principal	Comprobación crítica
Dos grupos independientes	t de Welch	Diferencia de medias	Independencia y residuos
Antes y después en los mismos animales	t pareada	Media de diferencias	Distribución de diferencias
Tres o más grupos	ANOVA	Contrastes de medias	Varianza y multiplicidad
Grupos con línea basal	ANCOVA	Diferencia ajustada	Forma funcional e interacción
Ordinal o rangos	Método no paramétrico	Diferencia distributiva	Pregunta estimada e independencia

5.7 ANCOVA y control estadístico de covariables

El análisis de covarianza combina comparación de grupos y regresión. Resulta especialmente útil cuando existe una medición basal del desenlace. En un ensayo sobre movilidad canina, ajustar la puntuación final por la puntuación inicial suele ser más eficiente que analizar el cambio bruto y evita que desequilibrios basales accidentales distorsionen la comparación.

El modelo básico incluye tratamiento, valor basal y covariables pretratamiento justificadas. La interpretación del coeficiente de tratamiento es una diferencia ajustada entre grupos para animales comparables en las covariables. Debe comprobarse la relación funcional, la linealidad cuando corresponde y la posible interacción tratamiento × basal. Si las pendientes difieren, una única diferencia ajustada puede ser engañosa y el efecto debe expresarse para valores basales relevantes.

No toda variable mejora el ajuste. Una covariable debe seleccionarse por el protocolo, la teoría causal y su relación con el desenlace, no porque alcance $p < 0{,}05$ en análisis univariados. Ajustar por una variable posterior al tratamiento puede bloquear parte del efecto que se desea estimar o introducir sesgo de colisionador. Por ejemplo, al estimar el efecto total de una terapia sobre recuperación, ajustar por una respuesta inflamatoria medida después de iniciar la terapia podría eliminar un mecanismo mediador.

En estudios observacionales, el término control estadístico puede generar falsa seguridad. Ajustar variables medidas no resuelve confusión no medida, errores de medición ni selección. Un modelo

con muchas covariables tampoco transforma datos observacionales en un ensayo. La elección debe apoyarse en un diagrama causal y una definición explícita del efecto objetivo.

5.8 Métodos no paramétricos

Los métodos no paramétricos son útiles cuando el parámetro de interés, la escala ordinal o la forma de la distribución no se adaptan a un modelo paramétrico sencillo. Mann-Whitney compara distribuciones mediante rangos; no siempre prueba una diferencia de medianas. Esa interpretación solo es defendible bajo condiciones adicionales, como formas y dispersiones similares. Wilcoxon con signo utiliza diferencias pareadas. Kruskal-Wallis extiende la comparación a más de dos grupos y Friedman se aplica a bloques o medidas repetidas simples.

Estos métodos no son métodos sin supuestos. Requieren independencia en las estructuras correspondientes y pueden responder una pregunta diferente a la comparación de medias. Además, pierden flexibilidad ante covariables, interacciones y jerarquías complejas. Para un conteo sesgado quizá sea preferible un modelo binomial negativo; para un desenlace ordinal, una regresión ordinal; para medidas repetidas, un modelo mixto generalizado.

Las pruebas de permutación construyen una distribución nula mediante reorganizaciones compatibles con el diseño. El esquema de permutación debe respetar bloques, pares o clusters. Permutar libremente animales de diferentes fincas cuando el tratamiento se asignó a la finca rompe la estructura experimental. El bootstrap, por su parte, estima incertidumbre mediante remuestreo. Si hay agrupación, el remuestreo debe ocurrir al nivel independiente pertinente.

Ejemplo. Se comparan apósitos convencionales y miel médica en 20 perros, con días hasta cicatrización muy asimétricos. Mann-Whitney puede ser razonable para una comparación exploratoria si los animales son independientes. El informe debería presentar mediana, rango intercuartílico, una medida del efecto basada en probabilidades o diferencia de localización e intervalo obtenido mediante un procedimiento apropiado. Informar solo U y p deja sin respuesta cuánto mejora el tiempo de cicatrización.

5.9 Regresión lineal y logística

La regresión lineal modela el valor medio de un desenlace continúo condicionado a predictores. Su coeficiente representa el cambio medio esperado por unidad de predictor, manteniendo constantes las demás variables del modelo. Esa frase no implica causalidad. La interpretación causal requiere que el conjunto de ajuste cierre las rutas de confusión y que se satisfagan supuestos adicionales.

El diagnóstico incluye residuos frente a valores ajustados, observaciones influyentes, forma funcional y colinealidad. Una asociación curvada no debe forzarse a una recta. La inclusión automática de todas las variables disponibles puede producir coeficientes inestables, sobreajuste y explicaciones contradictorias. La selección debe responder a un modelo científico; si el objetivo es predicción, se evalúa el rendimiento fuera de muestra y puede recurrirse a penalización.

La regresión logística modela el logit de la probabilidad de un desenlace binario. Al exponentiar un coeficiente se obtiene un odds ratio ajustado. Un odds ratio de 2 no significa que la probabilidad se duplique. Por ejemplo, pasar de 0,10 a odds duplicados produce una probabilidad cercana a 0,18; pasar de 0,50 a odds duplicados produce aproximadamente 0,67. Por ello es recomendable complementar los odds ratios con probabilidades predichas o efectos marginales para perfiles clínicos plausibles.

En un modelo de mastitis subclínica bovina, el desenlace puede definirse por cultivo y recuento celular. Predictores como número de lactancia, días en leche y producción deben medirse antes o en el momento de la decisión. Si se incluye un resultado de laboratorio disponible después del diagnóstico, la aparente precisión puede provenir de fuga de información. Para evaluación predictiva se requieren discriminación, calibración y validación externa; el área bajo la curva ROC no basta.

5.10 Modelos lineales generalizados

Los modelos lineales generalizados relacionan la media del desenlace con una combinación lineal de predictores mediante una función de enlace y especifican una distribución compatible con la

variable. La regresión lineal, logística y de Poisson pertenecen a esta familia. Elegir distribución y enlace por costumbre, sin examinar el mecanismo de datos, conduce a errores.

Para conteos, Poisson supone igualdad entre media y varianza condicional. Los datos veterinarios suelen presentar sobredispersión por heterogeneidad no observada; la binomial negativa incorpora un parámetro adicional de dispersión. Un modelo con enlace log produce razones de tasas. Si el tiempo de exposición varía entre animales, se incluye un offset. Contar episodios de diarrea sin considerar días de observación confunde frecuencia y seguimiento.

Para proporciones derivadas de ensayos binarios se utiliza la distribución binomial con numerador y denominador. Analizar porcentajes mediante ANOVA puede asignar la misma precisión a 2/4 y 50/100, aunque la segunda proporción contiene más información. Los modelos Gamma con enlace log son útiles para variables continuas positivas y sesgadas, como costos o ciertas concentraciones, siempre que no haya ceros.

El exceso de ceros puede surgir de procesos distintos. En parasitología, algunos animales carecen realmente de infección y otros están infectados pero la prueba no detecta huevos. Los modelos hurdle o cero-inflados formalizan dos procesos, aunque requieren una justificación biológica y suficiente información. La flexibilidad no debe sustituir una definición cuidadosa del desenlace ni una evaluación de sensibilidad.

Tabla 5.5

Distribuciones para desenlaces frecuentes

Desenlace	**Distribución**	**Enlace usual**	**Ejemplo**
Continuo aproximadamente simétrico	Normal	Identidad	Temperatura corporal
Binario	Binomial	Logit	Mastitis sí o no
Conteo	Poisson	Log	Número de episodios
Conteo sobredisperso	Binomial negativa	Log	Huevos por gramo
Continuo positivo sesgado	Gamma	Log	Costo de hospitalización

5.11 Modelos mixtos y datos agrupados

Los datos veterinarios rara vez son completamente independientes. Lechones comparten camada, vacas comparten hato, perros reciben atención en clínicas y cada animal puede aportar varias observaciones. Los modelos mixtos incluyen efectos fijos, que representan asociaciones o contrastes de interés, y efectos aleatorios, que representan variación entre unidades de agrupación.

Un intercepto aleatorio por animal permite que cada sujeto tenga su propio nivel basal y modela la correlación de medidas repetidas. Una pendiente aleatoria permite que las trayectorias cambien entre animales. En un estudio multicéntrico puede añadirse clínica como intercepto aleatorio, si el número de centros y el objetivo justifican tratarla como muestra de una población de centros. Con muy pocos clusters, la estimación de varianzas y errores estándar puede ser inestable.

El coeficiente de correlación intraclase cuantifica qué proporción de la variación corresponde a diferencias entre clusters. Cuando es distinto de cero, el tamaño muestral efectivo es menor que el número total de observaciones. Esta pérdida depende también del tamaño del cluster. Cientos de mediciones provenientes de pocas fincas no ofrecen la misma evidencia que cientos de fincas.

Los efectos aleatorios no reparan una pseudorreplicación del tratamiento. Si una dieta se asigna a dos galpones y cada galpón recibe una sola dieta, el número de unidades experimentales para el tratamiento es dos, aunque se midan 500 aves. El modelo debe reflejar el nivel de asignación y el diseño necesita replicación a ese nivel. En modelos mixtos generalizados, la interpretación condicional de coeficientes puede diferir de los efectos promedio poblacionales; el informe debe especificar cuál se presenta.

5.12 Análisis de supervivencia Kaplan Meier y modelo de Cox

El análisis de supervivencia estudia el tiempo hasta un evento. El evento debe definirse sin ambigüedad: muerte por cualquier causa, muerte atribuida a enfermedad, recaída, alta o fracaso terapéutico. La fecha de origen también debe ser común y clínicamente defendible, por ejemplo, cirugía, diagnóstico o inicio de tratamiento. Mezclar orígenes crea sesgo de tiempo inmortal.

La censura ocurre cuando el evento no se observa durante el seguimiento. Un animal vivo al cierre o perdido antes del evento aporta información hasta su última observación. Kaplan-Meier estima la función de supervivencia mediante probabilidades condicionales en cada tiempo de evento. La mediana es el tiempo en el que la supervivencia estimada alcanza 0,50; puede no estimarse si más de la mitad permanece sin evento.

La prueba log-rank compara globalmente curvas bajo una ponderación que funciona mejor cuando los riesgos son aproximadamente proporcionales. No cuantifica por sí sola la magnitud. El modelo de Cox estima hazard ratios sin especificar la forma del riesgo basal. Un HR de 0,50 indica que la tasa instantánea del evento, entre sujetos todavía en riesgo y bajo los supuestos del modelo, es la mitad; no implica que la mitad de los animales morirá ni que la supervivencia mediana se duplique.

El supuesto de riesgos proporcionales debe evaluarse con residuos de Schoenfeld, gráficos y razonamiento clínico. Si el efecto cambia con el tiempo, pueden usarse interacciones con tiempo, modelos estratificados o medidas como la diferencia en tiempo medio restringido. La censura debe ser no informativa condicionada a las variables del modelo. Si los tutores retiran con mayor frecuencia a animales que empeoran, esta condición puede fallar.

En oncología veterinaria también importan los riesgos competitivos. Un perro puede morir por una causa no relacionada antes de recaer. Tratar esa muerte como censura simple puede sobreestimar la incidencia acumulada de recaída. La elección entre supervivencia global, específica y análisis de riesgos competitivos debe corresponder a la pregunta y a la confiabilidad con que se adjudicó la causa.

5.13 PCA PERMANOVA y análisis multivariado

El análisis de componentes principales reduce un conjunto de variables continuas correlacionadas a componentes ortogonales que explican sucesivamente la mayor varianza. Requiere decidir si las variables se centran y escalan. En perfiles bioquímicos con unidades distintas, estandarizar suele evitar que una variable de gran escala domine. Las cargas indican cómo contribuye cada variable y las puntuaciones ubican cada animal en el espacio reducido.

PCA es descriptivo, lineal y sensible a valores extremos. No confirma grupos ni identifica causas. Si se observa separación entre tratamientos, debe comprobarse si proviene de lote, centro, edad u otra fuente. En datos composicionales, como abundancias relativas del microbioma, la geometría euclidiana sobre porcentajes crudos puede ser inadecuada; se requieren transformaciones y distancias acordes con la pregunta.

PERMANOVA compara centroides multivariados mediante una matriz de distancias y permutaciones. El resultado incluye un pseudo-F, un R^2 que expresa la proporción de variación asociada al factor y un valor p derivado de permutaciones. Una diferencia significativa puede reflejar cambios de localización, dispersión o ambos; por eso debe evaluarse la homogeneidad de dispersiones. El esquema de permutación debe respetar pareamiento, bloques y medidas repetidas.

Ejemplo simulado. Se comparan comunidades bacterianas fecales en perros que reciben una dieta con orégano y control, con mediciones antes y después. Una PERMANOVA que permuta libremente las 80 muestras ignora que dos muestras pertenecen a cada perro. La permutación debe restringirse por animal o el análisis debe modelar el cambio. El informe incluye distancia utilizada, transformación, número de permutaciones, R^2, evaluación de dispersión y una ordenación para visualizar, sin interpretar los ejes como prueba causal.

5.14 Modelos aditivos generalizados

Los modelos aditivos generalizados representan relaciones no lineales mediante funciones suaves estimadas de los datos. Su forma general combina un intercepto, efectos lineales y términos f(x). El grado de flexibilidad se controla por penalización para evitar curvas que sigan el ruido. El número efectivo de grados de libertad resume la complejidad de cada término: un valor cercano a uno sugiere una relación casi lineal.

Un GAM es útil cuando la biología sugiere curvatura pero no una forma paramétrica simple. La producción de leche puede descender poco a temperaturas moderadas y caer con rapidez después de un umbral térmico. Imponer una recta distribuye el descenso a lo largo de todo el rango y oculta la zona crítica. La curva debe acompañarse de bandas de incertidumbre y densidad de observaciones; los extremos con pocos datos son especialmente inciertos.

La autocorrelación temporal, la estacionalidad y la agrupación siguen siendo relevantes. En registros diarios de vacas, un suavizado de temperatura sin efecto por animal ni estructura temporal puede atribuir a la temperatura patrones de lactancia o estación. Los GAM mixtos permiten términos aleatorios; también pueden incluir suavizados cíclicos para mes o día del año.

La interpretación causal de una curva requiere más que buen ajuste. Temperatura, humedad, manejo y etapa de lactancia pueden estar relacionadas. El modelo estima una asociación condicional a las variables incluidas. Para predicción, el rendimiento debe evaluarse fuera de muestra y dentro del rango de covariables; extrapolar una función suave a temperaturas no observadas carece de soporte.

5.15 Predicción frente a explicación causal

Predicción y causalidad plantean preguntas distintas. La predicción estima un desenlace para un nuevo animal: ¿qué probabilidad tiene este gato de ser hospitalizado en siete días? La explicación causal pregunta qué ocurriría bajo una intervención. Un predictor puede ser útil sin ser causa. La clínica de origen o el modelo del equipo de imagen pueden anticipar un desenlace porque condensan información del sistema, pero intervenir sobre ellos no necesariamente modifica la enfermedad.

La inferencia causal exige definir exposición, comparador, población, desenlace y horizonte temporal. El ensayo aleatorizado controla, en expectativa, confusores conocidos y desconocidos. Con datos observacionales se requieren consistencia, intercambiabilidad condicional, positividad y medición adecuada. Un diagrama causal ayuda a decidir qué variables ajustar. Hernán y Robins (2020) sostienen que la pregunta debe formularse como comparación entre intervenciones bien definidas.

La IA puede estimar funciones complejas dentro de un análisis causal, pero no identifica el efecto por sí sola. Si ninguna clínica rural utiliza el nuevo tratamiento, un algoritmo extrapola sin apoyo empírico. Si el motivo de prescripción no está registrado, persiste confusión por indicación. En predicción, la validación debe separar animales, no filas; una validación temporal aproxima la implementación futura y una externa examina transportabilidad.

La explicación de un modelo predictivo tampoco equivale a causalidad. SHAP atribuye la predicción dentro del modelo, pero no demuestra el resultado de modificar una variable. La importancia de raza puede reflejar prevalencia, selección o prácticas diagnósticas. TRIPOD+AI reúne requisitos para modelos de regresión y aprendizaje automático, de modo que ambos deben informar datos, evaluación, limitaciones y validación con el mismo rigor (Collins et al., 2024).

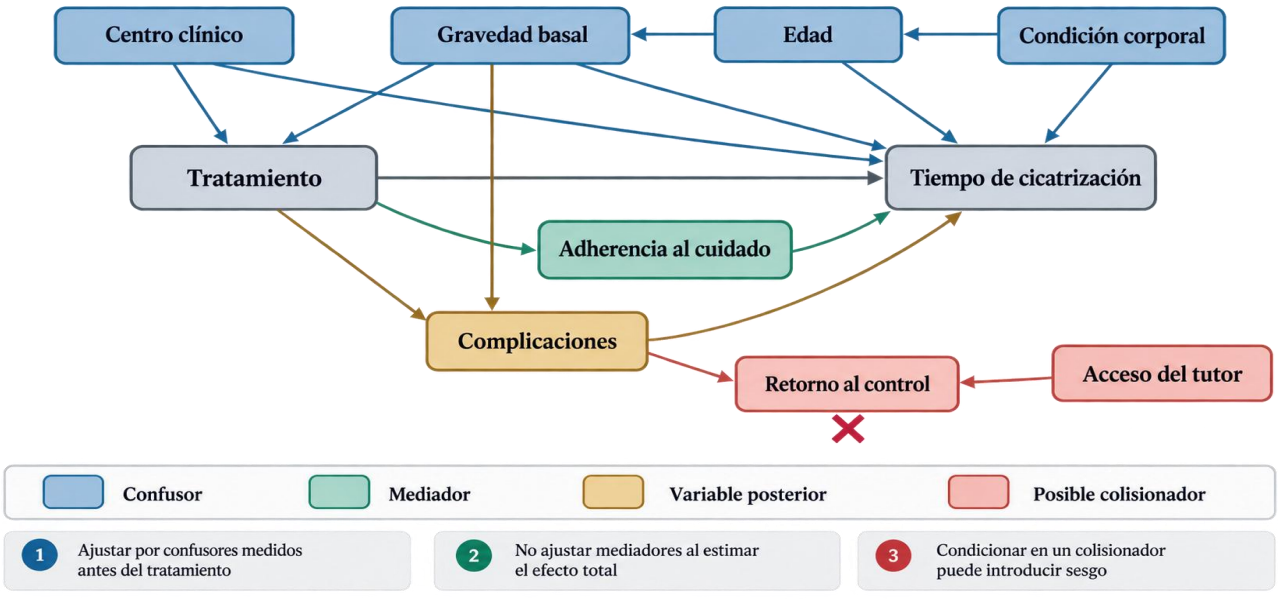
Tabla 5.6

Diferencias entre objetivos analíticos

Dimensión	Predicción	Inferencia asociativa	Inferencia causal
Pregunta	¿Qué ocurrirá?	¿Cómo se relacionan X e Y?	¿Qué ocurriría si se interviniera?
Criterio	Rendimiento fuera de muestra	Estimación condicional	Identificación del efecto
Variables	Pueden ser marcadores	Según modelo	Según estructura causal
Validación	Temporal o externa	Diagnóstico y sensibilidad	Diseño y supuestos causales
Error típico	Fuga y mala calibración	Sobreajuste interpretativo	Confusión y selección

Figura 5.3.

Diagrama causal de tratamiento y cicatrización



5.16 Caso didáctico supervivencia en oncología veterinaria

Caso didáctico simulado. Ningún dato de esta sección corresponde a pacientes reales. El ejercicio se construye para integrar estimación, censura, Kaplan-Meier, Cox, interpretación clínica y contraste con un objetivo predictivo.

Problema científico. Se desea evaluar si la administración adyuvante de vinblastina y prednisolona después de la resección de un mastocitoma cutáneo canino de grado II se asocia con mayor supervivencia libre de progresión que la observación posquirúrgica. La pregunta causal sería: en perros elegibles y sometidos a resección, ¿cuál sería la diferencia en supervivencia libre de progresión a 12 meses si todos recibieran el protocolo frente a si todos permanecieran en observación? La pregunta predictiva sería distinta: ¿qué probabilidad de progresión tiene un perro individual en 12 meses, dados tratamiento, estadio, índice mitótico y estado de márgenes?

Población y diseño. Se simula una cohorte retrospectiva de 48 perros atendidos en tres clínicas. Veintiséis recibieron vinblastina-prednisolona y 22 permanecieron en observación. El origen temporal es la fecha de cirugía. El evento es recurrencia local, metástasis documentada o muerte atribuida al mastocitoma, lo que ocurra primero. Los perros sin evento se censuran en la última visita. La unidad observacional y analítica es el perro; las clínicas constituyen una posible agrupación.

Variables. El tratamiento es la exposición principal. Se registran edad, sexo, tamaño tumoral, estadio clínico, índice mitótico y estado de márgenes. La variable de respuesta consta de tiempo en días y un indicador de evento. El índice mitótico y el estadio se consideran factores pronósticos anteriores a la elección final del tratamiento. Sin embargo, en una cohorte retrospectiva la indicación clínica puede depender de gravedad no registrada; el ajuste no garantiza intercambiabilidad.

Resultados simulados. Ocurrieron 10 eventos en el grupo tratado y 14 en observación. La mediana Kaplan-Meier fue de 510 días para tratamiento y 270 días para observación. La prueba log-rank produjo $\chi^2 = 5{,}83$ y $p = 0{,}016$. En un modelo de Cox ajustado por estadio, índice mitótico y márgenes, el hazard ratio del tratamiento fue 0,48, con IC 95 % de 0,26 a 0,89. El intervalo sugiere

una reducción compatible de la tasa instantánea entre 11 % y 74 %, pero no demuestra que el protocolo cause esa reducción.

Interpretación. El valor p no resume la evidencia. Las curvas muestran cuándo aparecen las diferencias, el número de perros en riesgo informa cuánto soporte queda en cada tiempo y el IC del HR expresa precisión. La mediana ofrece una medida temporal intuitiva, aunque no utiliza toda la forma de la curva. La interpretación clínica debe considerar toxicidad, calidad de vida, visitas, costos y la posibilidad de confusión por indicación. El tamaño pequeño limita interacciones y validación de supuestos.

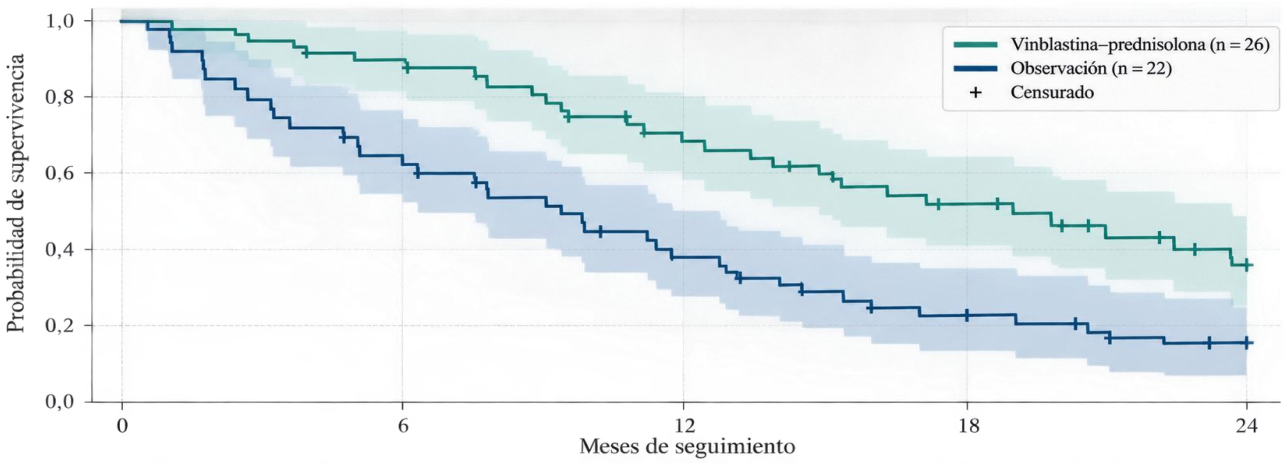
Tabla 5.7

Resumen de resultados del caso simulado

Indicador	**Tratamiento n 26**	**Observación n 22**	**Interpretación**
Eventos	10	14	No comparar porcentajes sin incorporar tiempo
Censurados	16	8	Aportan seguimiento hasta última visita
Mediana libre de progresión	510 días	270 días	Diferencia descriptiva de 240 días
Log-rank	$\chi^2 = 5,83$	p = 0,016	Evidencia de curvas diferentes bajo supuestos
Cox ajustado	HR = 0,48	IC 95 % 0,26 a 0,89	Asociación ajustada; no prueba causalidad

Figura 5.4.

Supervivencia de pacientes oncológicos según estrategia terapéutica



Pacientes en riesgo	0	6	12	18	24
Vinblastina–prednisolona (n = 26)	26	23	18	14	10
Observación (n = 22)	22	18	12	7	4

Las curvas describen el tiempo hasta el evento y consideran observaciones censuradas

La separación visual no sustituye la estimación del efecto, su IC ni la verificación de supuestos

Plan de análisis

1. Describir la cohorte por tratamiento con estimaciones e indicadores de datos faltantes, sin usar pruebas basales como filtro de covariables.
2. Verificar coherencia de fechas, origen, evento, censura y seguimiento. Describir pérdidas y sus motivos por grupo.
3. Estimar curvas Kaplan-Meier con intervalos de confianza y tabla de números en riesgo.
4. Comparar curvas mediante log-rank como contraste global previamente especificado.
5. Ajustar un modelo de Cox parsimonioso con variables pronósticas pretratamiento definidas por conocimiento clínico.
6. Evaluar riesgos proporcionales con residuos de Schoenfeld y gráficos; si falla, modelar efecto dependiente del tiempo o estimar tiempo medio restringido.
7. Realizar sensibilidad por clínica, definición del evento, pérdidas y posibles confusores no medidos. No sobrecargar el modelo con pocos eventos.

Código R reproducible

El siguiente código supone un archivo denominado mastocitoma_simulado.csv con una fila por perro. Su propósito es mostrar la secuencia analítica; requiere revisión de nombres, codificación y supuestos antes de utilizarse en otro conjunto de datos.

```
library(survival)
library(survminer)

datos <- read.csv("mastocitoma_simulado.csv")
datos$tratamiento <- factor(datos$tratamiento,
                            levels = c("observacion", "vinblastina_pred"))

# Objeto de tiempo a evento: evento = 1; censura = 0
S <- with(datos, Surv(tiempo_dias, evento))

# Curvas Kaplan-Meier y prueba log-rank
km <- survfit(S ~ tratamiento, data = datos)
ggsurvplot(km, data = datos, conf.int = TRUE, risk.table = TRUE,
           xlab = "Días desde la cirugía",
           ylab = "Supervivencia libre de progresión")
survdiff(S ~ tratamiento, data = datos)

# Modelo de Cox ajustado
cox <- coxph(S ~ tratamiento + estadio + indice_mitotico + margenes,
             data = datos, x = TRUE)
summary(cox)

# Evaluación del supuesto de riesgos proporcionales
ph <- cox.zph(cox)
print(ph)
plot(ph)
```

Lectura crítica del código

La función Surv no decide si la censura es informativa ni si el origen temporal es válido. Esas condiciones provienen del diseño. survdiff calcula la comparación log-rank. coxph estima coeficientes en escala log hazard; summary presenta HR al exponentiarlos. cox.zph examina una forma de incumplimiento de riesgos proporcionales, pero un valor p no sustituye la inspección de patrones ni el razonamiento clínico.

El modelo propuesto contiene cuatro términos con 24 eventos, por lo que existe riesgo de estimaciones inestables. No se debe aplicar una regla rígida de eventos por variable como garantía. Conviene limitar grados de libertad, evitar selección automática y considerar penalización o un modelo más parsimonioso. La clínica podría incorporarse mediante errores robustos o efectos compartidos si el número de centros y la información lo permiten; con solo tres clínicas, la inferencia entre centros es débil.

Si el objetivo fuera predicción individual, este análisis sería insuficiente. Se necesitarían un modelo preespecificado, evaluación de calibración, discriminación, error global, utilidad clínica, validación interna que reproduzca todo el proceso de modelado y validación externa. El HR de tratamiento no es una métrica de rendimiento predictivo. Tampoco debe usarse el mismo conjunto para seleccionar variables, ajustar y anunciar desempeño sin corrección del optimismo.

5.17 Síntesis autoevaluación y actividad práctica

Síntesis

- La unidad experimental y la dependencia determinan el análisis antes que el número de filas.
- La descripción debe mostrar distribución, datos faltantes, agrupación y trayectorias, no solo promedios.
- Las estimaciones y los intervalos de confianza responden por magnitud y precisión; el valor p no mide importancia clínica.
- Las pruebas t y ANOVA comparan medias bajo una estructura concreta; ANCOVA ajusta covariables pretratamiento justificadas.
- Los métodos no paramétricos tienen supuestos y estimandos propios; no son una solución automática a cualquier asimetría.
- Los GLM adaptan la distribución al desenlace; los modelos mixtos reconocen jerarquías y medidas repetidas.
- Kaplan-Meier y Cox incorporan censura, pero requieren origen, evento y supuestos claramente definidos.
- PCA describe estructura; PERMANOVA prueba diferencias multivariadas y debe acompañarse de análisis de dispersión.
- Los GAM modelan curvatura, aunque no eliminan confusión, autocorrelación ni necesidad de validación.
- Un modelo predictivo útil puede apoyarse en marcadores no causales. La inferencia causal requiere intervención definida, diseño y supuestos identificables.

- La IA no corrige un diseño defectuoso ni convierte asociaciones en efectos. Su evaluación debe incluir fuga, calibración, transportabilidad y utilidad clínica.

Autoevaluación

1. Un estudio registra 1 200 observaciones provenientes de 40 vacas durante 30 días. ¿Cuál es la unidad independiente y por qué una regresión ordinaria sería problemática?
2. Explique la diferencia entre desviación estándar y error estándar mediante un ejemplo de peso canino.
3. Un tratamiento produce p = 0,03 y una diferencia media de 0,2 puntos de dolor con IC 95 % de 0,02 a 0,38. ¿Qué información falta para juzgar relevancia clínica?
4. ¿Por qué p = 0,06 no demuestra ausencia de efecto? Formule una interpretación basada en un intervalo de confianza.
5. ¿Cuándo utilizaría t pareada y cuándo un modelo mixto en mediciones antes y después?
6. En un ANOVA con cuatro dietas, ¿por qué no es apropiado ejecutar seis pruebas t sin ajuste?
7. Describa una situación en la que ajustar una covariable posterior al tratamiento introduzca sesgo.
8. ¿Qué prueba no paramétrica usaría para tres grupos independientes con desenlace ordinal y qué limita la interpretación?
9. Interprete un OR de 2,0 sin afirmar que el riesgo se duplicó.
10. ¿Qué evidencia sugeriría sobredispersión en un modelo Poisson de conteos parasitarios?
11. ¿Por qué medir 100 aves en un solo galpón por tratamiento constituye pseudorreplicación?
12. Defina censura e indique cuándo podría ser informativa en un estudio oncológico.
13. Interprete un HR de 0,48 con IC 95 % de 0,26 a 0,89.
14. ¿Qué debe examinarse antes de atribuir una PERMANOVA significativa a diferencias entre centroides?
15. ¿Qué aporta un GAM frente a un polinomio de alto grado en una relación temperatura-producción?
16. Proponga un predictor de hospitalización que pueda ser útil sin ser causal.
17. ¿Por qué una explicación SHAP no demuestra el efecto de intervenir sobre una variable?
18. Distinga validación cruzada, validación temporal y validación externa.

19. ¿Qué significa positividad en una comparación causal de tratamientos?
20. Enumere cinco elementos estadísticos que deberían aparecer en el informe de un estudio veterinario reproducible.

Respuestas orientativas

1. La vaca es la unidad biológica repetida; las mediciones dentro de una vaca están correlacionadas. Se requiere un modelo longitudinal o mixto que represente esa dependencia.
2. La desviación estándar describe diferencias de peso entre perros; el error estándar describe la precisión con que la muestra estima la media poblacional.
3. Falta el umbral de cambio clínicamente importante, además del contexto de seguridad, escala, seguimiento y comparador.
4. Indica que, bajo el modelo nulo, el resultado no cruzó el umbral elegido. El IC muestra qué efectos siguen siendo compatibles y si descarta o no magnitudes relevantes.
5. La t pareada sirve para dos mediciones completas por sujeto y una comparación simple. Un mixto maneja varios tiempos, datos desequilibrados, covariables y estructuras de correlación.
6. Porque el error familiar aumenta. Deben definirse contrastes o aplicar un procedimiento como Tukey o Holm según el objetivo.
7. Ajustar por inflamación medida después de una terapia al estimar su efecto total sobre recuperación puede bloquear mediación e introducir sesgo.
8. Kruskal-Wallis. Compara distribuciones mediante rangos y no prueba automáticamente igualdad de medianas.
9. Las odds del evento son dos veces mayores; para comunicar riesgo se necesitan el riesgo basal o probabilidades predichas.
10. Varianza condicional muy superior a la media, residuos incompatibles y mayor frecuencia de conteos extremos que la esperada.
11. El tratamiento se asignó al galpón; las aves comparten exposición y no son réplicas independientes del tratamiento.

12. Es seguimiento incompleto sin evento observado. Puede ser informativa si animales que empeoran abandonan con más frecuencia.
13. La tasa instantánea ajustada se estima 52 % menor; los datos son compatibles con reducciones entre 11 % y 74 %. No equivale a riesgo relativo ni prueba causalidad.
14. La homogeneidad de dispersiones y el esquema de permutación compatible con el diseño.
15. Permite una función suave penalizada y evita imponer oscilaciones globales; aun así requiere seleccionar base, penalización y validación.
16. La clínica de origen puede anticipar ingreso por diferencias operativas, sin ser una causa biológica directa de enfermedad.
17. SHAP atribuye una predicción dentro de un modelo y distribución de datos; no identifica el resultado de una intervención.
18. La cruzada reutiliza particiones dentro del desarrollo; la temporal prueba un periodo posterior; la externa usa otra población o centro.
19. Que exista probabilidad positiva de recibir cada estrategia comparada para todos los perfiles de covariables relevantes.
20. Diseño y unidad experimental, estimando, modelo y supuestos, manejo de datos faltantes, estimación con IC, software y versión, y análisis de sensibilidad.

Actividad práctica integradora

Título. Del dato veterinario a una conclusión defendible.

Contexto. Un conjunto didáctico contiene información de 120 perros sometidos a cirugía ortopédica en cuatro clínicas. Se registran tratamiento analgésico, edad, peso, puntuación basal de dolor, puntuaciones a 6, 12, 24 y 48 horas, necesidad de rescate, tiempo hasta el alta y eventos adversos. El docente puede simular el archivo o adaptar una base anonimizada con autorización ética.

Producto. Cada equipo entregará un informe de 1 500 a 2 000 palabras y un script reproducible en R. El informe debe distinguir un objetivo inferencial y uno predictivo, justificar la unidad analítica, mostrar exploración, definir un estimando clínico, seleccionar el modelo, comprobar supuestos y redactar una conclusión limitada por el diseño.

1. Construya un diccionario de datos con unidad, escala, momento de medición, rango válido y mecanismo probable de ausencia.
2. Formule una pregunta de efecto del tratamiento sobre dolor a 24 horas y una pregunta predictiva de rescate analgésico.
3. Dibuje un diagrama causal mínimo para el efecto del tratamiento. Clasifique las variables como confusoras, mediadoras, pronósticas o posteriores al tratamiento.
4. Explore distribuciones y trayectorias por perro y clínica. Identifique valores imposibles, duplicados y patrones de pérdida.
5. Ajuste un modelo mixto para dolor repetido con tratamiento, tiempo, interacción y puntuación basal. Justifique efectos aleatorios.
6. Estime el efecto a 24 y 48 horas con IC 95 % y compárelo con un umbral clínico definido antes del análisis.
7. Ajuste una regresión logística predictiva para rescate. Separe los perros, no las filas, durante validación. Evalúe discriminación y calibración.
8. Realice un análisis de sensibilidad respecto de datos faltantes y diferencias entre clínicas.
9. Redacte dos conclusiones separadas: una sobre asociación/efecto bajo supuestos y otra sobre capacidad predictiva.
10. Audite el informe con ARRIVE 2.0 y, para la parte predictiva, con TRIPOD+AI.

Rúbrica breve

Tabla 5.8

Criterios de evaluación de la actividad

Criterio	Ponderación	Desempeño esperado
Pregunta y estimando	15 %	Población, exposición, comparador, desenlace y tiempo definidos
Estructura de datos	15 %	Unidad, repetición, clínica y ausencias correctamente tratadas
Modelo y supuestos	25 %	Elección justificada, diagnóstico y sensibilidad
Magnitud e incertidumbre	20 %	Efectos, IC y relevancia clínica interpretados
Predicción	15 %	Separación correcta, calibración y discriminación
Reproducibilidad	10 %	Código ejecutable, versiones y decisiones documentadas

Lista de verificación para el investigador

- ¿La pregunta distingue descripción, inferencia, predicción o causalidad?
- ¿El análisis respeta la unidad de asignación, el agrupamiento y las medidas repetidas?
- ¿El desenlace, comparador y horizonte temporal fueron definidos antes del análisis?
- ¿Se informan magnitud, intervalo de confianza y criterio de relevancia clínica?
- ¿Los supuestos se justifican con diseño, gráficos, diagnósticos y análisis de sensibilidad?
- ¿Las decisiones posteriores a observar los datos se identifican como exploratorias?
- ¿La validación evita fuga entre animales, centros o periodos?
- ¿La conclusión causal corresponde realmente al diseño y a los supuestos?
- ¿El código, las versiones de software y el flujo de datos permiten reproducir el resultado?
- ¿El uso de IA está documentado y sus salidas fueron verificadas por responsables humanos?

CAPÍTULO 6

GESTIÓN CALIDAD Y PREPARACIÓN DE DATOS VETERINARIOS

La solidez de un análisis veterinario depende menos de la sofisticación del algoritmo que de la correspondencia entre los datos, el proceso biológico y la pregunta científica. Una base extensa puede ser inadecuada si mezcla pacientes con visitas, confunde ausencia con normalidad, incorpora mediciones posteriores al desenlace o conserva duplicados producidos por la integración de sistemas. Por el contrario, un conjunto moderado, bien documentado y con una procedencia verificable permite estimaciones interpretables, modelos reproducibles y decisiones clínicas prudentes.

Este capítulo examina el ciclo que comienza en la fuente y termina en una matriz analítica lista para estadística o aprendizaje automático. Se abordan datos clínicos, laboratoriales, productivos, de sensores e imágenes; diccionarios, reglas de calidad, valores atípicos, ausencia, imputación, transformación, codificación, ingeniería de características, desbalance, interoperabilidad, metadatos y versiones. El criterio rector es que cada modificación debe responder a una razón clínica o metodológica, quedar registrada y poder repetirse sin alterar los datos originales.

Objetivos de aprendizaje

Al finalizar el capítulo, el lector podrá:

1. reconocer las principales fuentes de datos veterinarios y sus sesgos de generación;
2. distinguir paciente, episodio, visita, muestra, imagen, animal, lote y explotación como posibles unidades de análisis;
3. construir un diccionario de datos con definiciones operacionales, unidades, dominios y reglas de validación;
4. evaluar completitud, validez, consistencia, unicidad, oportunidad y representatividad;
5. diagnosticar duplicados, errores, valores atípicos y mecanismos de ausencia;
6. seleccionar una estrategia de imputación compatible con el diseño y el objetivo del análisis;
7. transformar variables sin introducir fuga de información;
8. preparar variables categóricas, series temporales, imágenes y datos agrupados;
9. afrontar el desbalance de clases sin reducir la evaluación a la exactitud global; y

10. documentar procedencia, metadatos y versiones para sostener la reproducibilidad.

6.1. Fuentes de datos en Ciencias Veterinarias

La investigación veterinaria integra fuentes creadas con propósitos distintos. La historia clínica se produce para atender a un paciente; la base productiva, para gestionar una explotación; el registro de laboratorio, para identificar y liberar resultados; el sensor, para vigilar continuamente una señal; y el cuaderno experimental, para responder una hipótesis predefinida. Esa finalidad primaria condiciona qué se mide, cuándo se mide, qué queda sin registrar y qué población está representada.

En animales de compañía son frecuentes las historias clínicas electrónicas, facturación, farmacia, imagenología, microbiología y patología. En producción animal se añaden genealogía, alimentación, ordeño, reproducción, tratamientos colectivos, mortalidad, decomisos, ambiente, movimientos y rendimiento. La vigilancia epidemiológica incorpora notificaciones, investigaciones de foco, resultados diagnósticos, vacunación, geolocalización y contactos. La fauna silvestre agrega captura, telemetría, observación, necropsias y variables del hábitat.

Una clasificación útil distingue datos **primarios**, obtenidos para el estudio, y **secundarios**, reutilizados desde atención, producción o vigilancia. Los primeros ofrecen mayor control sobre definiciones y tiempos; los segundos permiten cohortes grandes y evidencia del mundo real, pero arrastran decisiones clínicas, comerciales y operativas. Los registros electrónicos veterinarios pueden sostener farmacovigilancia y estudios de efectividad, aunque requieren validar fenotipos, temporalidad y procesos de extracción antes de interpretarlos como evidencia (Davies et al., 2025).

También importa la estructura. Los datos tabulares incluyen una fila por animal, visita, muestra o día; el texto clínico es no estructurado; imágenes y videos poseen matrices de píxeles y metadatos; acelerómetros generan series temporales; la genómica produce secuencias y conteos de alta dimensión; y las redes de contacto se expresan como nodos y vínculos. Convertirlos prematuramente a una tabla única puede destruir relaciones necesarias. Una radiografía pertenece a un estudio, una visita y un animal; varias visitas pertenecen al mismo perro; varios bovinos comparten hato, manejo y ambiente. Esa jerarquía debe conservarse.

Ejemplo veterinario 1. Para estudiar mastitis bovina pueden combinarse recuento de células somáticas, conductividad de leche, producción diaria, paridad, días en leche, tratamientos y temperatura ambiental. El animal-día es una unidad observacional, pero las observaciones se agrupan en vaca y hato. Una división aleatoria de filas permitiría que días contiguos de la misma vaca aparezcan en entrenamiento y prueba, generando una evaluación demasiado optimista.

Ejemplo veterinario 2. En un estudio de dolor posoperatorio canino, la escala Glasgow CMPS-SF, la administración analgésica y los signos vitales se registran repetidamente. Si el desenlace es rescate analgésico durante las primeras 24 horas, una medición tomada después del rescate no puede utilizarse como predictor basal. El problema no se corrige eliminando una variable: debe definirse el tiempo cero y congelarse la información disponible hasta ese instante.

Tabla 6.1.

Fuentes de datos y riesgos metodológicos frecuentes

Fuente	Unidad habitual	Fortalezas	Riesgos dominantes
Historia clínica	paciente, visita, episodio	contexto clínico y seguimiento	texto libre, codificación variable, sesgo de atención
Laboratorio	muestra, analito, panel	medición cuantitativa y control interno	unidades distintas, límites de detección, cambios de equipo
Imagenología	estudio, serie, imagen	señal anatómica rica	protocolos heterogéneos, etiquetas débiles, múltiples imágenes por animal
Sensor	animal-tiempo	continuidad y detección temprana	deriva, pérdida de señal, autocorrelación, batería
Producción	animal, lote, hato, ciclo	volumen, rendimiento y gestión	cambios de software, agregación, incentivos económicos
Vigilancia	evento, foco, territorio	alcance poblacional	subnotificación, demora y cambios de definición
Ensayo experimental	unidad experimental	control, aleatorización y cronología	tamaño pequeño, pérdidas, incumplimiento de protocolo
Ómicas	muestra biológica	alta resolución molecular	dimensionalidad, lote técnico, múltiples comparaciones

6.2. Historias clínicas electrónicas

Una historia clínica electrónica veterinaria suele contener identificación del animal y tutor, especie, raza, sexo, condición reproductiva, edad o fecha de nacimiento, peso, motivo de consulta, anamnesis, examen físico, diagnósticos, procedimientos, prescripciones, resultados, documentos e información de facturación. No todos los campos poseen el mismo estatus epistemológico.

“Vómito” puede ser motivo de consulta referido por el tutor, signo observado, problema activo, diagnóstico presuntivo o código usado para justificar una prueba. Antes de contar casos debe definirse qué evidencia constituye un caso.

Los registros se organizan alrededor de la atención, no necesariamente de episodios de enfermedad. Un episodio de dermatitis puede abarcar varias visitas, llamadas y recetas; una visita puede contener problemas independientes. El identificador de consulta no reemplaza al identificador del animal ni al del episodio. En investigación longitudinal deben conservarse las tres capas para evitar contar recaídas como casos nuevos o perder recurrencias reales.

La mezcla de campos estructurados y narrativas exige procedimientos diferentes. Los campos estructurados facilitan filtros, pero pueden usarse de forma incompleta. El texto contiene matices, negaciones y temporalidad: “sin diarrea”, “diarrea hace seis meses” y “vigilar aparición de diarrea” no equivalen a un evento presente. La extracción automática mediante procesamiento de lenguaje natural puede ayudar, pero necesita un estándar de referencia anotado por clínicos, reglas de adjudicación y evaluación por clase. En historias veterinarias de enteropatía felina, Wulcan et al. (2025) mostraron que la ambigüedad del texto también genera desacuerdo humano; por ello, el error del modelo no debe analizarse separado de la confiabilidad de la etiqueta.

Los registros de práctica primaria representan animales que acceden a clínicas participantes y cuyos tutores aceptan o pueden pagar atención. No son un censo de la población animal. Razas con seguimiento intensivo pueden mostrar más diagnósticos por mayor oportunidad de detección. Clínicas de referencia concentran cuadros graves. La prevalencia calculada en una red clínica debe describirse como prevalencia registrada o proporción de animales atendidos, según el diseño, y no extrapolarse sin discutir cobertura y selección.

La extracción debe conservar el estado original de los campos y crear variables derivadas en una capa aparte. Sustituir “raza: mestizo pequeño” por una raza específica inventa precisión. Convertir una edad aproximada en fecha exacta también la inventa. El valor analítico puede incorporar intervalos o categorías de certeza.

6.3. Datos de laboratorio sensores imágenes y producción

Datos de laboratorio

Un resultado de laboratorio no se reduce a un número. Requiere animal, muestra, fecha y hora de toma, matriz biológica, analito, método, instrumento, unidad, intervalo de referencia, estado preanalítico y, cuando corresponda, límite de detección. “Creatinina = 2,1” es insuficiente sin unidad y especie; el significado cambia entre mg/dL y µmol/L, entre perro y gato, y según edad, masa muscular e hidratación.

La armonización debe distinguir conversión matemática de equivalencia clínica. Dos métodos pueden reportar la misma unidad y mantener un sesgo sistemático. Cuando cambia el analizador, conviene registrar una variable de lote o periodo, revisar controles y evaluar comparabilidad. Los símbolos “<0,1” y “>500” representan censura analítica, no 0,1 y 500 exactos. El reemplazo mecánico por el límite introduce acumulaciones artificiales y distorsiona medias y regresiones.

Sensores y dispositivos

Collares, aretes, podómetros, acelerómetros, cámaras térmicas, micrófonos y estaciones ambientales generan datos de alta frecuencia. La agricultura y ganadería de precisión combinan sensor, algoritmo e interfaz, pero la disponibilidad de una señal no garantiza que mida de forma válida el comportamiento o estado sanitario previsto. Las revisiones de tecnologías para bienestar bovino muestran que la validación comercial y la validación científica no siempre coinciden para todos los contextos y desenlaces (Stygar et al., 2021).

Antes de extraer características deben revisarse frecuencia de muestreo, sincronización, zona horaria, posición del dispositivo, calibración, firmware, pérdida de paquetes, deriva y periodos sin uso. Un collar retirado durante tratamiento puede producir “inactividad” extrema; imputarla como reposo convertiría una falla operativa en fenotipo. Se recomienda conservar una bandera de calidad por ventana temporal y diferenciar cero observado de ausencia de señal.

Imágenes y video

En radiografía, ecografía, tomografía y resonancia deben preservarse archivos originales y metadatos. DICOM organiza imágenes e información asociada para almacenamiento e intercambio; exportar únicamente capturas JPEG puede perder profundidad, escala, orientación y datos del estudio. Antes de entrenar un modelo se revisan identificadores, lateridad, proyección, institución, dispositivo, parámetros de adquisición y posibles marcas visibles.

La unidad de división suele ser el animal, no la imagen. Varias proyecciones del mismo paciente son correlacionadas; colocarlas en entrenamiento y prueba permite reconocer características del estudio o del equipo. También debe evitarse que una anotación gráfica, el nombre del hospital o una etiqueta de urgencia revele indirectamente el diagnóstico. La anonimización exige retirar identificadores tanto del encabezado como de píxeles quemados.

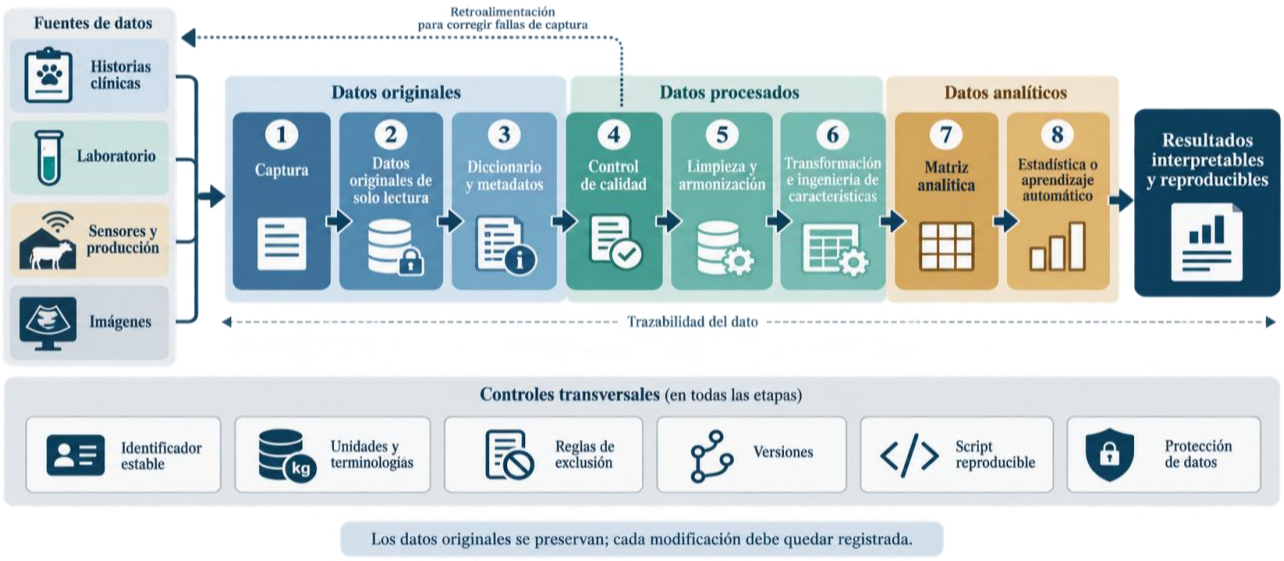
Datos productivos

Producción de leche, ganancia de peso, conversión alimenticia, fertilidad, postura, mortalidad y decomisos tienen denominadores y ventanas específicas. Un promedio mensual de leche puede ocultar días sin medición o cambios en el número de vacas ordeñadas. La mortalidad de lote requiere población en riesgo y tiempo de exposición. En granjas, “tratamiento” puede registrarse al comprar, dispensar o administrar; la investigación debe identificar cuál evento representa el dato.

Las integraciones productivas son sensibles a cambios de identificador. Arete electrónico, número interno y código del sistema de ordeño deben vincularse mediante una tabla maestra con fechas de vigencia. Reutilizar un número de corral para otro animal sin temporalidad genera trayectorias imposibles.

Figura 6.1.

Ciclo de transformación del dato veterinario



6.4. Creación de diccionarios de datos

El diccionario de datos es el contrato semántico del proyecto. Describe qué significa cada variable, cómo se obtuvo, qué valores admite y qué restricciones posee. No debe elaborarse al final para acompañar la publicación; se construye antes de recolectar o transformar y se actualiza con control de versiones.

Cada variable necesita un nombre estable y legible por máquina; etiqueta en lenguaje natural; definición operacional; tipo; unidad; nivel de medición; valores permitidos; códigos de ausencia; fuente; momento; responsable; reglas de validación; sensibilidad; y transformaciones previstas. En datos derivados se añade fórmula, ventana temporal, versión del código y variables de origen.

Tabla 6.2.

Fragmento de diccionario para una cohorte canina

Campo	Definición operacional	Tipo y dominio	Unidad	Regla de calidad
animal_id	identificador seudonimizado estable del perro	texto, no nulo	no aplica	único por animal, no reutilizable
visit_id	identificador de la consulta	texto, no nulo	no aplica	único por visita
visit_datetime	inicio documentado de la consulta	fecha-hora ISO 8601	UTC o zona declarada	no posterior a extracción

weight_kg	peso medido al ingreso	decimal > 0	kg	revisar <0,5 o >100, no borrar automáticamente
sex_neuter	sexo y estado reproductivo	F, FE, M, ME, desconocido	no aplica	catálogo cerrado
temperature_c	temperatura corporal de ingreso	decimal	°C	verificar método y valores <34 o >43
hospitalized	ingreso hospitalario tras evaluación inicial	binaria 0/1	no aplica	definida antes del modelado
creatinine_mg_dl	primera creatinina dentro de 6 h	decimal/censurada	mg/dL	convertir unidades y guardar indicador de censura

Un diccionario evita conflictos como usar “0” simultáneamente para “no”, “no medido” y “desconocido”. También obliga a precisar denominadores. “Edad” puede ser edad en la visita, edad al diagnóstico o edad al cierre; “duración” puede expresarse en horas completas o días calendario. La definición debe ser reproducible por una persona que no participó en la recolección.

El catálogo de valores se mantiene separado cuando es extenso. Para raza puede conservarse el texto original, una forma armonizada, el grupo racial y una bandera de incertidumbre. La columna armonizada no reemplaza la fuente. Si se actualiza la taxonomía, los datos originales permiten reconstruir la codificación.

6.5. Estandarización y codificación de variables

Estandarizar significa representar el mismo concepto del mismo modo; no equivale a convertir todas las variables a puntuaciones z. La normalización semántica precede al escalamiento matemático.

Los nombres deben seguir una convención: minúsculas, separador único y sin símbolos ambiguos. Las fechas se almacenan en un formato inequívoco, preferentemente ISO 8601, con zona horaria cuando importa. Las unidades se guardan en campos explícitos o se convierten a una unidad canónica preservando el valor original, la unidad original y la regla de conversión. Las categorías utilizan códigos estables, no colores, posiciones en una hoja o abreviaturas locales indocumentadas.

En veterinaria, especie, raza, sexo y estado reproductivo requieren especial cuidado. Un campo “sexo” con M, H, MC, HE, castrado y 0 mezcla dos atributos. Puede descomponerse en sexo

biológico registrado y condición reproductiva, conservando “desconocido” cuando no hay evidencia. Las razas escritas libremente necesitan reglas de mayúsculas, tildes, sinónimos y cruces; agruparlas por tamaño solo es válido si el objetivo justifica esa pérdida de detalle.

La codificación clínica controlada mejora comparabilidad. VeNom ofrece términos estandarizados para diagnósticos, motivos y procedimientos en animales de compañía (VeNom Coding Group, n.d.); la extensión veterinaria de SNOMED CT integra contenido animal con una terminología clínica formal y facilita el intercambio entre sistemas (Veterinary Terminology Services Laboratory, 2026). La adopción local debe mantener código, versión, etiqueta original y mapeo, porque las terminologías cambian.

Regla práctica. Nunca recodificar sobre la única columna existente. Crear `diagnostico_original`, diagnostico_normalizado, codigo_sistema, version_vocabulario y certeza_diagnostica. Esta estructura permite auditoría y análisis de sensibilidad.

6.6. Evaluación de calidad de los datos

La calidad no es una propiedad absoluta; es adecuación para un uso. Una base puede servir para facturación y ser insuficiente para estimar incidencia. La evaluación debe vincular cada dimensión con la pregunta científica.

Las dimensiones más útiles son:

- **Completitud:** proporción de valores esperados disponibles, evaluada por variable, clínica, periodo y subgrupo.
- **Validez:** cumplimiento del tipo, dominio, rango y formato definidos.
- **Consistencia:** ausencia de contradicciones internas, temporales o entre fuentes.
- **Unicidad:** cada entidad o evento aparece el número correcto de veces.
- **Exactitud:** proximidad al valor verdadero, estimada mediante fuente de referencia o revisión.
- **Oportunidad:** demora entre evento, registro, disponibilidad y extracción.

- **Concordancia:** acuerdo entre observadores, sistemas o métodos.
- **Representatividad:** correspondencia entre la población registrada y la población objetivo.
- **Trazabilidad:** capacidad de reconstruir origen y transformaciones.

Penev (2024) identificó una gran variación en marcos y métricas usados para evaluar historias electrónicas. De allí se desprende que una puntuación global de calidad suele ser poco informativa. Es preferible un perfil por dimensión acompañado de umbrales, justificación y consecuencias analíticas.

Tabla 6.3.

Reglas de calidad aplicadas a datos clínicos

Dimensión	Indicador	Ejemplo de regla	Acción ante incumplimiento
Completitud	% con peso basal	esperado en visitas de anestesia	describir por centro y periodo; investigar flujo
Validez	% dentro del dominio	frecuencia cardiaca numérica y positiva	revisar fuente; marcar no resoluble
Consistencia	contradicciones por 1000	fecha de alta ≥ fecha de ingreso	corregir con evidencia o excluir intervalo
Unicidad	duplicados exactos	una fila por visit_id en tabla de visitas	consolidar mediante regla documentada
Oportunidad	mediana de demora	toma de muestra a validación	incorporar retraso si afecta predicción
Representatividad	distribución comparada	especie, edad, región y tipo de clínica	limitar inferencia o ponderar con cautela

El informe de calidad se realiza antes y después de la depuración. Debe incluir recuento de filas, entidades únicas, periodos, distribución de desenlaces, porcentajes de ausencia, rangos, duplicados y exclusiones. Los umbrales automáticos sirven para señalar registros, no para borrarlos sin revisión. Un potasio de 9 mmol/L puede ser error preanalítico, pero también un hallazgo crítico; el contexto decide.

6.7. Detección de duplicados errores y valores atípicos

Duplicados

Un duplicado exacto repite todas las columnas; uno funcional representa el mismo evento con diferencias menores. Los primeros pueden surgir por concatenar exportaciones; los segundos por correcciones, reenvíos de laboratorio, cambios de nombre o múltiples sistemas. La clave de deduplicación debe corresponder a la entidad: animal, visita, muestra, receta o imagen.

El emparejamiento determinista utiliza identificadores confiables. Cuando faltan, el enlace probabilístico combina nombre seudonimizado, especie, fecha de nacimiento, tutor, microchip y fechas, con umbrales y revisión manual. No se debe fusionar solo por nombre: “Luna” puede pertenecer a varios animales. Toda fusión conserva los identificadores de origen y una puntuación o razón de enlace.

Errores

Los errores pueden ser sintácticos, semánticos o temporales. “31/02/2025” es inválido; “gato Holstein” es inconsistente; una necropsia anterior al nacimiento es temporalmente imposible. Algunas reglas combinan variables: dosis total debe ser compatible con dosis por kilogramo y peso, aunque tolerancias y redondeos deben considerarse.

La corrección requiere evidencia. Si la historia muestra 220 kg para un Chihuahua y la nota dice 2,2 kg, puede corregirse conservando valor original, nuevo valor, fuente y responsable. Si no existe evidencia, se marca como no resoluble; no se reemplaza por un promedio para ocultar el problema.

Valores atípicos

Un valor atípico es inusual respecto de una distribución, no necesariamente erróneo. Puede representar enfermedad extrema, especie distinta, unidad equivocada, error de digitación o una subpoblación legítima. Se detecta mediante reglas clínicas, gráficos, residuos, distancia robusta y comparación longitudinal. El rango intercuartílico identifica extremos univariados, pero no incorpora especie, edad o estado clínico.

La decisión se separa en cuatro posibilidades: confirmar y conservar; corregir con fuente; representar como censurado; o excluir con criterio predefinido. La winsorización automática puede borrar casos graves precisamente relevantes para medicina de urgencias. El análisis de sensibilidad

—con y sin observaciones discutibles— muestra cuánto dependen las conclusiones de esas decisiones.

6.8. Datos faltantes y mecanismos de ausencia

Un valor faltante puede significar que no se midió, no se registró, no aplica, se perdió, quedó por debajo del límite o se ocultó. Un único código `NA` en el archivo analítico puede ser necesario para el software, pero el sistema de origen debería conservar la razón cuando se conoce.

La clasificación clásica distingue ausencia completamente al azar (MCAR), al azar condicionada a variables observadas (MAR) y no al azar (MNAR). MCAR implica que la probabilidad de ausencia no depende de datos observados ni no observados; suele ser poco plausible en clínica. MAR permite que dependa de información registrada: por ejemplo, la presión arterial se mide más en gatos mayores o con creatinina elevada. MNAR aparece cuando la ausencia depende del propio valor no observado, incluso tras condicionar: tutores pueden rechazar una prueba por sospecha de gravedad no capturada o un sensor puede fallar durante actividad intensa.

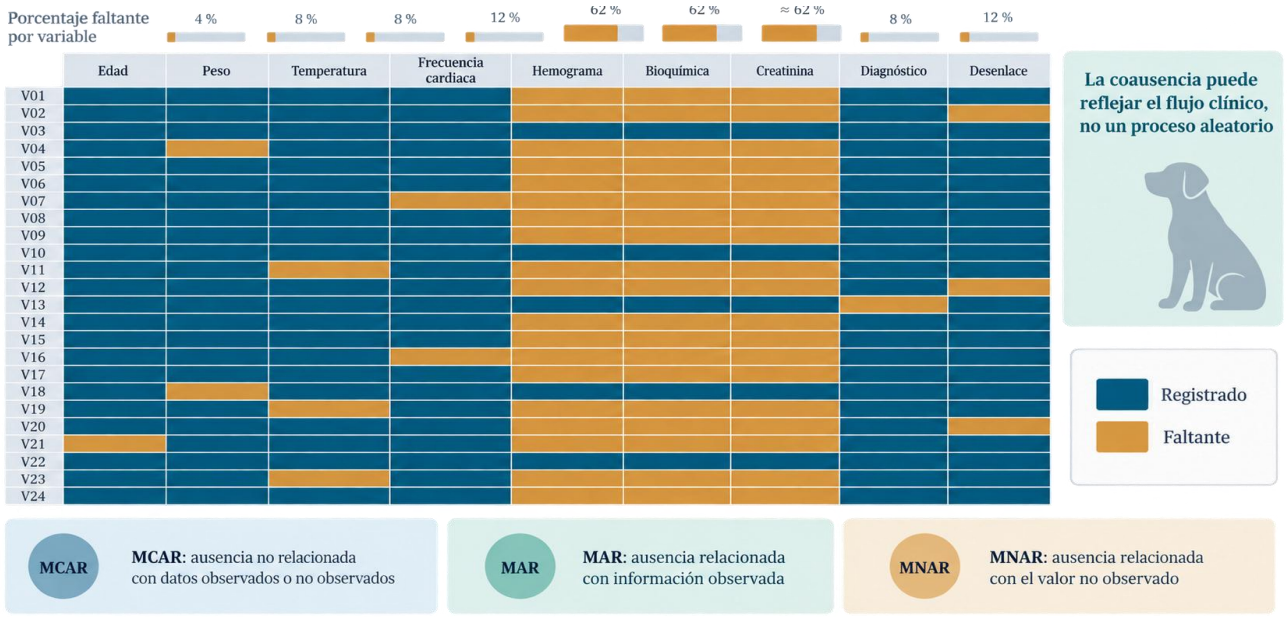
Estos mecanismos son supuestos, no etiquetas demostrables solo con la base. Se investigan comparando patrones, flujos de trabajo, centros, periodos y variables auxiliares; después se realizan análisis de sensibilidad. Un mapa de ausencia puede mostrar que hemograma y bioquímica faltan juntos porque pertenecen a un panel no solicitado. Esa coausencia es informativa sobre el proceso clínico.

Es crucial distinguir variables que no aplican. “Número de partos” en un macho no es un dato perdido; su dominio no aplica. Convertirlo a cero podría ser válido para una variable bien definida como “partos previos”, pero no para “dato reproductivo disponible” sin documentar el razonamiento.

La eliminación completa de filas es válida bajo supuestos restrictivos y puede reducir precisión y representatividad. Antes de decidir se informa el porcentaje por variable y registro, el patrón conjunto y la relación con desenlace, centro, especie, edad y tiempo. En modelos predictivos debe verificarse que el mecanismo de ausencia futuro sea similar al del entrenamiento.

Figura 6.2.

Mapa de ausencia en historias clínicas caninas



6.9. Imputación de datos

Imputar consiste en reemplazar valores ausentes por valores plausibles bajo un modelo explícito. No crea información observada ni elimina incertidumbre. La imputación simple por media o mediana reduce artificialmente la varianza, altera asociaciones y produce concentraciones poco naturales. Puede ser aceptable dentro de un pipeline predictivo como referencia, pero sus limitaciones deben evaluarse.

Para inferencia, la imputación múltiple genera varios conjuntos completos, analiza cada uno y combina estimaciones para incorporar incertidumbre. El modelo de imputación debe incluir el desenlace, predictores, variables auxiliares relacionadas con ausencia y estructura del diseño. En datos longitudinales o agrupados debe respetar animal, clínica o hato; ignorar esa estructura puede producir combinaciones imposibles y errores estándar inadecuados. La obra de van Buuren (2018) constituye una referencia metodológica para imputación flexible mediante ecuaciones encadenadas.

En predicción, la imputación se aprende exclusivamente con entrenamiento y se aplica sin recalcular a validación y prueba. Hacerla antes de dividir datos permite que la distribución de prueba influya en entrenamiento. Lo mismo ocurre con selección de variables, escalamiento y sobremuestreo. Collins et al. (2024) exigen describir el tratamiento de datos faltantes y todas las etapas del desarrollo de modelos predictivos.

Algunos algoritmos aceptan ausencia de manera nativa, pero eso no exime de comprenderla. El patrón de `NA` puede convertirse en atajo institucional: si solo una clínica solicita lactato a pacientes críticos, el modelo aprende la práctica de esa clínica, no la fisiología. Un indicador de ausencia puede mejorar discriminación local y deteriorar transportabilidad.

Tabla 6.4.

Selección orientativa de estrategias de ausencia

Situación	Estrategia posible	Precaución
<2% de ausencia aparentemente aleatoria	análisis completo y sensibilidad	verificar pérdida por subgrupo
Variable continua MAR para inferencia	imputación múltiple	incluir desenlace, auxiliares y estructura
Predictor para modelo desplegable	imputador dentro del pipeline	ajustar solo con entrenamiento
Serie temporal con hueco corto	interpolación contextual	no atravesar cambios clínicos ni largos periodos
Resultado bajo límite de detección	método para censura	no tratar como cero automático
Variable no aplicable	categoría o regla estructural	no confundir con ausencia
MNAR plausible	análisis de sensibilidad	variar supuestos y reportar impacto

Una buena imputación se evalúa comparando distribuciones observadas e imputadas, relaciones multivariadas, plausibilidad clínica y estabilidad de resultados. No se valida por la semejanza de una media aislada. Tampoco se imputan identificadores, el desenlace desconocido que se intenta predecir ni mediciones posteriores para reconstruir el pasado.

6.10. Transformación normalización y escalamiento

Las transformaciones cambian la representación para satisfacer supuestos, estabilizar variabilidad o facilitar optimización. No deben aplicarse por hábito. La transformación logarítmica es útil para

concentraciones positivas sesgadas, pero requiere decidir cómo tratar ceros y explicar la interpretación multiplicativa. La raíz cuadrada puede ayudar con conteos; para proporciones cercanas a 0 o 1 suelen preferirse modelos con distribución adecuada antes que transformaciones arbitrarias.

El escalamiento z resta la media y divide por desviación estándar. Resulta relevante para regresión penalizada, máquinas de vectores de soporte, redes neuronales, k vecinos y métodos basados en distancia. Los árboles de decisión y bosques aleatorios no suelen necesitarlo. El escalamiento robusto con mediana y rango intercuartílico reduce la influencia de extremos, pero no sustituye la revisión de errores.

La normalización min–max lleva una variable a un intervalo definido y es sensible a nuevos valores fuera del rango de entrenamiento. Todos los parámetros —media, desviación, mínimo, máximo o cuantiles— se estiman en entrenamiento. Si hay validación cruzada, se recalculan dentro de cada pliegue. Este detalle evita fuga y reproduce las condiciones reales de uso.

Las variables temporales merecen tratamiento propio. La edad se calcula respecto a la fecha índice; la duración desde inicio de síntomas puede truncarse o ser aproximada; hora y estación pueden representarse cíclicamente con seno y coseno cuando su proximidad circular es relevante. No debe inferirse una periodicidad biológica solo porque una transformación la permita.

6.11. Codificación de variables categóricas

Las categorías nominales no poseen orden. Asignar perro = 1, gato = 2 y caballo = 3 introduce distancias inexistentes. La codificación one-hot crea indicadores binarios y es transparente, aunque aumenta dimensionalidad. En categorías ordinales —por ejemplo, condición corporal con escala validada— pueden usarse códigos ordenados si el modelo y la interpretación respetan que la distancia entre niveles quizá no sea constante.

Las categorías raras pueden agruparse bajo una regla preespecificada y clínicamente coherente. Unificar todas las razas infrecuentes como “otras” puede mezclar tamaños y predisposiciones muy distintas. Alternativas son grupo racial, peso adulto, función zootécnica o embeddings, siempre justificando la representación. La categoría “desconocido” no equivale a “otros”.

La codificación por frecuencia o por media del desenlace puede ser útil en alta cardinalidad, pero la segunda presenta alto riesgo de fuga. Debe aprenderse dentro de cada pliegue y regularizarse. Nunca se calcula con todo el conjunto ni directamente sobre prueba.

Para texto clínico, una bolsa de palabras ofrece una línea base interpretable; representaciones densas pueden capturar contexto, pero requieren recursos, validación y gobernanza de privacidad. Las negaciones, temporalidad, especie y abreviaturas locales exigen un corpus representativo. El código automatizado no reemplaza la adjudicación clínica de una muestra.

6.12. Ingeniería y selección de características

Una característica es una representación analítica derivada de datos disponibles en el momento de decisión. La ingeniería combina conocimiento veterinario, temporalidad y propiedades del algoritmo. Su objetivo no es fabricar muchas columnas, sino expresar señales relevantes sin anticipar el desenlace.

Ejemplos clínicos incluyen cambio porcentual de peso, relación neutrófilos/linfocitos, tendencia de creatinina, número de consultas en 90 días, variabilidad de glucosa y dosis acumulada antes del tiempo índice. En producción pueden calcularse pendiente de leche, desviación respecto a la curva individual, duración de rumia nocturna y anomalía térmica respecto al hato. En imágenes se derivan textura o representaciones profundas; en sensores, energía por bandas, duración de episodios y entropía.

La ventana define el significado. “Consultas previas” debe excluir la visita índice; “antibiótico en 30 días” requiere fechas de administración, no solo facturación. Una variable disponible después del diagnóstico constituye fuga aunque parezca clínicamente relacionada. Hospitalización, orden de radiografía, derivación y código de alta pueden revelar una decisión tomada después del punto donde se pretende predecir.

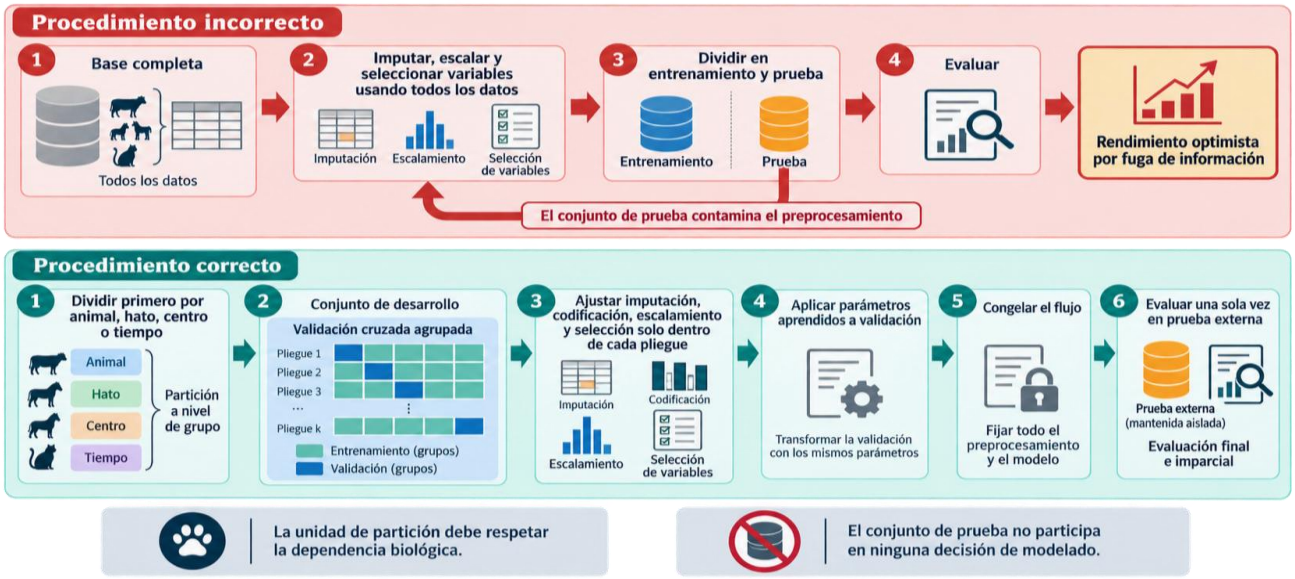
La selección de características puede basarse en conocimiento, estabilidad, redundancia y desempeño dentro de validación. Los filtros univariados ignoran interacciones; los métodos embebidos como LASSO o importancia de árboles dependen del modelo; la eliminación recursiva aumenta riesgo de sobreajuste si se realiza fuera del remuestreo. Para explicación causal,

seleccionar solo por poder predictivo puede ajustar mediadores o colisionadores; se requiere un modelo causal previo.

La alta correlación no implica que una variable sea inútil. Temperatura corporal en °C y °F son redundantes; hematocrito y hemoglobina están relacionados biológicamente, pero discrepancias pueden indicar calidad o enfermedad. La selección debe combinar estadística con plausibilidad y propósito.

Figura 6.3.

Preprocesamiento sin fuga de información



Lista de control contra fuga de información:

1. fijar el tiempo índice;
2. listar qué información existía entonces;
3. agrupar divisiones por animal, hato o clínica;
4. separar temporalmente si se predice el futuro;
5. ajustar imputación, escalamiento, selección y balanceo solo en entrenamiento;
6. revisar proxies administrativos del desenlace; y

7. probar el pipeline completo en datos externos o posteriores.

6.13. Datos desbalanceados

Existe desbalance cuando una clase es mucho menos frecuente que otra, como mortalidad, reacción adversa o enfermedad emergente. No es un defecto que deba eliminarse siempre: puede reflejar la prevalencia real. El problema aparece cuando el algoritmo ignora la clase minoritaria o cuando las métricas ocultan el fallo.

Con 2% de positivos, predecir "negativo" para todos produce 98% de exactitud y ninguna utilidad clínica. Deben informarse sensibilidad, especificidad, valores predictivos, curva precisión-recall, discriminación y calibración. Los valores predictivos dependen de prevalencia; cambiar la muestra mediante sobremuestreo altera la proporción de entrenamiento y puede descalibrar probabilidades.

Las opciones incluyen ponderar la función de pérdida, submuestrear la mayoría, sobremuestrear la minoría o generar observaciones sintéticas. SMOTE interpola entre casos minoritarios (Chawla et al., 2002), pero puede crear combinaciones clínicamente imposibles y no debe cruzar fronteras de animal, hato o tiempo. Cualquier remuestreo se realiza dentro del conjunto de entrenamiento y de cada pliegue, nunca antes de separar prueba.

También puede ajustarse el umbral según costos. Para tamizaje de parvovirosis puede priorizarse sensibilidad y confirmación posterior; para una intervención invasiva el costo de falsos positivos cambia. El umbral no se fija mirando la prueba final. Se selecciona en validación, se evalúa en prueba y se acompaña de análisis de utilidad clínica.

La clase rara puede ser heterogénea. "Muerte" reúne etiologías distintas; "complicación" puede incluir eventos leves y graves. Antes de balancear conviene revisar la etiqueta, adjudicar casos dudosos y asegurar suficientes eventos por subgrupo. Más copias de una etiqueta débil no producen mejor evidencia.

6.14. Interoperabilidad y estándares veterinarios

La interoperabilidad permite que dos sistemas intercambien datos y conserven significado. Incluye capas técnicas formatos, interfaces y transporte, semánticas terminologías y unidades,

organizacionales responsabilidades y flujos y legales permisos y finalidades. Exportar dos archivos CSV no garantiza interoperabilidad si “fecha” significa toma en uno y validación en otro.

HL7 FHIR representa información sanitaria mediante recursos y APIs (Health Level Seven International, 2023). Aunque fue desarrollado principalmente para salud humana, sus patrones pueden adaptarse con perfiles veterinarios cuidadosamente gobernados. DICOM es central en imagenología (National Electrical Manufacturers Association, n.d.); LOINC identifica observaciones y pruebas (Regenstrief Institute, n.d.); SNOMED CT y su extensión veterinaria representan conceptos clínicos. Ningún estándar resuelve por sí solo el mapeo local ni la calidad del registro.

Los identificadores son la columna vertebral. Un sistema interoperable distingue animal, propietario o responsable, establecimiento, profesional, dispositivo, muestra, orden, resultado, producto y ubicación. En producción se añade lote, explotación y movimiento. Los identificadores deben ser persistentes, no contener información sensible evidente y tener reglas para fusiones y bajas.

La estrategia europea de grandes datos regulatorios veterinarios reconoce que las fuentes reguladoras maduras se benefician de estándares acordados, mientras otras fuentes requieren mayor armonización (European Medicines Agency, 2022). En un proyecto latinoamericano, la adopción debe considerar infraestructura, conectividad, vocabularios en español, catálogos nacionales y capacidad de mantenimiento. Traducir etiquetas sin preservar códigos rompe equivalencia.

Tabla 6.5.

Estándares y usos posibles

Estándar o vocabulario	Función principal	Aplicación veterinaria	Limitación práctica
DICOM	imágenes y metadatos	radiología, CT, MRI, ecografía	campos incompletos y anonimización compleja
HL7 FHIR	intercambio por recursos	paciente animal, observación, informe, medicación	requiere perfiles y extensiones veterinarias
LOINC	identificación de pruebas	laboratorio y observaciones	método, especie y muestra deben mapearse bien

SNOMED CT VetSCT	terminología clínica	diagnósticos, hallazgos, procedimientos	licencias, gobernanza y capacitación
VeNom	términos clínicos veterinarios	práctica primaria y referencia	cobertura y adopción varían por región
ISO 8601	fecha y hora	cronología entre sistemas	zona horaria y precisión deben declararse

La interoperabilidad se verifica con casos de uso: ¿puede reconstruirse qué animal, muestra, método y unidad produjeron un resultado? ¿Se distingue una vacuna planificada de una administrada? ¿Se conserva la negación? La prueba de conformidad técnica se complementa con validación clínica de mensajes y mapeos.

6.15. Trazabilidad metadatos y control de versiones

La trazabilidad describe el recorrido desde la fuente hasta el resultado. Una matriz analítica sin procedencia impide saber qué exportación, filtros o correcciones la generaron. Los principios FAIR —localizables, accesibles, interoperables y reutilizables— subrayan la importancia de identificadores y metadatos procesables por máquinas, pero FAIR no significa que datos clínicos deban ser públicos; el acceso puede estar autenticado y restringido (Wilkinson et al., 2016).

Los metadatos del conjunto incluyen título, propósito, responsables, cobertura, población, fechas, fuentes, versión, licencia o condiciones de acceso, diccionario, método de seudonimización, limitaciones y contacto. A nivel variable se documentan definición, unidad y transformación. En imágenes y sensores se agregan dispositivo, configuración, calibración y firmware. En laboratorio, método, lote, límites y control de calidad.

Se recomienda una arquitectura por capas:

- `raw`: copia inmutable de la extracción;
- `staging`: estructura y nombres armonizados;
- `clean`: errores resueltos y reglas aplicadas;
- `analytic`: una versión por pregunta y unidad de análisis;
- `outputs`: tablas, modelos y figuras derivados.

El código, no la edición manual, debe realizar transformaciones repetibles. Cada ejecución guarda fecha, versión, parámetros, recuentos de entrada y salida, exclusiones y verificaciones. Git puede versionar código y documentación; los datos grandes o sensibles requieren repositorios y controles apropiados. No se suben historias identificables a repositorios públicos.

El registro de decisiones es tan importante como el script. Debe explicar por qué se agrupó una raza, se corrigió una unidad, se excluyó una clínica o se modificó una definición. Para modelos se registra versión de datos, variables, partición, semilla, paquetes, hiperparámetros y artefacto final. Plataformas como Anna ilustran que integrar clasificadores con historias veterinarias exige una infraestructura explícita entre el modelo y el sistema clínico, no solo un archivo entrenado (Kong et al., 2025).

6.16. Caso didáctico depuración de una base de historias clínicas caninas

Planteamiento

Caso didáctico simulado. Una red de tres clínicas aporta historias de urgencias caninas de enero de 2023 a diciembre de 2025. El objetivo es desarrollar un modelo que, con información disponible al finalizar la evaluación inicial, estime la probabilidad de hospitalización. Los datos y resultados siguientes son ficticios y se presentan únicamente con fines formativos.

La extracción inicial contiene 3.250 filas, 2.180 perros y 74 columnas. Cada fila pretende representar una visita, pero incluye duplicados de facturación. El desenlace `hospitalized` aparece como Sí, SI, 1, No, 0, observación y vacío. Peso, temperatura, frecuencia cardiaca, creatinina, motivo de consulta y hora de llegada son candidatos predictivos.

Pregunta e hipótesis metodológica

Pregunta: ¿puede una base multicéntrica depurada generar estimaciones reproducibles del riesgo de hospitalización sin usar información posterior a la decisión?

Hipótesis: un pipeline que preserve la unidad perro-visita, normalice semántica y unidades, trate la ausencia dentro del remuestreo y divida por animal y tiempo producirá una evaluación menos optimista, pero más válida, que una limpieza directa sobre toda la base.

Paso 1. Congelación y auditoría inicial

Se guarda la extracción original como archivo de solo lectura con suma de verificación. El informe inicial identifica:

- 3.250 filas y 3.041 `visit_id` distintos;
- 128 duplicados exactos y 81 filas repetidas por reenvío de facturación;
- 2.180 `animal_id`, de los cuales 487 tienen más de una visita;
- 17 formatos de raza y 11 representaciones de sexo/estado reproductivo;
- peso ausente en 9,8%, temperatura en 28,4% y creatinina en 62,1%;
- 35 pesos registrados en gramos bajo una columna rotulada kg;
- 22 temperaturas en Fahrenheit y 14 valores incompatibles con vida;
- una clínica cambia de analizador de creatinina en julio de 2024;
- la orden de hospitalización y el código de cama se registran después de la decisión y revelan el desenlace.

Paso 2. Definición de entidades y tiempo índice

Se establece una fila por visita de urgencias. Varias facturas se agregan a la visita, pero no se usan como predictores porque algunas se emiten después del ingreso. El tiempo índice es la firma de evaluación inicial. Solo se aceptan signos vitales medidos desde 60 minutos antes hasta ese momento y el primer resultado de laboratorio disponible antes del índice.

La división de desarrollo incluye visitas de 2023–2024; 2025 se reserva como prueba temporal. Dentro del desarrollo, la validación cruzada agrupa por `animal_id`, de modo que todas las visitas de un perro permanecen en el mismo pliegue. Una segunda evaluación deja una clínica fuera para explorar transportabilidad geográfica.

Paso 3. Deduplicación

Los 128 duplicados exactos se eliminan conservando un registro en la capa limpia. Las 81 filas de facturación se consolidan por `visit_id`; se retiene la versión clínica más reciente previa al cierre y se conserva una tabla hija de transacciones. Tres pares con identificadores diferentes pero mismo microchip y hora se revisan manualmente: dos son duplicados y uno corresponde a dos perros del mismo tutor.

Cada acción se registra con `record_id_original`, regla aplicada, decisión, fecha y responsable. El resultado no borra la evidencia de origen.

Paso 4. Armonización

Se separan sexo y estado reproductivo. Las etiquetas de raza se mapean a un catálogo, manteniendo texto original y una categoría "cruce no especificado" cuando no existe precisión. Pesos con unidad explícita en gramos se convierten a kg; los valores dudosos se contrastan con notas. Temperaturas entre 90 y 115, acompañadas por dispositivo Fahrenheit, se convierten a Celsius; los valores sin evidencia quedan como ausentes con razón `error_no_resuelto`.

La creatinina se convierte a mg/dL cuando procede y se añade `analyzer_period`. No se fuerza equivalencia entre analizadores; el periodo se conserva para evaluar desplazamiento. Los diagnósticos se normalizan sin usar el diagnóstico de alta como predictor.

Paso 5. Reglas de calidad y atípicos

Se crean banderas, no eliminaciones automáticas. Peso <0,5 o >100 kg; temperatura <34 o >43 °C; frecuencia cardiaca <20 o >300 lpm; y creatinina >20 mg/dL pasan a revisión. De 96 alertas, 61 son plausibles por especie y gravedad, 21 se corrigen con la nota y 14 quedan sin resolver.

Los 61 extremos plausibles se conservan. Se comparará un escalador estándar con uno robusto dentro del pipeline. Esta decisión protege pacientes críticos de una limpieza basada solo en cuantiles.

Paso 6. Ausencia e imputación

La creatinina falta sobre todo en casos leves y en una clínica sin laboratorio nocturno. Por ello, el patrón no es MCAR. Para un modelo de despliegue se incluyen una bandera `creatinine_ordered_before_index` y un imputador por mediana ajustado solo en entrenamiento, estratificado por clínica cuando existe suficiente información. Como análisis alternativo se prueba un modelo sin creatinina, útil para centros sin laboratorio.

Peso y temperatura se imputan dentro de cada pliegue mediante ecuaciones encadenadas en un análisis de sensibilidad y mediante mediana en el pipeline principal. Se comparan calibración y desempeño. No se imputa el desenlace: 23 visitas sin resolución verificable se excluyen del desarrollo del modelo y se describen.

Paso 7. Codificación e ingeniería

Las variables categóricas se codifican one-hot. Razas con pocos casos se representan también mediante grupo de tamaño definido antes del análisis. Se derivan edad al tiempo índice, horario nocturno, número de visitas previas en 90 días, frecuencia cardiaca relativa a grupos de edad y cambio de peso respecto a la última visita disponible.

Se excluyen código de cama, orden de ingreso, tarifa hospitalaria, diagnóstico de alta y medicación administrada después del índice. Una comparación exploratoria muestra que incluir `bed_order` eleva artificialmente el AUROC; se documenta como demostración de fuga, no como candidato final.

Paso 8. Desbalance y evaluación

Tras depuración quedan 2.996 visitas con desenlace, de las cuales 252 fueron hospitalizadas (8,4%). Se usa ponderación de clases durante entrenamiento; no se altera la prevalencia de la prueba temporal. El umbral se selecciona en validación para obtener sensibilidad cercana a 0,85, y luego se fija.

Se informan AUROC, área bajo precisión-recall, sensibilidad, especificidad, valores predictivos, puntuación de Brier y curva de calibración. Las métricas se acompañan de intervalos de confianza

y resultados por clínica. El modelo no se declara clínicamente útil solo por discriminar: se requiere calibración, análisis de beneficio neto, validación externa y evaluación de impacto.

Tabla 6.6.

Resultado de la depuración simulada

Etapa	Filas	Perros	Eventos	Observación
Extracción inicial	3.250	2.180	no confiable	mezcla clínica y facturación
Tras deduplicación	3.039	2.178	263	dos animales fusionados correctamente
Tras criterios temporales	3.019	2.166	258	se excluyen visitas fuera de ventana
Con desenlace verificable	2.996	2.153	252	prevalencia 8,4%
Desarrollo 2023–2024	2.087	1.587	181	remuestreo agrupado por perro
Prueba temporal 2025	909	715	71	prevalencia preservada

Interpretación crítica

La reducción de 3.250 a 2.996 filas no representa una “mejora” por sí misma. Cada exclusión debe justificarse. El mayor riesgo inicial no eran los valores extremos, sino la mezcla de unidades, duplicados y variables posteriores al desenlace. Un procedimiento menos cuidadoso habría producido una exactitud aparente elevada al aprender decisiones administrativas.

La alta ausencia de creatinina no autoriza a eliminar 62% de visitas ni a imputar ignorando el flujo clínico. El modelo alternativo sin laboratorio permite evaluar equidad de implementación entre centros. La validación temporal reconoce que protocolos, personal y equipos cambian. La agrupación por perro evita que el historial del mismo animal contamine la evaluación.

Esquema reproducible del pipeline

1. verificar integridad de la extracción;
2. separar tablas por entidad;
3. aplicar criterios de inclusión y tiempo índice;
4. dividir desarrollo y prueba por tiempo y animal;

5. ajustar transformaciones dentro de entrenamiento;
6. validar reglas clínicas y revisar alertas;
7. entrenar y ajustar umbral con validación;
8. bloquear el pipeline;
9. evaluar una sola vez en prueba temporal; y
10. archivar datos, código, diccionario, decisiones y resultados.

6.17. Síntesis autoevaluación y actividad práctica

Síntesis

Los datos veterinarios son productos de sistemas clínicos, productivos, diagnósticos y experimentales. Su valor científico depende de reconocer quién o qué generó cada registro, bajo qué proceso, en qué tiempo y con qué incertidumbre. La preparación responsable conserva originales, separa entidades, documenta significados, examina calidad por dimensión y evita correcciones no verificables.

Los valores atípicos no son sinónimo de error; la ausencia no es una categoría única; la imputación no recupera información perdida; y el escalamiento no resuelve mala medición. En aprendizaje automático, todas las transformaciones guiadas por datos se estiman dentro del entrenamiento. La división debe respetar animal, hato, clínica y tiempo. Los datos desbalanceados exigen métricas apropiadas y umbrales ligados al uso clínico.

Interoperabilidad y reproducibilidad requieren terminologías, identificadores, metadatos, versiones y código ejecutable. Ningún estándar reemplaza el juicio clínico, pero sin estándares el significado se degrada al intercambiar información. La calidad debe planificarse desde la captura: depurar al final es más costoso y nunca puede reconstruir completamente aquello que no se registró.

Preguntas de autoevaluación

1. ¿Por qué una historia clínica electrónica no puede interpretarse automáticamente como una base epidemiológica representativa?
2. Diferencie paciente, visita, episodio y muestra mediante un ejemplo clínico.
3. ¿Qué campos mínimos debe incluir un resultado de laboratorio para ser reutilizable?
4. Explique por qué varias radiografías del mismo perro no deben repartirse entre entrenamiento y prueba.
5. Diseñe cinco campos de un diccionario para estudiar mastitis subclínica bovina.
6. ¿Qué diferencia existe entre validez, consistencia y exactitud?
7. Proponga una regla para detectar duplicados de visitas sin fusionar animales homónimos.
8. ¿Cuándo un valor extremo de creatinina debe conservarse?
9. Compare MCAR, MAR y MNAR con ejemplos veterinarios.
10. ¿Por qué imputar antes de dividir entrenamiento y prueba constituye fuga de información?
11. ¿Cuándo elegiría escalamiento robusto en lugar de estandarización z?
12. ¿Qué problema produce codificar razas con números enteros consecutivos?
13. Mencione tres características derivadas que podrían anticipar mastitis sin usar información futura.
14. ¿Por qué 98% de exactitud puede ser inútil en una enfermedad con 2% de prevalencia?
15. ¿Qué diferencia existe entre interoperabilidad técnica y semántica?
16. ¿Qué elementos permiten reconstruir una matriz analítica seis meses después?
17. Identifique cuatro variables con fuga potencial en un modelo de hospitalización.
18. ¿Cómo evaluaría si la ausencia de temperatura difiere por clínica y gravedad?
19. ¿Qué daños podría causar aplicar SMOTE antes de separar la prueba?
20. Explique por qué FAIR no implica publicar historias clínicas identificables.

Respuestas orientativas

1. Porque se genera durante la atención, depende del acceso y de decisiones clínicas, y contiene medición y codificación selectivas.
2. El perro es paciente; cada consulta es visita; varias consultas por la misma dermatitis forman un episodio; cada tubo remitido es una muestra.
3. Animal, muestra, tiempo de toma, analito, método, valor, unidad, límites y estado de validación.
4. Comparten paciente, adquisición y contexto; dividirlas filtra información y sobrestima generalización.
5. Por ejemplo: `cow_id`, fecha de muestreo, recuento celular, cuarto mamario y método diagnóstico, con unidad y dominio.
6. Validez es cumplir el dominio; consistencia es no contradecir otros datos; exactitud es proximidad al valor verdadero.
7. Combinar identificador estable o microchip, hora, clínica y evidencia de episodio, con revisión de conflictos.
8. Cuando es verificable, clínicamente plausible y no proviene de unidad o transcripción errónea.
9. MCAR no depende de información; MAR depende de observados; MNAR conserva dependencia del no observado tras condicionar.
10. Porque los parámetros de prueba influyen en la transformación aprendida por el modelo.
11. Ante distribuciones con extremos legítimos que afectarían fuertemente media y desviación.
12. Introduce orden y distancias artificiales entre categorías nominales.
13. Descenso de producción previo, cambio de conductividad y reducción de rumia en una ventana anterior.
14. Puede corresponder a predecir siempre la clase mayoritaria y tener sensibilidad nula.
15. La primera permite transmitir; la segunda conserva significado.

16. Fuentes congeladas, diccionario, código, versiones, parámetros, registro de decisiones y sumas de verificación.
17. Código de cama, orden de ingreso, diagnóstico de alta y medicación posterior.
18. Comparando porcentajes y modelos de ausencia por clínica, signos basales, horario y desenlace, sin afirmar que ello prueba MAR.
19. Generaría sintéticos relacionados con prueba y produciría una estimación optimista; además alteraría su prevalencia.
20. FAIR admite autenticación y restricciones; la reutilización debe respetar ética, confidencialidad y gobernanza.

Actividad práctica integradora

Título: Auditoría y preparación de una base veterinaria para inferencia y predicción.

Escenario: El docente entrega una base simulada de 1.200 consultas de perros y gatos con identificadores de animal y visita, clínica, fecha, especie, raza, sexo, edad, peso, temperatura, hemograma, diagnóstico, tratamiento y reconsulta a 7 días. La base incluye duplicados, unidades mezcladas, ausencia, extremos, categorías inconsistentes y variables posteriores.

Producto esperado: informe técnico de 8 a 12 páginas, diccionario de datos, base limpia, script reproducible y diagrama del flujo.

Tareas:

1. formular una pregunta inferencial y otra predictiva;
2. definir población, unidad de análisis, tiempo índice y desenlace;
3. construir un diccionario para al menos 20 variables;
4. elaborar un perfil de calidad inicial por dimensión;
5. detectar duplicados con clave y reglas justificadas;
6. armonizar fechas, unidades y categorías conservando valores originales;

7. clasificar razones de ausencia y discutir mecanismos plausibles;
8. comparar análisis completo, imputación simple y una estrategia múltiple o dentro de pipeline;
9. crear tres características con fundamento clínico y descartar dos por fuga;
10. dividir datos respetando el animal y, cuando proceda, el tiempo;
11. evaluar desbalance y seleccionar métricas;
12. presentar un registro de decisiones y limitaciones.

Rúbrica de evaluación

Criterio	Peso	Evidencia de logro
Definición metodológica	20%	pregunta, unidad y tiempo índice coherentes
Diccionario y semántica	15%	definiciones, unidades, dominios y procedencia completas
Auditoría de calidad	20%	indicadores, patrones y reglas reproducibles
Tratamiento de ausencia y extremos	15%	supuestos explícitos y sensibilidad
Prevención de fuga	15%	partición correcta y pipeline encapsulado
Reproducibilidad	10%	script, versiones y registro de decisiones
Interpretación veterinaria	5%	conclusiones clínicas prudentes y limitaciones

Cierre crítico

Antes de aceptar una base como “lista”, el investigador debe poder responder: ¿qué representa cada fila?, ¿qué información existía en el tiempo de decisión?, ¿qué valores se modificaron y por qué?, ¿qué población quedó fuera?, ¿cómo se trataron la ausencia y los extremos?, ¿puede otra persona repetir el proceso? Si alguna respuesta depende de memoria personal o de cambios manuales no registrados, el conjunto todavía no está preparado para producir evidencia científica confiable.

CAPÍTULO 7

APRENDIZAJE AUTOMÁTICO SUPERVISADO EN INVESTIGACIÓN VETERINARIA

Introducción

El aprendizaje automático supervisado estudia relaciones entre observaciones cuyas respuestas son conocidas para construir una función capaz de predecir la respuesta de nuevos individuos. En investigación veterinaria, una observación puede ser un perro atendido en urgencias, una vaca durante un periodo de lactancia, una granja porcina, una muestra histológica o un intervalo temporal captado por un sensor. El término supervisado no alude a la vigilancia humana permanente del algoritmo, sino a la disponibilidad de una etiqueta o desenlace contra el cual se estima el error.

Una predicción útil comienza antes de programar. Exige precisar quién será evaluado, en qué momento se emitiría el pronóstico, qué información estaría disponible entonces y qué decisión podría modificarse. Si el objetivo consiste en anticipar hospitalización al finalizar el triaje, no deben entrar como predictores tratamientos, diagnósticos o resultados obtenidos después de la decisión de ingresar. Incorporarlos produce fuga de información y una estimación artificialmente alta del rendimiento.

Los métodos descritos en este capítulo abarcan regresión, árboles, ensambles, máquinas de vectores de soporte y vecinos cercanos. Ningún algoritmo domina en todos los conjuntos de datos. La regresión logística puede superar a un ensamble complejo cuando la muestra es modesta, la señal es aproximadamente aditiva y la medición contiene ruido. A la inversa, random forest o boosting pueden capturar umbrales e interacciones difíciles de especificar de antemano. La comparación válida requiere el mismo conjunto de pacientes, idénticos remuestreos, preprocesamiento aprendido solo con datos de entrenamiento y métricas alineadas con el costo clínico de los errores.

La literatura veterinaria reciente muestra usos plausibles y también advertencias. Se han empleado XGBoost para estimar el riesgo de recuento elevado de células somáticas después del parto, modelos multimodales para pronosticar mortalidad en animales de compañía y aprendizaje automático sobre historias electrónicas para predecir enfermedad valvular mitral canina (Burton et al., 2024; Kim et al., 2023; Thompson et al., 2023). Estos trabajos ilustran que la validación

externa, la calibración y la definición temporal del desenlace importan tanto como la elección del algoritmo.

Objetivos de aprendizaje

1. Traducir una pregunta veterinaria en una tarea de regresión o clasificación con población, horizonte temporal y uso previstos.
2. Distinguir predictores disponibles en el momento de uso de variables posteriores que causarían fuga de información.
3. Explicar el funcionamiento, los supuestos y las limitaciones de regresión, árboles, random forest, boosting, SVM y k vecinos.
4. Diseñar particiones de entrenamiento, validación y prueba que respeten animales, granjas, centros y tiempo.
5. Ajustar hiperparámetros dentro de un esquema de remuestreo sin contaminar la evaluación final.
6. Diagnosticar sobreajuste y subajuste mediante brechas de rendimiento y curvas de aprendizaje.
7. Comparar modelos con discriminación, calibración, error, utilidad clínica e incertidumbre.
8. Documentar un estudio predictivo conforme con principios de transparencia y evaluación del riesgo de sesgo.

7.1 Formulación de un problema predictivo

La unidad de análisis de un problema predictivo debe coincidir con la entidad sobre la cual se tomará una decisión. Si se desea predecir mastitis clínica en los próximos 30 días, la unidad puede ser la vaca-lactancia o la vaca-periodo, no cada medición de conductividad como si fuese un animal independiente. En imágenes radiográficas, la unidad clínica suele ser el paciente o el estudio, aunque cada paciente aporte varias proyecciones. La partición por imagen mezclaría vistas del mismo individuo entre entrenamiento y prueba y permitiría que el modelo reconociera rasgos propios del paciente o del equipo en lugar de la lesión.

Una formulación completa responde siete preguntas: población diana, momento cero, desenlace, horizonte, predictores disponibles, contexto de uso y acción vinculada. El momento cero es la frontera temporal desde la cual se observa el futuro. En urgencias de pequeños animales podría ser el cierre del examen inicial; en reproducción bovina, la inseminación; en vigilancia epidemiológica, el final de una semana. Las variables registradas después de ese punto no pueden emplearse para una predicción prospectiva salvo que el modelo se redefina como actualización dinámica.

Conviene distinguir clasificación de diagnóstico y clasificación de pronóstico. El diagnóstico intenta estimar un estado presente, por ejemplo infección intramamaria al momento del muestreo. El pronóstico estima un evento posterior, como recuento celular elevado durante los 30 días posteriores al parto. La distinción cambia el conjunto de predictores admisibles, el patrón de validación y la interpretación. También evita denominar temprana a una detección basada en pruebas que ya confirman el desenlace.

El producto del modelo puede ser una clase, una probabilidad o un valor continuo. Para apoyar decisiones, la probabilidad suele ser preferible a la etiqueta rígida porque permite elegir umbrales según recursos y consecuencias. Un hospital con pocas jaulas podría priorizar especificidad; una unidad que pretende no omitir pacientes sépticos puede exigir sensibilidad alta. El umbral pertenece a la política clínica, no es una propiedad universal del algoritmo.

Tabla 7.1

Plantilla para formular una pregunta predictiva veterinaria

Componente	Pregunta operativa	Ejemplo
Población	¿En qué animales se aplicará?	Perros y gatos atendidos en urgencias de primer ingreso
Momento cero	¿Cuándo se emite la predicción?	Al terminar triaje y examen físico inicial
Desenlace	¿Qué se predice y cómo se verifica?	Hospitalización por al menos 12 horas según registro de ingreso
Horizonte	¿Dentro de qué periodo?	Decisión durante las seis horas siguientes
Predictores	¿Qué datos existen en ese momento?	Especie edad signos vitales perfusión estado mental y pruebas rápidas

Uso	¿Quién recibe la estimación?	Veterinario de urgencias como apoyo no sustitución del juicio
Acción	¿Qué cambia según el riesgo?	Prioridad de evaluación monitorización y preparación de recursos

La estimación del efecto causal de una intervención es otra pregunta. Un modelo que predice hospitalización a partir de oxigenoterapia no demuestra que el oxígeno cause ingreso; probablemente capta gravedad. Para decidir si una intervención modifica el desenlace se requiere un diseño causal, con temporalidad y control de confusión. La capacidad de anticipar y la capacidad de explicar mecanismos no deben confundirse.

7.2 Variables predictoras y variable objetivo

Las variables predictoras, también llamadas características, atributos o covariables, representan información disponible para el algoritmo. Pueden ser continuas, ordinales, nominales, binarias, temporales, espaciales, textuales o derivadas de imágenes y señales. En un estudio de hospitalización, frecuencia cardiaca y temperatura son continuas; estado mental es ordinal; especie es nominal; presencia de disnea es binaria. La codificación debe conservar el significado y permitir que el procedimiento se aplique del mismo modo a nuevos pacientes.

La variable objetivo necesita una definición reproducible. Hospitalización podría significar permanencia superior a 12 horas, ingreso formal a cuidados intensivos o cualquier observación nocturna. Cada definición responde a un problema diferente. Si la decisión depende del criterio del clínico, el modelo aprende también patrones de práctica, disponibilidad de camas, capacidad económica del tutor y cultura institucional. Esos factores deben reconocerse porque pueden limitar la transportabilidad o reproducir desigualdades de acceso.

El error de etiqueta establece un techo al rendimiento. En dolor felino, por ejemplo, la etiqueta puede derivarse de una escala validada aplicada por observadores entrenados, pero conserva variación interevaluador. Steagall et al. (2023) combinaron localización de puntos faciales y XGBoost para estimar puntuaciones de la Feline Grimace Scale; el resultado muestra el valor del etiquetado estructurado, aunque la generalización depende de la diversidad de imágenes, pacientes y condiciones de captura.

La selección de predictores debe precederse por una auditoría temporal. Se clasifica cada variable como disponible antes del momento cero, disponible exactamente en ese momento, posterior, o imposible de obtener de forma rutinaria. Las variables posteriores se excluyen. Las variables costosas pueden conservarse para un modelo ampliado y compararse con un modelo parsimonioso. Esta comparación informa si la mejora justifica pruebas adicionales, tiempo o invasividad.

Los datos jerárquicos requieren especial atención. Varias visitas del mismo animal, animales dentro de granjas y estudios dentro de hospitales no son observaciones independientes. El identificador de agrupación debe guiar la partición. Un identificador administrativo rara vez debe entrar como predictor, pero sí debe mantenerse para impedir que una misma entidad aparezca en entrenamiento y prueba. La raza también exige juicio: puede captar predisposición biológica, pero a veces funciona como sustituto de estructura poblacional, prácticas de crianza o acceso a servicios.

Tabla 7.2

Riesgos frecuentes en la construcción de variables

Riesgo	Ejemplo veterinario	Consecuencia	Prevención
Fuga temporal	Usar duración de hospitalización para predecir ingreso	Rendimiento imposible en uso real	Congelar predictores en el momento cero
Duplicación	Varias filas de la misma visita	Intervalos demasiado optimistas	Definir una clave paciente visita
Dependencia	Imágenes del mismo animal en ambos conjuntos	Memorización de identidad o adquisición	Partir por paciente o centro
Etiqueta ambigua	Hospitalizado incluye observación breve y UCI	Objetivo heterogéneo	Definición clínica y auditoría manual
Disponibilidad diferencial	Prueba de lactato solo en casos graves	Ausencia informativa y sesgo de proceso	Modelar indicador de medición y validar
Proxy sensible	Código postal como acceso económico	Riesgo de inequidad	Evaluación por subgrupos y justificación

7.3 Regresión lineal y logística como modelos predictivos

Regresión lineal

La regresión lineal predice una respuesta continua mediante una combinación ponderada de predictores. Su forma básica es $\hat{y} = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$. En veterinaria podría estimar producción diaria de leche, duración de recuperación, concentración de un analito o peso de canal. Los coeficientes expresan el cambio esperado en la respuesta por unidad del predictor, manteniendo constantes los demás, siempre que la especificación sea adecuada. Esa interpretación condicional no convierte la asociación en causalidad.

Como modelo predictivo, la regresión lineal se evalúa fuera de muestra con error absoluto medio, raíz del error cuadrático medio y, con cautela, R^2. El MAE conserva las unidades clínicas y pondera linealmente los errores; el RMSE penaliza más los errores grandes. Un RMSE de 0,30 °C puede ser aceptable para vigilancia térmica bovina y, al mismo tiempo, insuficiente para detectar cambios muy pequeños. La relevancia depende del error de medición y del uso previsto.

La linealidad se refiere a la forma en que los parámetros entran al modelo, no a que toda relación biológica deba ser una recta. Transformaciones, términos polinomiales, splines e interacciones permiten representar curvatura. Sin embargo, cada grado de flexibilidad consume información. La regularización ridge o lasso estabiliza modelos con muchos predictores correlacionados; sus parámetros deben ajustarse dentro de validación cruzada.

Regresión logística

La regresión logística modela la probabilidad de un desenlace binario. Relaciona los predictores con el logit de la probabilidad: $\log[p/(1-p)] = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$. Al aplicar la función logística, la predicción queda entre 0 y 1. En el ejemplo de hospitalización, p representa la probabilidad estimada de ingreso para un paciente con un perfil determinado.

El exponencial de un coeficiente es una razón de momios condicional. Esta medida no equivale a riesgo relativo y puede parecer más grande cuando el desenlace es común. Para comunicación clínica conviene complementar coeficientes con probabilidades absolutas para perfiles plausibles. En un modelo centrado en predicción, la significación individual de cada coeficiente no constituye un criterio apropiado para eliminar variables; la selección paso a paso por valores p suele ser inestable.

La regresión logística es un punto de referencia exigente, no un método obsoleto. Puede entregar probabilidades bien calibradas, admite regularización, interacciones y términos no lineales, y facilita auditoría. TRIPOD+AI incluye tanto modelos basados en regresión como en aprendizaje automático dentro de un mismo marco de reporte, lo que ayuda a evitar la falsa oposición entre estadística e IA (Collins et al., 2024).

Tabla 7.3

Lectura predictiva de modelos de regresión

Aspecto	Lineal	Logística
Objetivo	Valor continuo	Probabilidad de clase
Ejemplo	Días hasta alta posoperatoria	Hospitalización sí o no
Función de pérdida común	Error cuadrático	Log loss o devianza binomial
Métricas	MAE RMSE R^2	AUC PR sensibilidad especificidad Brier calibración
Riesgo de especificación	Curvatura y heterocedasticidad	No linealidad en el logit e interacciones omitidas
Ventaja	Interpretación y línea base transparente	Probabilidades y recalibración sencilla

7.4 Árboles de decisión

Un árbol de decisión divide repetidamente el espacio de predictores. Cada nodo formula una regla, por ejemplo lactato mayor que cierto punto, y cada hoja produce un valor o una probabilidad. Para clasificación se buscan divisiones que reduzcan impureza, medida a menudo con Gini o entropía. Para regresión se minimiza la variación residual. El resultado se aproxima a una secuencia clínica de preguntas y permite representar umbrales e interacciones sin declararlos de antemano.

La interpretabilidad de un árbol depende de su tamaño. Un árbol corto puede discutirse en una sesión clínica; uno profundo contiene muchas excepciones, cambia ante pequeñas perturbaciones y deja de ser transparente. Los cortes aprendidos tampoco deben confundirse con umbrales fisiológicos universales. Si el algoritmo divide la frecuencia respiratoria en 38 respiraciones por minuto, ese punto optimiza una muestra y una función de pérdida; requiere validación y no reemplaza un criterio clínico establecido.

Los principales hiperparámetros son profundidad máxima, número mínimo de observaciones por división, tamaño mínimo de hoja y parámetro de complejidad para poda. Sin restricción, un árbol puede producir hojas con uno o pocos pacientes y ajustar ruido. La poda seleccionada mediante validación cruzada reduce varianza. Un árbol podado suele perder algo de ajuste aparente, pero gana estabilidad.

Los árboles manejan relaciones no lineales y variables de escalas distintas. Aun así, presentan alta varianza y probabilidades escalonadas. Por ello se utilizan con frecuencia como componentes de ensambles. Un árbol individual conserva valor como referencia interpretable, especialmente si se compara con regresión y ensambles bajo el mismo esquema de validación.

7.5 Random forest

Random forest construye muchos árboles sobre muestras bootstrap y, en cada división, examina un subconjunto aleatorio de predictores. La combinación reduce la correlación entre árboles y promedia su inestabilidad. En clasificación, cada árbol aporta una clase o probabilidad; en regresión, un valor continuo. Breiman (2001) formalizó este enfoque, que sigue siendo una referencia para datos tabulares con interacciones y no linealidades.

La aleatoriedad cumple dos funciones. El bootstrap cambia los individuos disponibles para cada árbol; el muestreo de variables evita que un predictor dominante determine todos los árboles. Los casos no incluidos en una muestra bootstrap forman observaciones out of bag y permiten una estimación interna del error. Esa estimación resulta útil durante exploración, pero no sustituye una prueba externa ni corrige particiones mal hechas por animal, granja o tiempo.

Los hiperparámetros incluyen número de árboles, cantidad de variables candidatas por nodo, profundidad y tamaño mínimo de hoja. Más árboles suelen estabilizar la estimación sin incrementar el sobreajuste del mismo modo que profundizar un árbol único, aunque elevan costo computacional. El tamaño de hoja controla suavidad: hojas pequeñas capturan detalles y ruido; hojas mayores producen probabilidades menos extremas.

La importancia por disminución de impureza favorece variables continuas o con muchas categorías. La importancia por permutación responde a cuánto cae el rendimiento al romper una

variable, pero puede repartir o distorsionar importancia cuando existen predictores correlacionados. Ninguna medida demuestra causalidad. En una base hospitalaria, que la administración de un fármaco aparezca como importante puede indicar gravedad y no un efecto nocivo del medicamento.

Random forest es apropiado como competidor general en datos tabulares, sobre todo si se esperan interacciones. Requiere menos escalamiento que SVM o kNN y tolera mezclas de relaciones. Sin embargo, puede necesitar calibración de probabilidades y presentar dificultades ante extrapolación. Para un valor continuo fuera del rango observado, el bosque no puede extrapolar de forma lineal porque promedia valores de hojas existentes.

7.6 Gradient boosting XGBoost y LightGBM

El boosting ajusta modelos débiles en secuencia. Cada nuevo árbol corrige parte del error acumulado por el conjunto anterior. En la formulación de gradient boosting, la corrección sigue el gradiente de una función de pérdida. Esta estrategia suele lograr alto rendimiento en datos tabulares, pero exige control del aprendizaje y de la complejidad porque muchos árboles profundos pueden adaptar ruido.

XGBoost incorpora regularización, manejo eficiente de estructuras dispersas, submuestreo y optimización computacional (Chen & Guestrin, 2016). LightGBM utiliza histogramas y crecimiento de árboles por hojas, con técnicas destinadas a reducir costo en grandes volúmenes (Ke et al., 2017). La rapidez no convierte a un método en científicamente superior. En muestras veterinarias pequeñas, una búsqueda extensa de hiperparámetros puede seleccionar configuraciones adaptadas al remuestreo.

Los hiperparámetros actúan de forma conjunta. Una tasa de aprendizaje baja requiere más árboles; profundidad, número de hojas y observaciones mínimas por hoja controlan complejidad; submuestreo de filas y columnas introduce aleatoriedad; regularización penaliza hojas extremas. Ajustar cada parámetro de manera aislada ignora estas interacciones. Una búsqueda aleatoria o bayesiana dentro de validación cruzada suele explorar el espacio con mayor eficiencia que una cuadrícula exhaustiva.

En bovinos lecheros, Thompson et al. (2023) desarrollaron un modelo XGBoost con datos rutinarios para estimar el riesgo de recuento elevado de células somáticas tras el parto y realizaron validación en un conjunto separado de hatos. El estudio informó calibración además de discriminación, una práctica especialmente pertinente cuando las probabilidades se emplean para agrupar animales según riesgo. Alsahaf et al. (2023) utilizaron XGBoost para predecir distocia antes de la inseminación; el desenlace fue poco frecuente y el AUC en datos no vistos fue 0,73, recordatorio de que una tarea útil puede conservar incertidumbre considerable.

XGBoost y LightGBM pueden modelar valores ausentes mediante reglas aprendidas, pero esto no elimina la necesidad de comprender el mecanismo de ausencia. Que una prueba no se haya solicitado puede contener información sobre decisiones previas del clínico. Si el patrón de solicitud cambia entre hospitales, el modelo puede degradarse. Debe compararse un enfoque con imputación explícita, indicadores de ausencia y análisis por centro.

Tabla 7.4

Parámetros de complejidad en ensambles de árboles

Familia	Parámetro	Efecto principal	Riesgo si es extremo
Random forest	mtry o max_features	Diversidad entre árboles	Muy bajo pierde señal muy alto correlaciona árboles
Random forest	min_n o min_samples_leaf	Tamaño mínimo de hojas	Muy pequeño produce predicciones inestables
Boosting	learning_rate	Magnitud de cada actualización	Alto sobreajusta bajo requiere muchos árboles
Boosting	n_estimators	Número de etapas	Excesivo con tasa alta adapta ruido
Boosting	max_depth o num_leaves	Interacciones y complejidad	Valores altos crean particiones específicas
Boosting	subsample y colsample	Regularización estocástica	Valores muy bajos elevan sesgo

7.7 Máquinas de vectores de soporte

Una máquina de vectores de soporte busca un límite de decisión con el mayor margen entre clases. En el caso lineal, solo algunas observaciones cercanas al límite, los vectores de soporte, determinan la solución. El parámetro C regula el compromiso entre un margen amplio y la penalización de clasificaciones erróneas. Un C muy alto intenta corregir casi todos los casos de entrenamiento y puede reducir generalización; un C bajo admite errores para obtener un límite más suave.

El truco kernel permite representar fronteras no lineales sin construir explícitamente todas las transformaciones. El kernel radial es frecuente y añade gamma, que controla la escala de influencia de cada observación. Gamma alto crea regiones locales y sinuosas; gamma bajo produce una frontera más lisa. C y gamma deben ajustarse conjuntamente dentro del remuestreo.

SVM necesita escalamiento porque distancia y producto interno dependen de unidades. Si el sodio se mide en centenas y una puntuación ordinal en unidades, la primera variable dominará sin estandarización. Media y desviación se estiman solo en el conjunto de entrenamiento de cada pliegue. Variables categóricas requieren codificación, usualmente one hot.

La salida natural de una SVM es una puntuación de distancia al margen, no una probabilidad. La conversión a probabilidades mediante calibración debe aprenderse sin reutilizar la prueba. SVM resulta competitiva en muestras moderadas y espacios de alta dimensión, por ejemplo perfiles espectrales o moleculares. En bases tabulares con muchas variables mixtas y ausencias, el preprocesamiento y la comunicación clínica pueden ser más laboriosos.

7.8 Métodos de vecinos cercanos

El método k vecinos más cercanos clasifica una observación según las etiquetas de los casos próximos o predice el promedio de sus respuestas. Su aparente sencillez concentra una decisión crítica: qué significa cercanía. La distancia euclidiana solo es razonable después de escalar variables numéricas y tratar categorías. También puede ser inadecuada cuando abundan dimensiones irrelevantes o correlacionadas.

Con k pequeño, el modelo sigue detalles locales y tiene varianza alta; con k grande, suaviza la frontera y puede ignorar subgrupos reales. Se puede ponderar cada vecino por la inversa de su distancia. No obstante, una alta proximidad matemática no garantiza semejanza clínica. Dos pacientes pueden parecer próximos en signos vitales y diferir en especie, edad o enfermedad de base. La selección de variables y una métrica compatible con datos mixtos deben justificarse.

La maldición de la dimensionalidad describe la pérdida de contraste entre distancias cuando aumenta el número de variables. En ese escenario, casi todos los puntos parecen lejanos. Reducción de dimensión o selección de variables puede ayudar, pero debe incluirse dentro de validación cruzada para evitar optimismo. kNN también conserva los datos de entrenamiento y puede ser lento al predecir si la base es grande.

En el estudio BrachySound, Oren et al. (2023) combinaron configuraciones de kNN y un árbol dentro de un ensamble para analizar sonidos respiratorios de perros braquicéfalos. El ejemplo muestra que un método basado en vecinos puede aportar señal cuando las características acústicas se han extraído y normalizado cuidadosamente. La unidad de partición y la generalización entre razas siguen siendo aspectos centrales.

7.9 Selección del algoritmo según el problema científico

La selección comienza con las restricciones del estudio, no con una clasificación de popularidad. Se consideran el tipo de desenlace, tamaño efectivo de muestra, número y naturaleza de predictores, estructura de agrupación, no linealidad esperada, costo de errores, necesidad de probabilidad calibrada, velocidad de predicción y capacidad de mantenimiento. Un método ligeramente menos discriminativo puede ser preferible si es estable, calibrado y desplegable en la clínica objetivo.

El tamaño efectivo no siempre coincide con el número de filas. Diez mil mediciones de 40 vacas contienen menos información independiente que diez mil vacas. De modo similar, miles de imágenes de cien perros no equivalen a miles de pacientes. La complejidad del algoritmo debe juzgarse frente al número de unidades independientes y eventos. PROBAST+AI orienta la evaluación de calidad, riesgo de sesgo y aplicabilidad de modelos construidos con regresión o técnicas de IA (Moons et al., 2025).

Se recomienda definir un conjunto pequeño de candidatos antes de observar la prueba: un modelo de referencia simple, uno o dos métodos flexibles y, cuando exista, una regla clínica establecida. La inclusión indiscriminada de decenas de algoritmos multiplica oportunidades de selección fortuita. La hipótesis científica puede sugerir la arquitectura: una relación temporal puede requerir rasgos longitudinales; un problema tabular no necesita una red profunda por defecto.

Tabla 7.5

Guía razonada para seleccionar candidatos

Condición del problema	Candidato inicial	Motivo	Comprobación necesaria
Respuesta continua relación aproximadamente suave	Regresión lineal regularizada	Línea base interpretable y estable	Residuos no linealidad error en unidades clínicas
Clasificación con probabilidad para decidir	Logística regularizada	Probabilidad directa y recalibración simple	Calibración interacciones y no linealidad
Reglas cortas exigidas por operación	Árbol podado	Secuencia comprensible	Estabilidad y pérdida de rendimiento
Datos tabulares con interacciones	Random forest o boosting	Flexibilidad sin especificar interacciones	Calibración y validación externa
Alta dimensión muestra moderada	SVM regularizada	Margen y kernels	Escalamiento ajuste de C y gamma
Patrones locales con pocas variables	kNN	Referencia no paramétrica	Métrica distancia escala y dimensionalidad

La interpretabilidad debe definirse por usuario y tarea. Un coeficiente, una regla, una contribución local y una explicación global responden preguntas distintas. SHAP puede describir cómo un predictor contribuye a una predicción en un modelo dado, pero no valida que el modelo use relaciones clínicamente legítimas. Burton et al. (2024) integraron texto y datos tabulares para predecir mortalidad en animales de compañía y extendieron explicaciones a ambas modalidades; este tipo de enfoque requiere auditar tanto el rendimiento como la plausibilidad de los patrones lingüísticos aprendidos.

7.10 Entrenamiento validación y prueba

El conjunto de entrenamiento se usa para estimar parámetros. El de validación orienta hiperparámetros, selección de variables, umbrales y elección entre modelos. El conjunto de prueba se reserva para una evaluación final. Si se consulta la prueba después de cada cambio, deja de ser prueba y se convierte en otro conjunto de validación.

Una partición aleatoria individual es adecuada solo cuando las futuras observaciones son intercambiables y no existen agrupaciones. En estudios multicéntricos, separar por hospital evalúa transportabilidad. En producción, separar por hato impide que rasgos propios de una granja aparezcan en ambos lados. En datos longitudinales, entrenar con periodos anteriores y probar en posteriores imita el uso prospectivo y detecta cambios de prevalencia, manejo o equipos.

La estratificación conserva aproximadamente la proporción de desenlaces, pero no debe romper grupos. Primero se asignan animales, hatos o centros; después se comprueba la distribución. Si hay pocos centros, una validación leave one centre out permite explorar heterogeneidad, aunque cada estimación tendrá incertidumbre. La prueba temporal o geográfica suele ser más informativa que apartar al azar un 20 por ciento de la misma fuente.

Tabla 7.6

Estrategias de partición

Estructura	Partición recomendada	Error que evita
Una fila por animal sin agrupación	Aleatoria estratificada	Desbalance accidental de clases
Varias visitas por animal	Por identificador de animal	El mismo paciente en entrenamiento y prueba
Animales dentro de granjas	Por granja o validación por granja	Aprender firma de manejo compartida
Varios hospitales	Centro externo o leave one centre out	Optimismo por protocolos y equipos comunes
Series temporales	Entrenamiento pasado prueba futura	Uso de información futura y deriva temporal
Imágenes múltiples por estudio	Por paciente o estudio	Duplicación de vistas y memorizar adquisición

Figura 7.1.

Flujo completo de aprendizaje supervisado veterinario



El preprocesamiento forma parte del modelo. Imputación, escalamiento, codificación, reducción de dimensión y selección de variables se estiman con entrenamiento y luego se aplican sin recalcular a validación y prueba. Una receta que calcula la media global antes de dividir filtra información. La misma regla se aplica a balanceo: SMOTE o submuestreo solo dentro de cada conjunto de entrenamiento, nunca antes de la partición.

7.11 Validación cruzada

La validación cruzada k fold divide el conjunto de desarrollo en k partes. Entrena con k−1 y evalúa en la restante, repitiendo hasta que cada parte haya servido de validación. El promedio estima rendimiento interno; la dispersión muestra sensibilidad a la partición. Cinco o diez pliegues son opciones comunes, no leyes. Repetir el esquema reduce dependencia de una sola división, con mayor costo computacional.

La validación cruzada debe reflejar la estructura de uso. Grouped cross validation mantiene todas las observaciones de un animal o hato en un mismo pliegue. La validación temporal usa ventanas que avanzan. La validación anidada reserva un ciclo exterior para estimar rendimiento y un ciclo

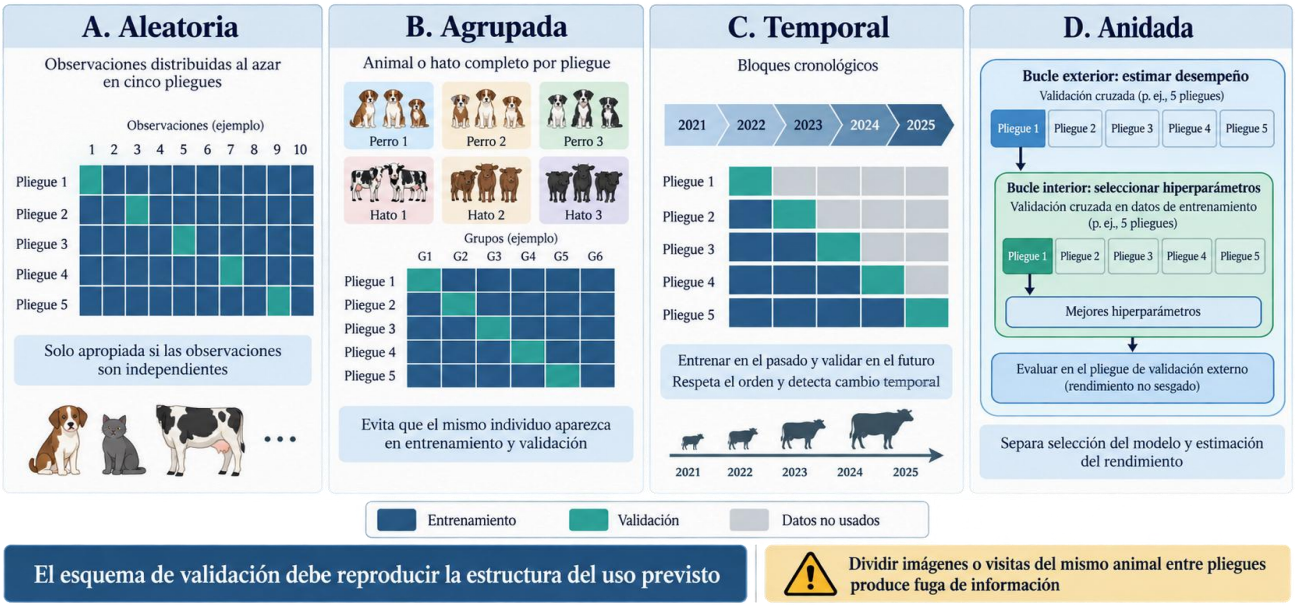
interior para ajustar hiperparámetros. Es especialmente útil cuando la muestra es limitada y se comparan muchas configuraciones.

Promediar métricas puede ocultar fallos. Conviene conservar resultados por pliegue y examinar si un centro o periodo concentra errores. Los intervalos de confianza deben respetar dependencia. Un intervalo estrecho obtenido tratando cientos de imágenes del mismo animal como independientes ofrece falsa precisión.

La validación interna mide reproducibilidad bajo la distribución disponible. No demuestra transportabilidad. La validación externa aplica el modelo congelado a animales de otra institución, región, periodo o sistema de registro. Thompson et al. (2023) separaron hatos para evaluar el modelo de salud de ubre, una decisión coherente con la aplicación a explotaciones distintas.

Figura 7.2.

Validación cruzada aleatoria, agrupada, temporal y anidada



7.12 Ajuste de hiperparámetros

Los parámetros se estiman directamente de los datos, como coeficientes logísticos o valores de hojas. Los hiperparámetros controlan cómo se realiza ese aprendizaje: penalización, profundidad,

número de árboles, tasa de aprendizaje, C, gamma o k. Elegirlos con el rendimiento de prueba produce sesgo de selección.

La búsqueda en cuadrícula prueba combinaciones predefinidas. Es transparente, pero desperdicia evaluaciones cuando muchos parámetros tienen poca influencia. La búsqueda aleatoria asigna un presupuesto y muestrea combinaciones, lo que explora más valores relevantes. La optimización bayesiana usa resultados previos para proponer nuevas configuraciones; reduce costo, aunque añade complejidad y todavía puede adaptar el remuestreo si se ejecuta sin límite.

El espacio de búsqueda debe justificarse por conocimiento y escala. Tasa de aprendizaje y regularización suelen explorarse en escala logarítmica. Los valores inviables se eliminan antes de correr el análisis. El presupuesto de ajuste se fija de antemano y se aplica por igual a métodos comparados, o se informa la diferencia. Dar miles de intentos a boosting y cinco a regresión no constituye una comparación neutral.

Después de seleccionar la configuración, se vuelve a entrenar el flujo completo con todos los datos de desarrollo y se evalúa una sola vez en prueba. Si se elige un umbral de clasificación, este se fija usando desarrollo. En modelos de probabilidad, la calibración también se aprende dentro del ciclo de entrenamiento; isotonic o Platt scaling no deben ajustarse en prueba.

Tabla 7.7

Ejemplos de hiperparámetros y escalas de búsqueda

Modelo	Hiperparámetros	Búsqueda inicial razonable	Observación
Logística regularizada	penalización y lambda	Lambda logarítmica	Estandarizar predictores numéricos
Árbol	profundidad min_n complejidad	Rangos cortos clínicamente plausibles	Podar mediante remuestreo
Random forest	mtry min_n árboles	mtry desde pequeño a fracción de p	Usar suficientes árboles para estabilizar
XGBoost LightGBM	tasa árboles profundidad hojas subsample	Aleatoria o bayesiana acotada	Vigilar interacción tasa número de árboles

SVM radial	C y gamma	Escala logarítmica	Escalamiento dentro del flujo
kNN	k ponderación distancia	Valores impares y rango limitado	Revisar dimensionalidad y escala

7.13 Sobreajuste y subajuste

Existe sobreajuste cuando el modelo aprende peculiaridades del conjunto de entrenamiento que no se repiten en nuevos datos. Se observa como rendimiento muy alto en entrenamiento y menor en validación o prueba. El subajuste ocurre cuando el modelo no representa suficiente señal y rinde mal incluso en entrenamiento. Ambos conceptos dependen del objetivo: un modelo puede discriminar bien y estar mal calibrado.

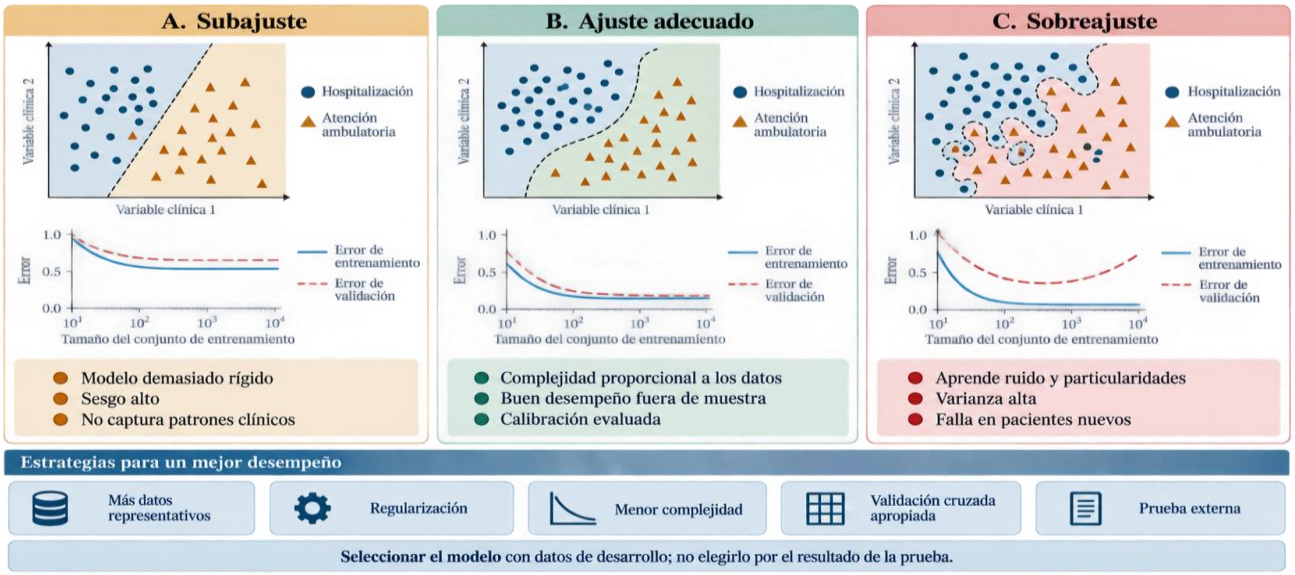
El sobreajuste surge por demasiada flexibilidad, pocos eventos, selección reiterada, fuga, etiquetas ruidosas o búsqueda extensa. La solución no consiste solo en cambiar algoritmo. Puede requerir más unidades independientes, mejor medición, regularización, reducción de variables, restricción de profundidad, detención temprana o un objetivo más estable. Un aumento masivo de filas correlacionadas no reemplaza diversidad.

Las curvas de aprendizaje representan desempeño frente a tamaño de entrenamiento. Si entrenamiento y validación son pobres y próximos, domina el sesgo; conviene mejorar características o flexibilidad. Si entrenamiento es alto y validación bajo, domina varianza; más datos independientes o regularización pueden ayudar. Si ambos se estabilizan separados, añadir pocas observaciones quizá no resuelva el problema.

Señales clínicas también revelan sobreajuste: probabilidades extremas sin fundamento, cambios abruptos ante una mínima variación, importancia dominada por identificadores o resultados demasiado buenos para la dificultad de la tarea. La auditoría de errores por especie, raza, sexo, edad, centro y gravedad permite detectar atajos. Un modelo puede mostrar AUC global aceptable y fallar en gatos si perros dominan la muestra.

Figura 7.3.

Subajuste, ajuste adecuado y sobreajuste



7.14 Comparación justa entre modelos

Los modelos se comparan sobre las mismas observaciones y los mismos pliegues. Cada flujo recibe idéntica información admisible, salvo que la comparación examine deliberadamente conjuntos de variables. Se documentan transformaciones específicas. Las métricas se eligen antes de conocer la prueba y deben corresponder a la decisión.

En clasificación binaria, AUC ROC mide la probabilidad de ordenar un caso positivo por encima de uno negativo y no depende de un umbral único. Con desenlaces raros puede parecer favorable aunque el valor predictivo positivo sea bajo; la curva precisión recall aporta otra perspectiva. Sensibilidad y especificidad describen funcionamiento en un umbral. Los valores predictivos dependen de prevalencia, por lo que cambian entre hospitales.

La calibración contrasta probabilidades con frecuencias observadas. El intercepto ideal es cero y la pendiente ideal es uno. El Brier promedia el error cuadrático de probabilidades y combina calibración con discriminación. Una curva de calibración identifica sobreconfianza o subconfianza. Si el modelo se usa para decidir, una buena AUC sin calibración puede asignar recursos de manera incorrecta.

La utilidad se evalúa relacionando errores con consecuencias. El análisis de curvas de decisión expresa beneficio neto para distintos umbrales y compara el modelo con intervenir en todos o en

ninguno. En veterinaria, los costos incluyen bienestar, retraso terapéutico, uso de camas, pruebas, tiempo del personal y carga económica para el tutor. Estos elementos no se reducen a una sola métrica.

Tabla 7.8

Métricas y significado veterinario

Métrica	Pregunta que responde	Limitación
AUC ROC	¿Ordena mayor riesgo en positivos que en negativos?	No evalúa calibración ni un umbral concreto
AUC PR	¿Mantiene precisión al recuperar positivos?	Depende de prevalencia
Sensibilidad	¿Qué proporción de hospitalizados identifica?	Aumenta al bajar el umbral
Especificidad	¿Qué proporción de no hospitalizados descarta?	Aumenta al subir el umbral
VPP y VPN	¿Qué significa una predicción para esta población?	Cambian con prevalencia y selección
Brier	¿Qué tan próximas son las probabilidades al desenlace?	Combina propiedades y requiere contexto
Pendiente calibración	¿Las probabilidades son demasiado extremas?	Necesita suficientes casos
Beneficio neto	¿Supera estrategias de tratar a todos o a ninguno?	Exige umbrales y consecuencias defendibles

Las diferencias de rendimiento deben acompañarse de incertidumbre. Cuando dos modelos se evalúan en los mismos pacientes, el análisis es pareado. Bootstrap por paciente o grupo puede estimar intervalos de la diferencia. Una diferencia de AUC de 0,01 quizá sea compatible con azar y no compense mayor complejidad. La elección final puede favorecer el modelo más simple si el desempeño clínicamente relevante es equivalente.

El reporte transparente debe incluir población, fuentes, fechas, exclusiones, manejo de faltantes, partición, hiperparámetros, métricas con intervalos, calibración, subgrupos, disponibilidad del modelo y limitaciones. TRIPOD+AI ofrece una lista de 27 ítems para estudios de desarrollo y

evaluación, mientras PROBAST+AI permite valorar calidad, riesgo de sesgo y aplicabilidad (Collins et al., 2024; Moons et al., 2025).

7.15 Caso didáctico predicción de hospitalización en pacientes veterinarios

Caso didáctico simulado. Todos los datos, resultados, métricas y decisiones numéricas de esta sección son ficticios y se presentan únicamente con fines formativos. No corresponden a una clínica real ni deben emplearse como regla asistencial.

Problema científico y propósito

Un hospital docente desea estimar, al concluir el triaje, la probabilidad de que un perro o gato sea hospitalizado durante las seis horas siguientes. El propósito operativo consiste en priorizar la evaluación, anticipar monitorización y preparar recursos. La predicción no decide el ingreso, no reemplaza al veterinario y no debe utilizarse para negar atención.

Pregunta: entre perros y gatos que acuden por primera vez al servicio de urgencias, ¿qué rendimiento tienen una regresión logística regularizada, un random forest y XGBoost para estimar hospitalización usando solo información disponible al finalizar el examen inicial? La hipótesis predictiva establece que los modelos flexibles pueden mejorar discriminación al capturar interacciones, pero la regresión puede conservar calibración y rendimiento comparable.

Población diseño y unidad de análisis

Se plantea una cohorte retrospectiva de visitas de urgencia registradas entre enero de 2023 y diciembre de 2025. La unidad es la visita índice. Si un animal presenta varias visitas, todas se asignan al mismo conjunto para evitar contaminación. Se excluyen traslados con ingreso ya decidido, pacientes muertos al arribo y visitas sin registro verificable del desenlace. El conjunto de desarrollo usa 2023 y 2024; 2025 se reserva como prueba temporal.

La base simulada contiene 2 400 visitas de 2 080 animales. El desenlace ocurre en 672 visitas, una prevalencia de 28 %. Desarrollo reúne 1 680 visitas y prueba temporal 720. La partición conserva animales completos. El tamaño no se justifica por una regla fija de eventos por variable; se

examinaría con métodos de tamaño muestral para modelos predictivos, considerando prevalencia, número de parámetros y sobreajuste esperado.

Definición del desenlace y predictores

Hospitalización se define como ingreso formal con permanencia prevista de al menos 12 horas, documentado dentro de las seis horas posteriores al momento cero. La etiqueta se contrasta con órdenes de ingreso y ocupación de cama. Dos revisores auditan una muestra aleatoria y los casos discordantes se resuelven por consenso.

Tabla 7.9

Diccionario resumido del caso didáctico simulado

Variable	Tipo	Momento	Tratamiento
hospitalización	Binaria objetivo	Dentro de 6 h	1 ingreso 0 alta o manejo ambulatorio
especie	Nominal	Registro	One hot
edad años	Continua	Registro	Spline o estandarización según modelo
frecuencia cardiaca	Continua	Triaje	Mediana plausible indicador de ausencia
frecuencia respiratoria	Continua	Triaje	Mediana plausible indicador de ausencia
temperatura	Continua	Triaje	Imputación dentro de pliegue
mucosas	Nominal	Examen	Normal pálida congestiva cianótica
llenado capilar	Ordinal	Examen	Categorías clínicas predefinidas
estado mental	Ordinal	Examen	Alerta deprimido estupor comatoso
disnea	Binaria	Examen	0 ausente 1 presente
dolor	Ordinal	Examen	Escala clínica documentada
lactato rápido	Continua	Antes del momento cero	Imputación e indicador de prueba solicitada
hospital y año	Agrupación	Sistema	Solo partición y auditoría no predictor

Se excluyen diagnóstico definitivo, tratamientos administrados después del triaje, duración de consulta, interconsulta, jaula asignada y facturación final porque ocurren después del momento cero o reflejan la decisión de ingreso. El costo estimado inicial se excluye del modelo principal por su relación con acceso económico; se reserva para análisis de equidad y proceso.

Plan de análisis

- Bloquear la prueba temporal antes de explorar modelos.
- Realizar validación cruzada agrupada de cinco pliegues, repetida tres veces, dentro del conjunto de desarrollo.
- Construir un flujo de preprocesamiento por pliegue: límites plausibles, imputación, indicadores de ausencia, codificación y escalamiento cuando corresponda.
- Comparar regresión logística elastic net, random forest y XGBoost con el mismo conjunto de predictores.
- Optimizar hiperparámetros por log loss y vigilar AUC y Brier; fijar el presupuesto de búsqueda antes de entrenar.
- Elegir el modelo con prioridad en calibración, sensibilidad a umbral clínico y parsimonia, no solo AUC.
- Fijar en desarrollo un umbral que logre sensibilidad cercana a 0,85, sujeto a una especificidad operativamente aceptable.
- Reentrenar el flujo seleccionado en todo desarrollo y evaluar una sola vez en 2025.

Resultados simulados de validación

Tabla 7.10

Rendimiento en prueba temporal resultados simulados

Modelo	AUC ROC	IC 95 %	Brier	Sensibilidad	Especificidad	Pendiente de calibración
Logística elastic net	0,842	0,811 a 0,873	0,148	0,85	0,66	0,97
Random forest	0,854	0,824 a 0,884	0,151	0,86	0,65	0,82
XGBoost	0,861	0,832 a 0,890	0,149	0,85	0,68	0,86

Los tres modelos discriminan de forma parecida. XGBoost alcanza la mayor AUC, pero sus probabilidades son demasiado extremas, como sugiere una pendiente de 0,86. La logística presenta el menor Brier y una pendiente próxima a uno. El intervalo de la diferencia de AUC entre XGBoost y logística, estimado por bootstrap agrupado, incluiría cero en este ejemplo. Por tanto, no se afirma superioridad del algoritmo complejo.

Para el umbral fijado en desarrollo, la regresión clasifica 720 visitas de prueba. Con 202 hospitalizaciones observadas y 518 no hospitalizaciones, los conteos simulados son 172 verdaderos positivos, 30 falsos negativos, 176 falsos positivos y 342 verdaderos negativos. La

sensibilidad es 172/202 = 0,851; la especificidad, 342/518 = 0,660; el VPP, 172/348 = 0,494; y el VPN, 342/372 = 0,919.

Tabla 7.11

Matriz de confusión del modelo logístico simulado

Predicción	Hospitalizado	No hospitalizado	Total
Riesgo alto	172	176	348
Riesgo bajo	30	342	372
Total	202	518	720

El VPN alto significa que, en una población con prevalencia de 28 %, una clasificación de riesgo bajo se asocia con una probabilidad reducida de hospitalización. No autoriza alta automática: 30 pacientes hospitalizados quedaron bajo el umbral. La revisión clínica de falsos negativos encuentra concentración de gatos con signos inespecíficos y temperatura no medida, lo que motiva revisar el proceso de triaje y evaluar rendimiento por especie.

Evaluación por subgrupos y utilidad

En los datos simulados, la sensibilidad es 0,88 en perros y 0,79 en gatos. La diferencia puede reflejar menor tamaño felino, manifestaciones clínicas distintas o medición menos completa. Antes de desplegar, se requiere ampliar la muestra, revisar variables felinas y confirmar la brecha con intervalos. Ajustar umbrales por especie solo sería justificable si mejora beneficio y no introduce trato inequitativo.

El análisis de decisión simulado muestra beneficio neto frente a ingresar a todos o a ninguno para umbrales de 0,18 a 0,42. Esa franja coincide con decisiones de priorización, no con una orden de ingreso. El hospital selecciona 0,24 para activar revisión temprana, después de considerar capacidad y costo de omitir un paciente grave. El protocolo conserva evaluación médica obligatoria para todos.

Modelo elegido y justificación

Se elige la regresión logística elastic net. Su AUC es apenas menor, las diferencias son inciertas, la calibración es mejor y su implementación resulta más sencilla. XGBoost se conserva como modelo secundario de investigación y se recalibra antes de cualquier estudio prospectivo. Esta elección ilustra que el algoritmo con el número más alto no siempre ofrece la mejor solución científica.

Antes del uso asistencial se requiere validación prospectiva silenciosa, evaluación en otro hospital, revisión ética y de gobernanza, documentación de cambios del sistema y un plan de vigilancia. Se monitorizarían prevalencia, datos ausentes, calibración, sensibilidad, carga de alertas y resultados por especie. Si cambia la política de ingreso o se incorpora una prueba rápida nueva, el modelo debe reevaluarse.

Código orientativo en R

El siguiente fragmento muestra la lógica de un flujo reproducible con tidymodels. Es un esquema didáctico; los nombres de variables y rangos deben adaptarse al diccionario real. La división por animal y la prueba temporal deben construirse antes del ajuste.

```
library(tidymodels)

receta <- recipe(hospitalizacion ~ especie + edad + fc + fr + temperatura +
                 mucosas + llenado_capilar + estado_mental + disnea +
                 dolor + lactato + lactato_medido,
                 data = desarrollo) |>
  step_impute_median(all_numeric_predictors()) |>
  step_unknown(all_nominal_predictors()) |>
  step_dummy(all_nominal_predictors()) |>
  step_zv(all_predictors()) |>
  step_normalize(all_numeric_predictors())

modelo <- logistic_reg(penalty = tune(), mixture = tune()) |>
  set_engine("glmnet") |>
  set_mode("classification")

flujo <- workflow() |> add_recipe(receta) |> add_model(modelo)

# Los pliegues deben agrupar por id_animal.
rejilla <- grid_latin_hypercube(
  penalty(range = c(-5, 0)), mixture(), size = 30
)

ajuste <- tune_grid(
  flujo, resamples = pliegues_agrupados, grid = rejilla,
  metrics = metric_set(roc_auc, pr_auc, mn_log_loss, brier_class)
)
```

```
mejor <- select_best(ajuste, metric = "mn_log_loss")
final <- finalize_workflow(flujo, mejor) |> fit(data = desarrollo)

# prueba_2025 no intervino en selección ni ajuste.
pred <- bind_cols(
  prueba_2025 |> select(hospitalizacion),
  predict(final, prueba_2025, type = "prob")
)
```

El flujo integra preprocesamiento y modelo, por lo que cada operación se estima dentro de los pliegues. En una implementación completa se añadirían análisis de calibración, intervalos por bootstrap agrupado, curvas de decisión y registro de versiones de datos y paquetes.

Limitaciones del caso

- El desenlace refleja decisiones clínicas y disponibilidad, no solo necesidad biológica de hospitalización.
- La naturaleza retrospectiva permite medición incompleta y cambios de protocolo.
- La prueba temporal procede del mismo hospital y no constituye validación geográfica externa.
- La prevalencia y los valores predictivos pueden cambiar en clínicas generales o centros con distinta casuística.
- Los resultados simulados no sustentan uso clínico y solo ilustran razonamiento metodológico.

7.16 Síntesis autoevaluación y actividad práctica

Síntesis

Un problema supervisado bien formulado define población, momento cero, desenlace, horizonte, predictores disponibles y acción. Esa definición controla la temporalidad y previene fuga. La unidad de partición debe seguir la independencia real de animales, visitas, granjas, hospitales e imágenes.

Regresión, árboles, random forest, boosting, SVM y kNN aportan sesgos inductivos distintos. La regresión ofrece una referencia estable y probabilidades fáciles de recalibrar; los árboles expresan reglas; los ensambles capturan interacciones; SVM resulta útil en alta dimensión; kNN representa semejanza local. La complejidad se justifica por mejora reproducible y utilidad, no por novedad.

El rendimiento honesto exige separar entrenamiento, validación y prueba; aprender todo preprocesamiento dentro del remuestreo; ajustar hiperparámetros sin consultar la prueba; y mantener grupos completos. La validación cruzada interna no reemplaza evaluación externa. Un modelo se juzga por discriminación, calibración, error, consecuencias, estabilidad y equidad entre subgrupos.

La predicción veterinaria permanece subordinada al método científico. Un algoritmo no corrige etiquetas vagas, mediciones deficientes, pseudorreplicación ni sesgo de selección. Tampoco convierte asociaciones en efectos causales. La transparencia recomendada por TRIPOD+AI y la evaluación de riesgo de sesgo de PROBAST+AI deben planificarse desde el protocolo, no añadirse después de obtener resultados.

Autoevaluación

1. ¿Qué elementos mínimos debe contener una pregunta predictiva veterinaria?
2. ¿Por qué el momento cero determina qué variables pueden utilizarse?
3. Diferencie parámetro de hiperparámetro con un ejemplo de regresión y otro de random forest.
4. ¿Por qué la regresión logística debe incluirse como línea base en muchos problemas tabulares?
5. Explique cómo random forest reduce la varianza de un árbol individual.
6. ¿Qué relación existe entre tasa de aprendizaje y número de árboles en gradient boosting?
7. ¿Por qué SVM y kNN requieren escalamiento de predictores?
8. ¿Qué partición usaría si cada perro aporta cinco radiografías y por qué?
9. Diferencie validación cruzada agrupada de validación externa.
10. ¿Cómo puede SMOTE producir fuga si se aplica antes de dividir los datos?
11. ¿Qué patrón de rendimiento sugiere sobreajuste?
12. ¿Por qué un AUC alto no garantiza probabilidades útiles?
13. ¿Cómo cambia el valor predictivo positivo cuando cambia la prevalencia?
14. ¿Qué información aporta la pendiente de calibración?

15. ¿Por qué una variable muy importante para XGBoost no debe interpretarse como causa?
16. ¿Cuándo elegiría un modelo más simple aunque tenga una AUC ligeramente menor?
17. ¿Qué grupos examinaría para evaluar desempeño diferencial en un hospital de pequeños animales?
18. ¿Qué debe permanecer congelado durante la evaluación final?
19. ¿Qué diferencia existe entre predicción de hospitalización y efecto causal de hospitalizar?
20. Mencione cuatro elementos que reportaría siguiendo TRIPOD+AI.

Respuestas orientativas

1. Población, momento cero, desenlace verificable, horizonte, predictores disponibles, contexto, usuario y acción prevista.
2. Porque simula el instante de uso; cualquier variable posterior introduce información que todavía no existiría.
3. Un coeficiente logístico es parámetro; la intensidad de penalización es hiperparámetro. En un bosque, los valores de hojas son parámetros y mtry es hiperparámetro.
4. Porque proporciona una referencia transparente, regularizable y frecuentemente bien calibrada.
5. Promedia árboles entrenados con muestras bootstrap y subconjuntos de variables, disminuyendo correlación e inestabilidad.
6. Una tasa menor suele requerir más árboles; ambos se ajustan conjuntamente.
7. Porque sus cálculos dependen de productos internos o distancias y las unidades pueden dominar el resultado.
8. Partición por perro, manteniendo todas sus imágenes juntas para impedir memorización del paciente.
9. La agrupada remuestrea la misma fuente respetando grupos; la externa prueba otra fuente, centro, región o periodo.

10. Las observaciones sintéticas pueden usar vecinos que luego quedan en validación, transfiriendo información.
11. Rendimiento alto en entrenamiento y claramente inferior en validación o prueba.
12. Porque AUC mide ordenamiento, no correspondencia entre probabilidad y frecuencia ni utilidad de un umbral.
13. En general aumenta con prevalencia, aun si sensibilidad y especificidad se mantienen.
14. Indica si las probabilidades son demasiado extremas o demasiado moderadas; el ideal es uno.
15. La importancia describe dependencia predictiva dentro del modelo y puede reflejar confusión, proxy o proceso de atención.
16. Cuando la diferencia es incierta o pequeña y el modelo simple ofrece mejor calibración, estabilidad o implementación.
17. Especie, edad, sexo, raza o conformación, centro, gravedad, turno y disponibilidad de pruebas, según pertinencia y ética.
18. El conjunto de prueba, la definición de métricas y el flujo seleccionado antes de la evaluación final.
19. La primera anticipa una decisión observada; la segunda pregunta qué ocurriría bajo una intervención y exige inferencia causal.
20. Fuentes y fechas, participantes y exclusiones, desenlace y predictores, faltantes, partición, ajuste, métricas con intervalos, calibración y disponibilidad.

Actividad práctica

Título: desarrollo y auditoría de un modelo de hospitalización veterinaria. Modalidad: equipos de tres o cuatro integrantes. Producto: protocolo analítico breve, cuaderno reproducible y reporte técnico de seis a ocho páginas. Puede utilizarse una base docente suministrada por la asignatura o una base simulada creada con un diccionario aprobado. Ningún dato clínico identificable debe cargarse en servicios no autorizados.

Situación de trabajo

El equipo recibe una base con visitas de perros y gatos, signos vitales, hallazgos de triaje, pruebas rápidas, identificadores de animal y fecha. Debe comparar al menos regresión logística regularizada y un método no lineal. La variable objetivo es hospitalización. Antes de programar, el equipo define el momento cero y marca variables posteriores.

Fases

- Formulación: redactar pregunta, población, unidad, horizonte, uso, desenlace y hipótesis predictiva.
- Auditoría: elaborar diccionario, detectar duplicados, rangos imposibles, faltantes, dependencia y fuga.
- Partición: reservar una prueba temporal o por centro y crear pliegues agrupados por animal.
- Modelado: construir un flujo reproducible para cada algoritmo con preprocesamiento dentro del remuestreo.
- Ajuste: definir por anticipado métricas, espacio de hiperparámetros y presupuesto de búsqueda.
- Evaluación: informar AUC ROC, AUC PR, Brier, calibración, sensibilidad, especificidad, VPP y VPN en un umbral justificado.
- Auditoría de errores: revisar falsos negativos y desempeño por especie y otro subgrupo pertinente.
- Decisión: elegir un modelo y justificarlo con rendimiento, incertidumbre, utilidad y factibilidad.
- Comunicación: preparar una ficha que indique uso previsto, población no cubierta, limitaciones y plan de vigilancia.

Preguntas de análisis

- ¿Qué variable parecía útil pero fue eliminada por ocurrir después del momento cero?
- ¿La mejor configuración cambia entre AUC y log loss?
- ¿Qué modelo presenta menor brecha entre validación y prueba?
- ¿Las probabilidades están calibradas o requieren recalibración?
- ¿Qué consecuencias tiene el umbral para camas, alertas y falsos negativos?
- ¿Qué cambio de prevalencia o protocolo podría degradar el modelo?
- ¿Qué estudio adicional sería necesario antes de uso clínico?

Tabla 7.12

Rúbrica de evaluación de la actividad

Criterio	Ponderación	Evidencia esperada
Formulación científica	15 %	Población momento cero desenlace horizonte y uso coherentes
Calidad y fuga	15 %	Diccionario auditoría temporal duplicados faltantes y grupos
Diseño de validación	20 %	Prueba intacta remuestreo agrupado y preprocesamiento interno
Modelado reproducible	15 %	Línea base método flexible hiperparámetros y semilla documentados
Evaluación	20 %	Discriminación calibración umbral incertidumbre y subgrupos
Interpretación clínica	10 %	Errores consecuencias utilidad limitaciones y no causalidad
Transparencia	5 %	Código versiones flujo y declaración de datos simulados o permisos

CAPÍTULO 8

APRENDIZAJE NO SUPERVISADO, SERIES TEMPORALES Y DETECCIÓN DE PATRONES

Introducción

Buena parte de los datos veterinarios no llega al investigador acompañada de una etiqueta fiable. Un acelerómetro registra miles de movimientos, una estación meteorológica acumula temperatura y humedad cada pocos minutos, un sistema de ordeño conserva curvas de producción y conductividad, y una historia clínica reúne combinaciones de signos, diagnósticos y tratamientos. En estos escenarios, el primer problema no siempre es predecir una categoría conocida, sino reconocer estructuras, trayectorias, regularidades y observaciones inusuales. El aprendizaje no supervisado, la minería de patrones y el análisis de series temporales proporcionan instrumentos para esa tarea, aunque responden a preguntas distintas y exigen formas de validación específicas.

El valor científico de estos métodos depende de la conexión entre el patrón matemático y el fenómeno biológico. Un algoritmo puede separar vacas según perfiles de actividad, pero no demuestra por sí solo que existan tipos naturales de comportamiento. Puede detectar una caída abrupta de rumia, sin distinguir automáticamente entre enfermedad, pérdida del collar, cambio de corral o fallo de transmisión. También puede pronosticar producción lechera con baja raíz del error cuadrático medio y, a la vez, ser poco útil si anticipa el descenso cuando la ventana de intervención ya terminó. Por eso, la exploración algorítmica debe integrarse con conocimiento clínico, diseño del estudio, metadatos, validación externa y evaluación de consecuencias.

Este capítulo articula tres familias de problemas. La primera comprende reducción de dimensionalidad y agrupamiento; la segunda, detección de anomalías y minería de coocurrencias; la tercera, series temporales y patrones espaciales o espaciotemporales. En todos los casos se mantiene una distinción central: describir una estructura encontrada en los datos no equivale a explicar su causa, y obtener un buen pronóstico no sustituye la inferencia sobre mecanismos.

Objetivos de aprendizaje

1. Diferenciar aprendizaje no supervisado, análisis exploratorio, predicción temporal e inferencia causal.
2. Aplicar e interpretar PCA, agrupamiento jerárquico, k-means, DBSCAN y métodos de reducción no lineal en datos veterinarios.

3. Diseñar procedimientos de detección de anomalías que separen eventos biológicos plausibles de fallos instrumentales.
4. Reconocer tendencia, estacionalidad, autocorrelación, dependencia longitudinal y rezagos en series de producción y fisiología.
5. Evitar fuga temporal y validar modelos mediante particiones cronológicas, evaluación por animal o finca y análisis prospectivo.
6. Integrar sensores, clima, ubicación y registros clínicos sin incurrir en pseudorreplicación.
7. Comunicar la incertidumbre y las limitaciones de grupos algorítmicos sin convertirlos en categorías biológicas rígidas.

8.1. Fundamentos del aprendizaje no supervisado

En el aprendizaje supervisado existe una variable objetivo observada durante el entrenamiento: mastitis sí/no, días de supervivencia, concentración de cortisol o producción diaria de leche. En el aprendizaje no supervisado esa etiqueta no guía el ajuste. El algoritmo utiliza similitudes, distancias, densidades, covarianzas o representaciones latentes para describir la geometría de los datos. Sus resultados habituales son componentes, grupos, reglas de asociación, representaciones de baja dimensión o puntuaciones de rareza.

La ausencia de etiquetas no implica ausencia de decisiones. El investigador elige variables, escala, métrica de distancia, número de grupos, tratamiento de valores faltantes y nivel de análisis. Estas decisiones pueden dominar el resultado. Si se agrupan registros horarios en lugar de animales, un mismo bovino puede aportar cientos de observaciones correlacionadas y aparentar evidencia masiva. Si se mezclan variables expresadas en litros, grados Celsius y conteos, la de mayor varianza numérica puede controlar la distancia euclídea. Si se incluyen identificadores de finca o marcas indirectas del manejo, los grupos pueden representar establecimientos y no fenotipos.

Un flujo defendible comienza con una pregunta exploratoria concreta, por ejemplo: ¿existen perfiles reproducibles de actividad diaria en vacas de transición? Luego se define la unidad analítica, se audita la calidad de los datos, se transforman variables sesgadas, se estandarizan magnitudes comparables y se prueban varias especificaciones razonables. La estabilidad se examina con remuestreo, perturbación de variables y replicación en otro periodo o rebaño. Finalmente, los patrones se contrastan con información externa no usada en el agrupamiento: examen clínico, condición corporal, paridad, diagnóstico o resultados productivos.

Tabla 8.1.

Diferencias entre objetivos analíticos relacionados

Objetivo	Pregunta típica	Salida	Criterio de evaluación
Exploración no supervisada	¿Qué estructura aparece sin usar una etiqueta?	Componentes, grupos, anomalías	Estabilidad, coherencia interna, utilidad biológica
Clasificación supervisada	¿A qué clase conocida pertenece el animal?	Probabilidad o clase	Sensibilidad, especificidad, AUC, calibración
Pronóstico temporal	¿Qué valor ocurrirá en un horizonte futuro?	Predicción con intervalo	MAE, RMSE, sesgo, cobertura temporal
Inferencia causal	¿Qué cambiaría bajo una intervención?	Efecto estimado	Identificación, control de confusión, sensibilidad

8.2. Análisis de componentes principales

El análisis de componentes principales (PCA) transforma p variables correlacionadas en combinaciones lineales ortogonales. El primer componente, PC1, maximiza la varianza proyectada; el segundo maximiza la varianza restante sujeto a ser perpendicular al primero, y así sucesivamente. Si X es la matriz centrada, cada componente se obtiene como z_k = Xw_k, donde w_k es un vector de cargas. Las cargas indican cómo contribuye cada variable y las puntuaciones ubican las observaciones en el nuevo espacio.

En una investigación metabólica bovina podrían medirse beta-hidroxibutirato, ácidos grasos no esterificados, glucosa, insulina, producción, consumo y condición corporal. El PCA puede resumir correlaciones en ejes tentativos de movilización energética o desempeño productivo. Esa interpretación requiere prudencia: el signo de un componente es arbitrario, una carga elevada no prueba causalidad y la varianza estadística no equivale a importancia clínica. Antes del análisis debe decidirse si se usa la matriz de covarianzas o la de correlaciones. Cuando las escalas difieren, centrar y dividir por la desviación estándar suele ser razonable; si todas las variables comparten unidad y variabilidad comparable, estandarizar puede borrar diferencias relevantes.

La retención de componentes no debe depender solo del criterio de autovalor mayor que uno. Conviene integrar gráfico de sedimentación, varianza acumulada, análisis paralelo, estabilidad de cargas e interpretabilidad. Una solución que explica 70% de la varianza puede ser suficiente para visualización, pero insuficiente para reconstruir señales fisiológicas sutiles. Los datos faltantes, atípicos y no lineales también alteran el PCA. En presencia de valores extremos se puede considerar PCA robusto; para composiciones, como proporciones microbianas, se requieren transformaciones apropiadas antes de aplicar distancias euclídeas.

Un biplot permite observar simultáneamente puntuaciones y cargas, pero su lectura depende de la escala gráfica. Animales cercanos poseen perfiles aproximados en las variables representadas; flechas con ángulo pequeño sugieren correlación positiva, opuestas correlación negativa y cercanas a 90 grados poca correlación lineal. El biplot no reemplaza la matriz de correlación ni un modelo confirmatorio.

8.3. Agrupamiento jerárquico

El agrupamiento jerárquico aglomerativo comienza con cada observación como grupo separado y fusiona, paso a paso, los grupos más próximos. El dendrograma conserva la secuencia de fusiones y permite explorar soluciones con diferentes números de conglomerados sin reajustar el modelo. La distancia define semejanza entre observaciones; el enlace define semejanza entre grupos. Enlace completo privilegia grupos compactos, enlace simple puede formar cadenas y Ward minimiza el incremento de variación intragrupo, usualmente con distancia euclídea al cuadrado.

Las variables mixtas son comunes en veterinaria: edad, sexo, especie, estado reproductivo, puntajes ordinales y biomarcadores. Aplicar distancia euclídea después de codificar categorías como números introduce un orden ficticio. La distancia de Gower resulta más adecuada cuando se combinan variables continuas, binarias, nominales y ordinales, siempre que se definan correctamente los tipos y pesos. Para secuencias o curvas completas pueden emplearse distancias entre trayectorias, correlación o dynamic time warping, aunque cada una responde a una noción diferente de similitud.

Cortar un dendrograma en tres ramas no demuestra que existan tres subpoblaciones. Se debe examinar si el patrón persiste al cambiar el enlace, retirar variables redundantes, remuestrear

animales y repetir el estudio en otra cohorte. La correlación cofenética ayuda a valorar cuánto conserva el árbol las distancias originales. Los índices de silueta o gap orientan, pero no sustituyen el juicio sustantivo. Un grupo pequeño puede ser una categoría clínica valiosa, un lote de animales de una sola finca o un conjunto de errores de medición.

8.4. K-means y otros métodos de clustering

K-means asigna cada observación al centroide más cercano y actualiza los centroides hasta reducir la suma de cuadrados dentro de los grupos. Funciona mejor con conglomerados aproximadamente esféricos, de tamaño y dispersión semejantes, en variables continuas escaladas. Requiere fijar K y es sensible a inicializaciones y valores extremos; por ello se usan múltiples arranques y una semilla documentada. Su aparente sencillez no elimina la necesidad de justificar la unidad de análisis y la métrica.

K-medoids reemplaza el centroide por una observación real y suele tolerar mejor valores extremos. Los modelos de mezcla gaussiana asignan probabilidades de pertenencia y permiten formas elípticas, aunque dependen de supuestos distribucionales. DBSCAN identifica regiones densas, admite grupos de forma irregular y marca ruido, pero sus parámetros epsilon y mínimo de puntos son difíciles de transferir entre escalas o densidades. El clustering espectral puede separar estructuras no convexas; los mapas autoorganizados y autoencoders ofrecen representaciones flexibles, a costa de mayor complejidad y menor transparencia.

Tabla 8.2.

Selección orientativa de métodos de agrupamiento

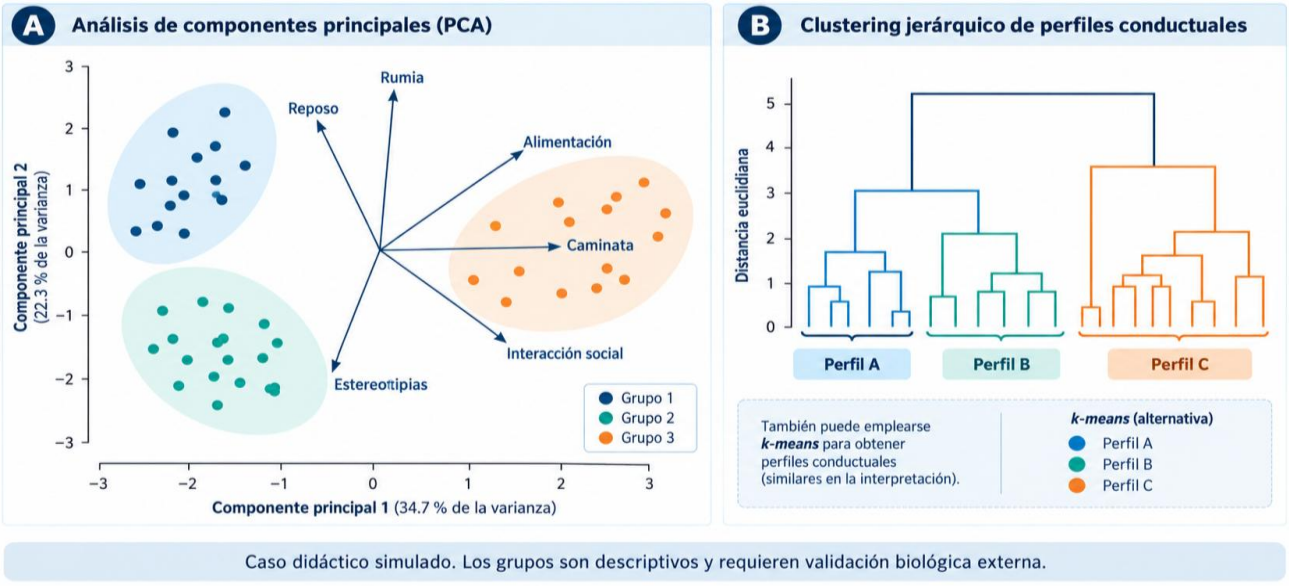
Método	Ventaja	Limitación principal	Ejemplo veterinario
Jerárquico	Dendrograma y exploración de varios cortes	Sensible a distancia y enlace	Perfiles metabólicos de vacas en transición
K-means	Rápido e interpretable mediante centroides	Supone grupos compactos; exige K	Patrones diarios de actividad resumidos
K-medoids	Mayor robustez a extremos	Costo computacional superior	Perfiles clínicos con distribuciones asimétricas

Mezcla gaussiana	Pertenencia probabilística	Supuestos y selección del número de componentes	Fenotipos continuos de riesgo
DBSCAN	Detecta ruido y formas irregulares	Parámetros sensibles a escala y densidad	Trayectorias GPS o episodios raros

La elección de K puede apoyarse en curva del codo, silueta, gap statistic, BIC en mezclas y estabilidad. Ninguno es una verdad automática. Si K=2 separa principalmente fincas tecnificadas y no tecnificadas, el hallazgo puede ser útil para gestión, pero no debe denominarse “dos fenotipos fisiológicos”. La descripción posterior debe informar distribuciones, no solo medias, y utilizar intervalos o remuestreo para cuantificar incertidumbre.

Figura 8.1.

Exploración de perfiles conductuales mediante PCA y clustering



8.5. Reducción de dimensionalidad

Reducir dimensionalidad puede perseguir compresión, visualización, eliminación de ruido o construcción de variables para otro modelo. PCA conserva estructura lineal global; t-SNE enfatiza vecindades locales y UMAP construye una representación basada en un grafo de proximidades. Estos métodos no lineales son útiles para inspeccionar espectros, señales, imágenes o perfiles

ómicos, pero una separación visual en dos dimensiones puede ser creada o exagerada por hiperparámetros, inicialización y densidad de muestreo.

En t-SNE, la distancia entre islas lejanas y el tamaño relativo de los grupos no deben interpretarse literalmente. UMAP suele preservar mejor parte de la estructura global y proyectar nuevos casos, aunque tampoco garantiza que los ejes posean significado biológico. Para una investigación reproducible deben reportarse preprocesamiento, métrica, número de vecinos, distancia mínima, semilla y versión del software. Resulta aconsejable colorear la proyección con variables externas —finca, fecha, lote, dispositivo— para descubrir confusión técnica.

La reducción supervisada y no supervisada no deben mezclarse inadvertidamente. Si se ajusta una transformación con todos los datos antes de una evaluación predictiva, se filtra información del conjunto de prueba. PCA, imputación y escalamiento deben estimarse dentro de cada pliegue de entrenamiento. En estudios longitudinales, la partición debe mantener juntos los registros del animal y respetar el orden temporal cuando el objetivo sea futuro.

8.6. Detección de anomalías

Una anomalía es una observación que se aparta de un patrón de referencia. Puede ser puntual —temperatura corporal imposible—, contextual —actividad baja durante un periodo normalmente activo— o colectiva —secuencia de cambios moderados que, en conjunto, precede a enfermedad—. La rareza estadística no equivale a patología. En datos de sensores, las causas incluyen desprendimiento, batería agotada, descalibración, desplazamiento del dispositivo, pérdida de conectividad, cambios de manejo y eventos biológicos.

Los enfoques sencillos incluyen límites fisiológicos, mediana y desviación absoluta mediana, residuos estandarizados y gráficos de control. Métodos multivariados como distancia de Mahalanobis requieren estimación estable de covarianza. Isolation Forest aísla con pocas particiones las observaciones raras; Local Outlier Factor compara densidad local; one-class SVM delimita la región normal; autoencoders usan error de reconstrucción. Los modelos complejos solo aportan si existe volumen y calidad suficientes, y si se comparan con reglas simples.

El umbral debe elegirse según el costo de falsas alertas y omisiones. En una granja, demasiadas alarmas generan fatiga y abandono del sistema; en vigilancia de una enfermedad exótica, tolerar más falsas alertas puede ser razonable. La evaluación necesita eventos confirmados y datos negativos representativos. Sensibilidad por episodio, anticipación, tasa de falsas alarmas por animal-día, valor predictivo positivo y tiempo hasta la detección son más informativos que una exactitud global. La revisión prospectiva, con el sistema funcionando antes del desenlace, es decisiva.

Tabla 8.3.

Auditoría de una alerta anómala

Pregunta	Evidencia que debe revisarse
¿Es físicamente posible?	Rango del sensor, unidades, calibración y saturación.
¿Es un fallo de dato?	Paquetes perdidos, duplicados, batería, posición del dispositivo.
¿Es contextual?	Hora, etapa productiva, temperatura ambiental, alimentación y manejo.
¿Es reproducible?	Otros sensores, observación directa, examen clínico, laboratorio.
¿Requiere acción?	Gravedad, anticipación, costo y protocolo de confirmación.

8.7. Minería de patrones en datos clínicos

La minería de patrones busca combinaciones frecuentes, secuencias y relaciones de coocurrencia. En expedientes clínicos podría identificar que determinados signos, pruebas y tratamientos aparecen juntos; en producción, que una caída de consumo suele preceder a menor producción y mayor temperatura. Las reglas de asociación se describen mediante soporte, confianza y lift. El soporte es la proporción de registros que contiene antecedente y consecuente; la confianza estima P(B|A); el lift compara esa probabilidad con la frecuencia basal de B. Un lift mayor que uno señala asociación, no causalidad.

Los datos clínicos son especialmente vulnerables al sesgo de documentación. Un diagnóstico registrado puede reflejar prácticas del hospital, disponibilidad de pruebas o preferencias del médico, no solo la enfermedad. Los códigos ausentes no significan ausencia clínica. Además, tratamientos y pruebas se ordenan después de que el profesional observa gravedad, generando confusión por indicación. Las reglas deben estratificarse por especie, servicio, edad, periodo y

centro cuando sea necesario, y proteger la privacidad mediante desidentificación, control de acceso y evaluación del riesgo de reidentificación.

En análisis secuencial, el tiempo importa. “Anorexia → ecografía → cirugía” no equivale a una cesta sin orden. Se requieren ventanas temporales justificadas, fechas fiables y separación entre episodios. Los patrones con soporte muy bajo pueden ser inestables; los de soporte alto, triviales. La validación útil consiste en comprobar el patrón en otro periodo o clínica y pedir a expertos que evalúen plausibilidad, novedad y posibilidad de acción.

8.8. Fundamentos de series temporales

Una serie temporal es una secuencia indexada por tiempo en la que el orden contiene información. Sus componentes habituales son nivel, tendencia, estacionalidad, ciclos, intervenciones y ruido. La autocorrelación expresa semejanza entre valores separados por un rezago. Ignorarla conduce a errores estándar demasiado pequeños y evaluaciones predictivas optimistas. Una serie de producción diaria también puede exhibir cambios de régimen por parto, enfermedad, modificación de dieta o secado.

La estacionariedad implica que propiedades como media, varianza y dependencia no cambian sustancialmente en el tiempo. Muchos modelos ARIMA trabajan con series estacionarias después de transformación o diferenciación. Los modelos de suavizamiento exponencial representan nivel, tendencia y estacionalidad; ARIMA modela dependencia autorregresiva y errores; modelos de espacio de estados permiten componentes latentes dinámicos; GAM incorporan relaciones no lineales suaves; modelos mixtos agregan efectos de animal y finca; redes recurrentes o transformers pueden aprender patrones complejos si existe un volumen considerable y una evaluación estricta.

La elección del horizonte modifica la pregunta. Pronosticar temperatura ruminal a 30 minutos no es igual que producción semanal. Deben definirse origen del pronóstico, horizonte, frecuencia y disponibilidad real de covariables. Una variable meteorológica observada no puede utilizarse para predecir siete días adelante salvo que en producción se disponga de su pronóstico. Este detalle, con frecuencia omitido, crea fuga de información.

La validación aleatoria convencional rompe el tiempo. Se prefieren evaluación con origen móvil, ventanas expansivas o deslizantes. El modelo se entrena hasta t, predice t+h y luego se desplaza el origen. Si existen varios animales, debe decidirse si la meta es predecir futuros registros de animales conocidos o generalizar a animales nuevos; son tareas distintas y requieren particiones diferentes.

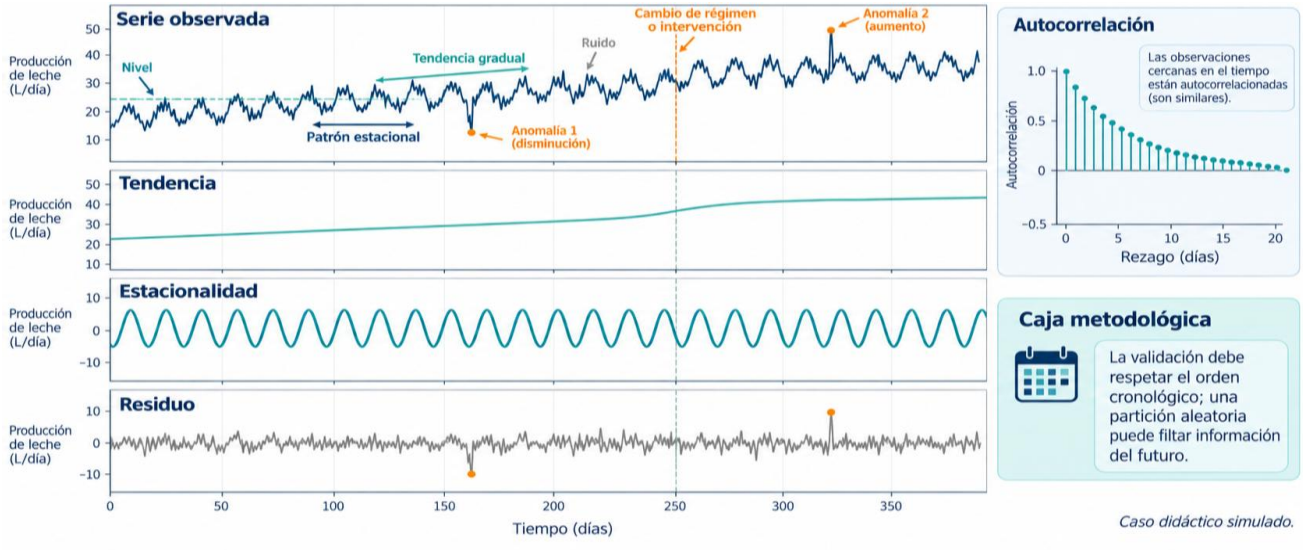
Tabla 8.4.

Fuentes de dependencia en datos temporales veterinarios

Dependencia	Ejemplo	Consecuencia metodológica
Serial	Producción de hoy correlaciona con la de ayer	Residuos autocorrelacionados y validación temporal
Intraanimal	Muchas mediciones de una misma vaca	Efecto aleatorio o partición por animal
Intrafinca	Manejo y clima compartidos	Efecto de finca o validación externa por establecimiento
Estacional	Calor, fotoperiodo o época de partos	Componentes estacionales y evaluación en varias épocas
Intervención	Cambio de dieta, tratamiento o sensor	Variable de intervención y análisis de cambio estructural

Figura 8.2.

Componentes de una serie temporal productiva



8.9. Predicción de producción y variables fisiológicas

La producción lechera es una serie determinada por días en leche, paridad, genética, gestación, alimentación, enfermedad, ambiente y manejo. Antes de usar aprendizaje automático conviene establecer líneas base: último valor observado, media estacional, curva de lactación y regresión o GAM. Un modelo complejo debe superar estas referencias en datos futuros y aportar una ventaja operativa. Bovo et al. (2021) mostraron la aplicación de random forest a producción bajo estrés térmico; Moore et al. (2023) documentaron asociaciones del calor con rendimiento y composición, y el metaanálisis de Chen et al. (2024) confirmó efectos sobre consumo, leche y eficiencia. Estos hallazgos sustentan la inclusión de exposición térmica, pero no autorizan a tratar el índice temperatura-humedad como causa única.

Las variables fisiológicas —temperatura, frecuencia cardiaca, rumia, pH ruminal— pueden tener ritmos circadianos y respuestas diferidas. La ingeniería de características debe preservar esa dinámica: rezagos, medias móviles, pendientes, amplitud diaria, tiempo por encima de umbrales y variabilidad. Las ventanas no deben cruzar el instante de predicción. El filtrado excesivo puede eliminar señales tempranas; el insuficiente confunde ruido con biología.

MAE expresa error absoluto en unidades originales y suele ser comprensible; RMSE penaliza más los errores grandes; MAPE falla cerca de cero y puede sesgar comparaciones; el error medio revela sobreestimación o subestimación. Además se requieren intervalos de predicción y cobertura. En decisiones clínicas, debe evaluarse si el error cambia conductas: un MAE de 1 kg de leche puede ser adecuado para planificación del tanque, pero insuficiente para detectar una caída individual temprana.

8.10. Datos longitudinales procedentes de sensores

Los sensores convierten la vida del animal en flujos de alta frecuencia. Un collar puede registrar aceleración triaxial, un podómetro pasos, una estación visitas y consumo, un bolo temperatura y pH, y GPS localización. La revisión de Aguilar-Lazcano et al. (2023) destaca la fusión de señales en monitoreo animal; estudios recientes han evaluado reconocimiento de conducta, cojera y metritis con datos portátiles (Neupane et al., 2024; Parsons et al., 2024; Vidal et al., 2023). La utilidad depende tanto del dispositivo y el etiquetado como del algoritmo.

El primer nivel de análisis es la señal: frecuencia de muestreo, unidades, orientación, sincronización, deriva y pérdida de paquetes. Luego se segmentan ventanas y se calculan rasgos en tiempo y frecuencia. Una etiqueta obtenida por observación cada cinco minutos no justifica clasificar cada segundo como verdad conocida. La concordancia entre observadores, el desfase de relojes y las conductas transicionales deben incorporarse al protocolo.

La jerarquía de los datos impide tratar cada ventana como independiente. Si ventanas del mismo animal aparecen en entrenamiento y prueba, el modelo puede reconocer características individuales en lugar de conducta. Una evaluación por animal estima generalización a individuos nuevos; una partición temporal por finca evalúa uso futuro; una validación externa entre dispositivos examina transportabilidad. Cabezas et al. (2022) combinaron acelerómetro y GPS para identificar conducta y eventos anómalos, un ejemplo de cómo la integración de fuentes puede mejorar contexto, sin eliminar la necesidad de confirmación.

Desde la perspectiva de bienestar, el monitoreo continuo puede detectar cambios antes que una inspección esporádica, pero también normalizar vigilancia sin protocolo de respuesta. Deben especificarse responsables, tiempos de atención, criterios de escalamiento, almacenamiento, propiedad y acceso. La calidad del sistema se mide por el ciclo completo desde la señal hasta una acción proporcional, no por una métrica aislada.

8.11. Análisis espacial y espaciotemporal

Los datos veterinarios están anclados en lugares: granjas, potreros, rutas de movilización, mercados, mataderos y hábitats. La proximidad espacial genera dependencia y puede revelar exposición compartida, conectividad o difusión. Un mapa descriptivo de casos no estima riesgo si omite la población en riesgo. Para comparar áreas se requieren denominadores, estandarización y consideración de vigilancia desigual.

La autocorrelación espacial global puede evaluarse con Moran I; los indicadores locales y estadísticos de barrido ayudan a detectar agrupamientos. La estimación de densidad kernel muestra concentración, pero depende del ancho de banda y no distingue intensidad de casos de densidad de animales. Los modelos espaciales de regresión, efectos aleatorios y procesos gaussianos

permiten ajustar covariables y dependencia residual. En vigilancia espaciotemporal se agregan fecha, duración, movilidad y cambios en cobertura diagnóstica.

Las redes de movimiento son diferentes de la distancia geográfica. Dos fincas lejanas pueden estar conectadas por comercio frecuente. Leung et al. (2023) revisaron la simulación de redes de contacto para epidemiología pecuaria, mostrando que la estructura de movimientos condiciona los resultados. Estudios de patrones espaciotemporales, como el análisis de ántrax en Vietnam de Walker et al. (2022), ilustran la integración de registros históricos y localización. La resolución espacial debe proteger confidencialidad: publicar coordenadas exactas de predios puede identificar propietarios y afectar su actividad económica.

Un análisis responsable incluye sensibilidad a escala —problema de unidad espacial modificable—, incertidumbre de geocodificación y cambios de frontera. La coincidencia espacial no prueba transmisión directa. Para sostener mecanismos se necesitan datos temporales, contactos, genómica, movimientos o evidencia ambiental.

8.12. Limitaciones de la interpretación de grupos algorítmicos

Los conglomerados son resultados condicionados por una representación. Cambiar la escala, las variables o la población puede modificar su número y composición. Denominarlos “fenotipos” exige evidencia adicional de estabilidad, reproducibilidad y asociación externa. Etiquetas como “alto riesgo” o “bajo bienestar” no deben asignarse únicamente porque el centroide posea determinados valores; primero se requiere una definición clínica y validación independiente.

La validación interna examina compactación y separación; la estabilidad pregunta si reaparecen grupos semejantes bajo remuestreo; la validación externa contrasta con variables no usadas; la utilidad evalúa si el agrupamiento mejora decisiones. También se debe investigar equidad: un grupo minoritario puede corresponder a una especie, raza, región o sistema productivo subrepresentado. El algoritmo podría tratar la diferencia como anomalía y perpetuar un estándar basado en la mayoría.

La visualización es una herramienta de comunicación con riesgos. Una proyección UMAP muy separada puede inducir a pensar en clases naturales aun cuando el espacio original sea continuo.

Conviene mostrar sensibilidad a parámetros, distribución de distancias, incertidumbre de pertenencia y perfiles individuales. Cuando no hay estabilidad, el resultado negativo también es científicamente valioso: indica que los datos no sostienen una taxonomía discreta.

Tabla 8.5.

Lista de comprobación para interpretar conglomerados

Dimensión	Pregunta crítica
Unidad analítica	¿Se agruparon animales, episodios, ventanas o fincas?
Representación	¿Qué variables, transformaciones y pesos definieron la similitud?
Estructura	¿El método admite la forma, densidad y tamaño observados?
Estabilidad	¿La solución persiste con remuestreo y especificaciones alternativas?
Validez externa	¿Se relaciona con evidencia clínica no utilizada al construirla?
Generalización	¿Se reproduce en otro periodo, finca, población o dispositivo?
Consecuencia	¿Qué decisión cambiaría y cuáles son los riesgos de etiquetar?

8.13. Caso didáctico: identificación de patrones conductuales en animales

Caso didáctico simulado. Una unidad de producción registra 60 vacas Holstein durante 28 días con collares de acelerometría triaxial y observación directa estratificada. La pregunta es si existen perfiles diarios reproducibles de actividad y si alguno se asocia con alteraciones clínicas posteriores. La unidad de agrupamiento será la vaca-día; para evitar que los días de una misma vaca dominen la evidencia, la estabilidad se evaluará remuestreando animales completos.

Variables derivadas por día: minutos de rumia, alimentación, decúbito y locomoción; número de transiciones posturales; amplitud circadiana; entropía de actividad; variación respecto de la mediana individual de los siete días previos. Se excluyen días con más de 20% de paquetes perdidos. Los valores imposibles se corrigen solo cuando existe evidencia del fallo; los restantes se marcan. Las variables se estandarizan con parámetros calculados dentro de cada muestra de entrenamiento.

Plan analítico. Primero se explora PCA para reconocer redundancia. Después se comparan Ward con distancia euclídea, k-means para K=2 a 6 y k-medoids. Se consideran silueta, gap y estabilidad Jaccard por remuestreo; la selección final exige además perfiles interpretables. Los diagnósticos y

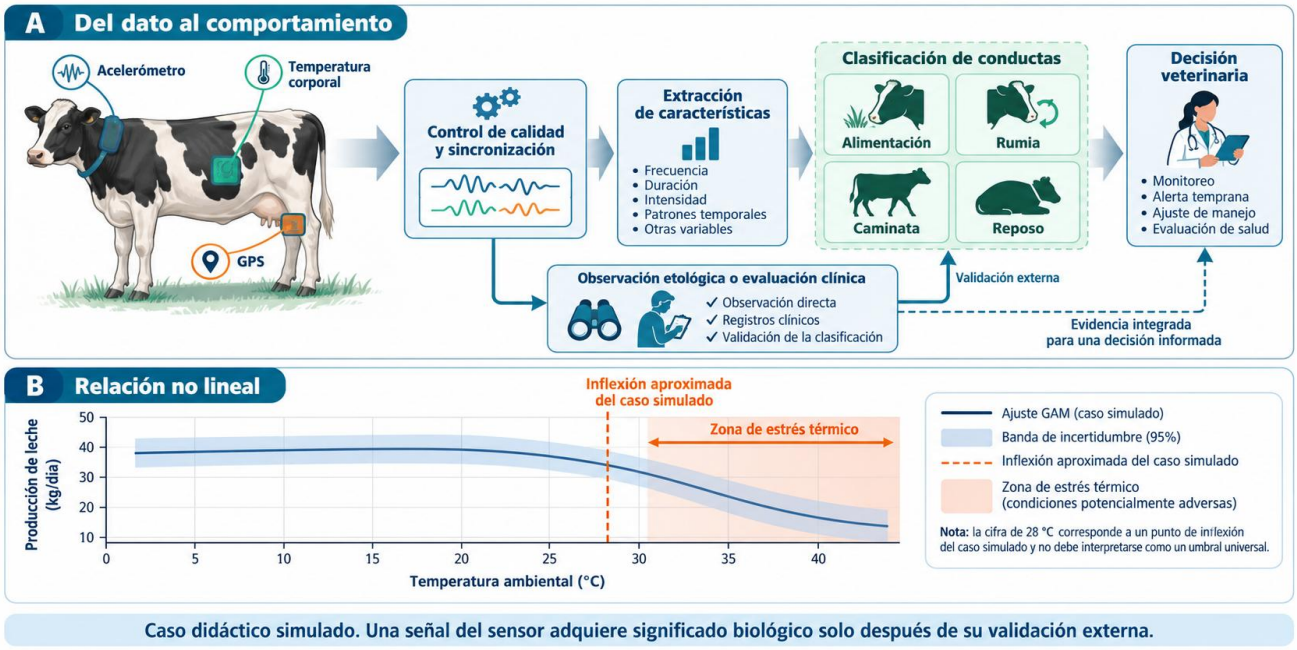
tratamientos no ingresan al agrupamiento: se reservan para validación externa. Una regresión mixta posterior puede estimar la asociación entre perfil y evento clínico en las 72 horas siguientes, con intercepto por vaca y ajuste por paridad, días en leche y temperatura ambiental. Esa asociación no se interpretará causalmente.

Resultado simulado. Una solución de tres perfiles muestra estabilidad moderada: A, actividad regular; B, mayor locomoción y menor decúbito; C, descenso de rumia, alimentación y amplitud circadiana. El perfil C presenta más eventos clínicos posteriores, pero está sobrerrepresentado en días de calor y en un corral con fallos de conectividad. Tras controlar covariables, la asociación disminuye. La interpretación correcta es que C constituye una señal exploratoria que requiere verificación prospectiva, no un diagnóstico automático.

Validación propuesta. Durante ocho semanas adicionales, el algoritmo se congela. Cada asignación a C genera una evaluación ciega del registro técnico y, si persiste, observación clínica protocolizada. Se reportan frecuencia de alerta por 100 vaca-días, sensibilidad por episodio, anticipación mediana, proporción atribuible a fallos y variación entre corrales. Un modelo solo es aceptable si aporta anticipación útil sin una carga excesiva de falsas alarmas.

Figura 8.3.

Del sensor al patrón biológico



Caso didáctico simulado. Una señal del sensor adquiere significado biológico solo después de su validación externa.

8.14. Caso didáctico: predicción de producción lechera y estrés térmico

Caso didáctico simulado. Se dispone de 24 meses de producción diaria de 420 vacas en cuatro fincas, temperatura y humedad horarias, días en leche, paridad, eventos sanitarios, dieta y calendario de enfriamiento. El objetivo operativo es predecir la producción individual del día siguiente y estimar riesgo de una reducción superior a 10% respecto de la trayectoria esperada. El objetivo predictivo no es estimar el efecto causal de instalar ventiladores.

El índice temperatura-humedad (THI) se calcula con una ecuación documentada y se derivan máximo diario, horas por encima del umbral, promedio nocturno y exposiciones rezagadas de uno a siete días. Se incorporan producción previa, pendiente de siete días y curva flexible de días en leche. La fecha del parto y la asignación de finca se verifican; los cambios de sensor se registran como intervenciones. Las covariables disponibles solo después del día pronosticado se excluyen.

Se comparan cuatro modelos: último valor observado; GAM con curva de lactación, paridad y suavizado de THI; modelo mixto con interceptos de vaca y finca; y gradient boosting. La evaluación utiliza origen móvil: entrenar meses 1–12, validar 13–15 y probar secuencialmente 16–24. Además se realiza validación dejando una finca fuera. MAE, RMSE, sesgo, cobertura del

intervalo y sensibilidad de la caída mayor a 10% se presentan por finca, paridad, estación y rango térmico.

Resultado simulado. El boosting reduce el RMSE interno frente al GAM, pero pierde ventaja en la finca excluida. El GAM muestra menor sesgo durante olas de calor y permite visualizar una respuesta no lineal con rezago. La decisión editorial y científica no debe proclamar ganador universal: para pronóstico local podría preferirse boosting; para comunicación y transporte entre fincas, el GAM resulta más estable. Ninguno identifica causalmente el beneficio del enfriamiento, pues su uso depende de condiciones y decisiones de manejo.

Aplicación. Las predicciones se agregan para planificar volumen y se genera una alerta individual solo cuando convergen caída esperada, exposición térmica sostenida y cambio de rumia. La alerta solicita inspección y no prescribe tratamiento. Se monitorea deriva: si cambia dieta, genética, sensor o patrón climático, se recalibran y reevalúan los modelos antes de actualizar el sistema.

8.15. Síntesis, autoevaluación y actividad práctica

Síntesis

- El aprendizaje no supervisado descubre representaciones condicionadas por variables, escala, distancia y algoritmo; no revela automáticamente categorías naturales.
- PCA resume variación lineal; t-SNE y UMAP facilitan visualización local, pero sus separaciones no deben interpretarse como evidencia confirmatoria.
- El agrupamiento requiere evaluación interna, estabilidad, validación externa y utilidad biológica. La solución debe probarse en otras poblaciones.
- Una anomalía puede ser enfermedad, evento de manejo o error técnico. Toda alerta necesita contexto y protocolo de confirmación.
- Las series temporales exigen respetar orden, autocorrelación, estacionalidad, rezagos y disponibilidad real de covariables.
- La validación debe corresponder al uso: futuro del mismo animal, animales nuevos, otra finca u otra región.
- Los patrones espaciales necesitan denominadores, control de vigilancia desigual, protección de ubicación y cautela causal.

Preguntas de autoevaluación

1. ¿Por qué un análisis sin variable objetivo no está libre de decisiones subjetivas?
2. ¿Cuándo conviene estandarizar variables antes de un PCA y qué información podría alterarse?
3. ¿Qué diferencia práctica existe entre enlace simple, completo y Ward?
4. ¿Por qué el valor máximo de silueta no basta para elegir el número de grupos?
5. ¿Qué precauciones exige interpretar una proyección t-SNE o UMAP?
6. Distinga anomalía puntual, contextual y colectiva con un ejemplo veterinario.
7. ¿Por qué exactitud global puede ser una métrica inadecuada para alertas sanitarias raras?
8. Explique soporte, confianza y lift en una regla de asociación.
9. ¿Qué es fuga temporal y cómo podría ocurrir al predecir producción lechera?
10. ¿Qué diferencia hay entre predecir nuevos días de vacas conocidas y generalizar a vacas nuevas?
11. ¿Por qué un mapa de casos sin denominador puede inducir conclusiones erróneas?
12. ¿Qué evidencia adicional se necesita antes de denominar fenotipo a un conglomerado?
13. ¿Cómo distinguiría una caída de actividad por enfermedad de un fallo del collar?
14. ¿Por qué un modelo de alta precisión no estima necesariamente el efecto de una intervención?
15. Proponga una validación externa para un sistema de detección temprana de cojera.

Claves orientativas

1. Porque selección de variables, escala, métrica, algoritmo, hiperparámetros y unidad analítica incorporan supuestos.
2. Cuando las magnitudes o unidades no son comparables; puede disminuir el peso de variables cuya varianza absoluta tiene significado biológico.
3. Simple usa la pareja más cercana y encadena; completo usa la más distante y compacta; Ward minimiza el aumento de variación intragrupo.
4. Porque mide estructura interna, no estabilidad, significado clínico, generalización ni utilidad.
5. No interpretar ejes, distancias globales, tamaños o islas como categorías naturales sin análisis de sensibilidad.

6. Puntual: temperatura imposible; contextual: baja rumia a una hora inusual; colectiva: descenso sostenido multivariable.
7. La clase normal domina; se necesitan sensibilidad por evento, valor predictivo positivo y falsas alarmas por animal-tiempo.
8. Soporte mide frecuencia conjunta, confianza la probabilidad condicional observada y lift la compara con la frecuencia basal.
9. Usar datos posteriores al origen del pronóstico, o ajustar preprocesamiento con todo el periodo.
10. La primera permite información histórica individual; la segunda exige partición por animal y evalúa transportabilidad.
11. Puede reflejar mayor densidad de animales, pruebas o notificación en vez de mayor riesgo.
12. Estabilidad, replicación, asociación externa, plausibilidad y utilidad, sin confundirlas con causalidad.
13. Auditar pérdidas, batería y posición; contrastar otros sensores, observación y examen clínico.
14. La predicción explota asociaciones; una intervención requiere comparador y control de confusión o aleatorización.
15. Congelar el modelo y evaluarlo prospectivamente en fincas, animales y dispositivos no usados, con confirmación clínica ciega.

Actividad práctica integradora

Propósito. Analizar un conjunto simulado de vacas lecheras para: (a) resumir perfiles conductuales mediante PCA; (b) comparar agrupamiento; (c) detectar días anómalos; y (d) construir una validación temporal básica de producción. Los resultados deben interpretarse como ejercicio, no como evidencia clínica.

Datos requeridos. Una fila por vaca-día con: id_vaca, finca, fecha, paridad, dias_leche, rumia_min, alimentacion_min, decubito_min, pasos, transiciones, temperatura, humedad, leche_kg y evento_clinico_72h. Debe crearse THI y rezagos usando únicamente datos pasados.

1. Auditar duplicados, cobertura por sensor y valores físicamente imposibles; documentar cada exclusión.
2. Estandarizar variables conductuales y ejecutar PCA. Interpretar cargas, varianza y biplot.
3. Comparar k-means K=2–5 con Ward. Calcular silueta y examinar estabilidad remuestreando vacas.
4. Describir los perfiles sin usar evento_clinico_72h; reservarlo para validación externa.
5. Detectar anomalías con mediana individual y MAD, y contrastarlas con pérdida de paquetes.
6. Crear una partición cronológica 70/30 para predecir leche_kg. Comparar último valor y modelo con rezagos, días en leche y THI.
7. Presentar MAE y sesgo por finca; explicar si el modelo generaliza o solo interpola condiciones conocidas.
8. Redactar una sección de limitaciones y un protocolo de confirmación de alertas.

Código orientativo en R

```
library(tidyverse)
library(lubridate)
library(cluster)
library(factoextra)
library(rsample)
library(yardstick)

datos <- read_csv('vacas_dia_simulado.csv') |>
  mutate(fecha = ymd(fecha),
         THI = (1.8 * temperatura + 32) -
               (0.55 - 0.0055 * humedad) * (1.8 * temperatura - 26)) |>
  arrange(id_vaca, fecha) |>
  group_by(id_vaca) |>
  mutate(leche_lag1 = lag(leche_kg), THI_lag1 = lag(THI)) |>
  ungroup()

vars <- datos |>
  select(rumia_min, alimentacion_min, decubito_min, pasos, transiciones) |>
  drop_na() |> scale()
pca <- prcomp(vars, center = TRUE, scale. = FALSE)
summary(pca); pca$rotation

set.seed(2026)
km <- kmeans(vars, centers = 3, nstart = 50)
sil <- silhouette(km$cluster, dist(vars))
mean(sil[, 'sil_width'])

# Partición temporal: el corte se define por fecha, nunca al azar.
corte <- quantile(datos$fecha, 0.70, na.rm = TRUE)
train <- filter(datos, fecha <= corte)
test  <- filter(datos, fecha > corte)

# Línea base: último valor observado.
pred_base <- test |> filter(!is.na(leche_lag1)) |>
  transmute(id_vaca, fecha, truth = leche_kg, estimate = leche_lag1)
mae(pred_base, truth = truth, estimate = estimate)
rmse(pred_base, truth = truth, estimate = estimate)
```

Producto esperado. Informe de 1.500 a 2.000 palabras con pregunta, unidad analítica, diagrama del flujo, justificación del preprocesamiento, dos visualizaciones, tabla de estabilidad, evaluación temporal, interpretación veterinaria y limitaciones. El archivo de código debe fijar semilla, registrar versiones y separar claramente exploración de confirmación.

Criterios de evaluación: coherencia entre pregunta y método (25%); prevención de fuga y pseudorreplicación (25%); interpretación biológica prudente (20%); reproducibilidad (15%); comunicación de incertidumbre y limitaciones (15%).

CAPÍTULO 9

DEEP LEARNING, VISIÓN COMPUTACIONAL Y DIAGNÓSTICO POR IMÁGENES

Introducción

Una imagen diagnóstica no es una fotografía neutra del paciente. Es el resultado de una cadena que comienza con una indicación clínica, continúa con posicionamiento, preparación, adquisición y procesamiento, y termina en una interpretación situada en la historia del animal. El aprendizaje profundo puede extraer regularidades de millones de píxeles, pero también puede aprender marcas laterales, bordes de colimación, protocolos de un hospital o decisiones previas del operador. Por ello, un modelo con una métrica elevada no equivale automáticamente a una prueba diagnóstica fiable.

La visión computacional veterinaria comprende tareas distintas: control de calidad, reconocimiento de proyecciones, clasificación de estudios, localización de lesiones, segmentación de estructuras y cuantificación. Algunas aplicaciones buscan ayudar al radiólogo o patólogo; otras pretenden priorizar estudios o estandarizar mediciones. En cada caso cambian la unidad de análisis, la referencia diagnóstica, la métrica apropiada y las consecuencias de un error. La pregunta científica debe preceder a la elección de la arquitectura.

La evidencia veterinaria ha crecido en radiología, oftalmología y patología digital, aunque muchos trabajos siguen siendo retrospectivos, unicéntricos y construidos con muestras seleccionadas. Estudios recientes demuestran factibilidad en evaluación automática de calidad radiográfica, detección de enfermedad ocular, clasificación histopatológica y segmentación de heridas, pero también exponen la necesidad de datos multicéntricos y validación externa (Banzato et al., 2023; Buschi et al., 2023; Fragoso-Garcia et al., 2023; Kim et al., 2022). El objetivo de este capítulo es enseñar a diseñar y juzgar estos estudios, no solo a describir redes neuronales.

Objetivos de aprendizaje

1. Explicar el funcionamiento esencial de una red neuronal y distinguir arquitectura, entrenamiento e inferencia.
2. Diferenciar clasificación, detección, segmentación y cuantificación en problemas veterinarios de imagen.
3. Relacionar las propiedades de radiografía, ecografía, tomografía, resonancia, histopatología y citología con decisiones de modelado.

4. Diseñar conjuntos de imágenes sin fuga de información entre animales, estudios, centros o cortes de una misma adquisición.
5. Planificar anotación, adjudicación y evaluación de concordancia entre observadores.
6. Seleccionar métricas compatibles con la tarea, la prevalencia y el costo clínico de los errores.
7. Evaluar aumento de datos, transferencia de aprendizaje, calibración, interpretabilidad y validación externa.
8. Reconocer sesgo de espectro, atajos algorítmicos y fallos de generalización entre equipos y centros.
9. Formular protocolos reproducibles para clasificación radiográfica y segmentación de lesiones cutáneas.

9.1. Fundamentos de las redes neuronales

Una red neuronal artificial es una función parametrizada compuesta por capas. Cada unidad recibe valores de entrada, calcula una combinación ponderada, añade un término independiente y aplica una función no lineal. En forma simplificada, una neurona produce $h = g(w^T x + b)$, donde x representa entradas, w los pesos, b el sesgo y g una función de activación. Al encadenar muchas transformaciones, la red puede representar relaciones no lineales que serían difíciles de especificar manualmente.

En una radiografía torácica, los valores de píxel constituyen la entrada inmediata, pero el objeto científico es un estudio perteneciente a un animal. La red no “ve” cardiomegalia como concepto clínico; ajusta filtros que responden a contornos, contrastes y configuraciones espaciales asociadas con las etiquetas disponibles. Si las radiografías positivas proceden sobre todo de un equipo y las negativas de otro, el algoritmo puede usar diferencias de procesamiento como sustituto de la enfermedad.

Las capas iniciales suelen representar bordes, gradientes y texturas; capas posteriores integran patrones más amplios. Esta jerarquía es útil, pero no garantiza que las características aprendidas sean anatómicas. El rendimiento debe probarse mediante particiones por paciente y análisis de errores. Imágenes repetidas, proyecciones del mismo estudio o cortes vecinos de una tomografía no pueden repartirse entre entrenamiento y prueba como observaciones independientes.

La salida depende de la tarea. En clasificación binaria se obtiene una probabilidad estimada; en clasificación multietiqueta, una probabilidad por hallazgo; en segmentación, una probabilidad por píxel; en detección, coordenadas, clase y confianza para cada objeto. Estas salidas no tienen el mismo significado que una probabilidad clínica posterior a la prueba. Para acercarse a ella se requiere calibración y evaluación en una población con prevalencia y espectro semejantes al uso previsto.

9.2. Arquitectura y entrenamiento de modelos profundos

La arquitectura define cómo circula la información. Profundidad, número de filtros, conexiones residuales, mecanismos de atención y resolución de entrada condicionan capacidad, memoria y susceptibilidad al sobreajuste. Redes residuales facilitan el entrenamiento de modelos profundos; EfficientNet equilibra profundidad, anchura y resolución; los vision transformers dividen la imagen en parches y modelan relaciones mediante atención (Dosovitskiy et al., 2021). La novedad arquitectónica rara vez compensa una cohorte mal construida.

Entrenar significa minimizar una función de pérdida. Para clasificación se usa con frecuencia entropía cruzada; ante desbalance pueden considerarse ponderación de clases o focal loss. Para segmentación se combinan pérdidas de entropía, Dice o Tversky. El optimizador actualiza los pesos mediante gradientes calculados por retropropagación. La tasa de aprendizaje, el tamaño de lote, la regularización, el número de épocas y el criterio de parada son hiperparámetros que deben ajustarse exclusivamente con datos de entrenamiento y validación.

La división debe respetar la estructura veterinaria. Si un perro aporta tres proyecciones, todas pertenecen al mismo subconjunto. Si se pretende generalizar a clínicas nuevas, una validación aleatoria dentro de un hospital es insuficiente: conviene reservar centros completos o realizar validación externa temporal y geográfica. También se debe registrar la semilla aleatoria, versiones de bibliotecas, hardware, preprocesamiento y pesos seleccionados.

El sobreajuste aparece cuando la red reproduce peculiaridades de entrenamiento y pierde rendimiento fuera de muestra. La brecha entre pérdida de entrenamiento y validación es una señal, no una prueba completa. Early stopping, regularización L2, dropout, aumento de datos y modelos más pequeños pueden ayudar. La decisión final debe apoyarse en un conjunto de prueba

bloqueado, intervalos de confianza y análisis estratificado por raza, tamaño corporal, edad, centro, equipo y gravedad.

Tabla 9.1.

Componentes del entrenamiento profundo y su interpretación veterinaria

Componente	Pregunta metodológica	Riesgo si se maneja mal
Cohorte	¿Representa a los animales y centros donde se usará el modelo?	Sesgo de espectro y estimaciones optimistas.
Unidad de partición	¿La separación se hizo por animal, estudio o centro?	Fuga por imágenes correlacionadas.
Etiqueta	¿Qué referencia diagnóstica y qué incertidumbre se aceptaron?	El modelo aprende desacuerdo o documentación deficiente.
Pérdida	¿Penaliza de modo coherente falsos negativos, falsos positivos y bordes?	Optimización de una meta distinta del uso clínico.
Hiperparámetros	¿Se ajustaron sin consultar la prueba?	Sobreajuste al conjunto de prueba.
Umbral	¿Se eligió según prevalencia y consecuencias?	Sensibilidad o carga de alertas inadecuadas.
Validación externa	¿Incluye otros equipos, protocolos y poblaciones?	Desempeño local que no se reproduce.

9.3. Redes neuronales convolucionales

Las redes neuronales convolucionales (CNN) aplican filtros compartidos que recorren la imagen. Un filtro produce un mapa de características al responder a patrones locales; las capas sucesivas amplían el campo receptivo y combinan señales. El uso compartido de pesos reduce parámetros respecto de una red completamente conectada y aprovecha la estructura espacial. El pooling o las convoluciones con paso reducen resolución y aumentan invariancia, aunque pueden eliminar detalles pequeños.

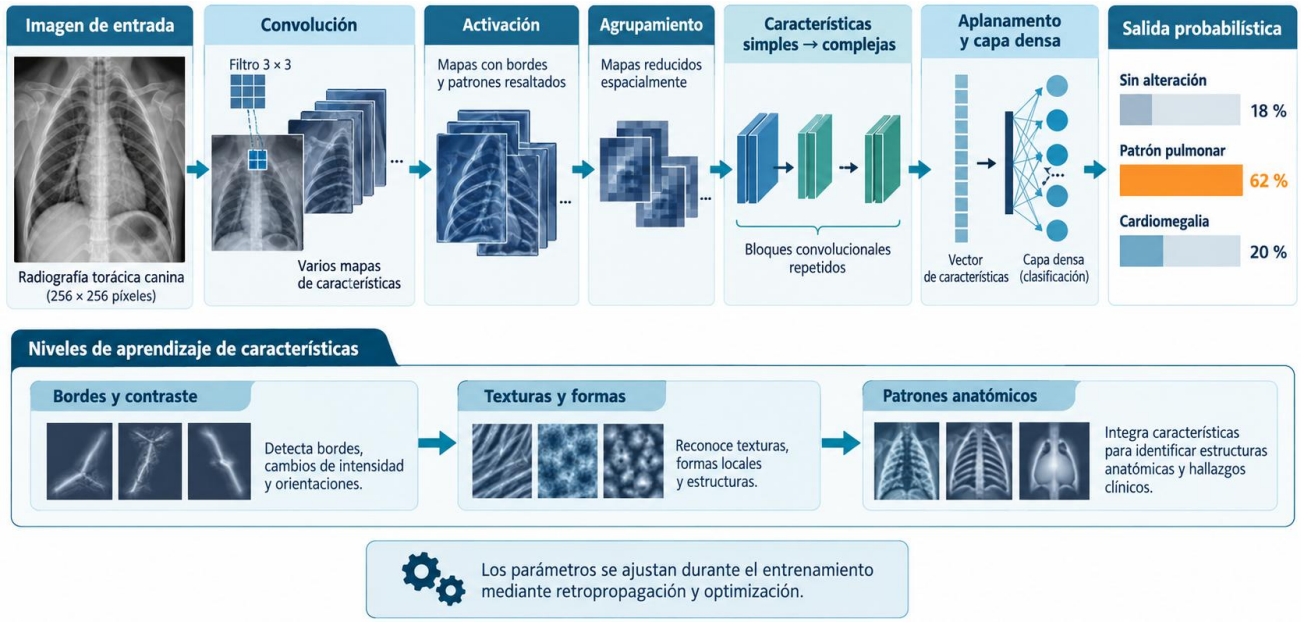
En radiología, una CNN puede aprender silueta cardiaca, patrón pulmonar y distribución de gas. En histopatología, los parches permiten analizar imágenes de portaobjetos completos, pero crean una jerarquía: miles de parches pertenecen a una lámina y varias láminas a un animal. El tamaño muestral científico no es el número de parches. Evaluar con parches del mismo caso en entrenamiento y prueba produce una ilusión de abundancia e independencia.

Las arquitecturas encoder-decoder, como U-Net, comprimen la imagen y luego recuperan resolución combinando información profunda y detalles de capas tempranas. Son frecuentes para delinear órganos, masas y heridas. Las conexiones residuales ayudan a entrenar redes profundas. Los transformers visuales ofrecen campos de interacción amplios, pero suelen necesitar gran volumen de datos o preentrenamiento; en muestras veterinarias pequeñas, una CNN ajustada con transferencia puede ser más defendible.

Un estudio comparativo debe mantener constantes particiones, resolución, aumentos y estrategia de ajuste. Comparar una CNN cuidadosamente optimizada con otro modelo usando hiperparámetros por defecto no es una comparación justa. Debe informarse variabilidad entre repeticiones y no solo el mejor resultado de múltiples intentos.

Figura 9.1.

Flujo de una red convolucional aplicada a imágenes veterinarias



9.4. Clasificación, detección y segmentación de imágenes

La clasificación asigna una etiqueta a la imagen o al estudio: presencia de edema pulmonar, tipo de tumor o grado de degeneración. La detección localiza objetos mediante cajas delimitadoras, por ejemplo cuerpos extraños o parásitos. La segmentación asigna una clase a cada píxel y permite

calcular área, volumen o forma. La segmentación semántica no separa instancias de la misma clase; la segmentación de instancias sí distingue cada objeto.

La unidad clínica rara vez coincide con una imagen aislada. Un estudio torácico incluye proyecciones ortogonales y debe integrarse con historia y exploración. Un clasificador por imagen puede emitir resultados discordantes entre proyecciones. Conviene definir una regla de agregación a nivel de estudio y evaluar tanto el rendimiento por imagen como por animal.

Las métricas dependen de la tarea. En clasificación se requieren sensibilidad, especificidad, valores predictivos, curva ROC, curva precisión-recall y calibración. En detección se usan intersección sobre unión y precisión media promedio, pero deben acompañarse de sensibilidad por lesión. En segmentación, Dice e IoU miden solapamiento; la distancia de Hausdorff o error de superficie informa sobre bordes. Un Dice alto en una lesión grande puede coexistir con un error clínicamente relevante en el margen.

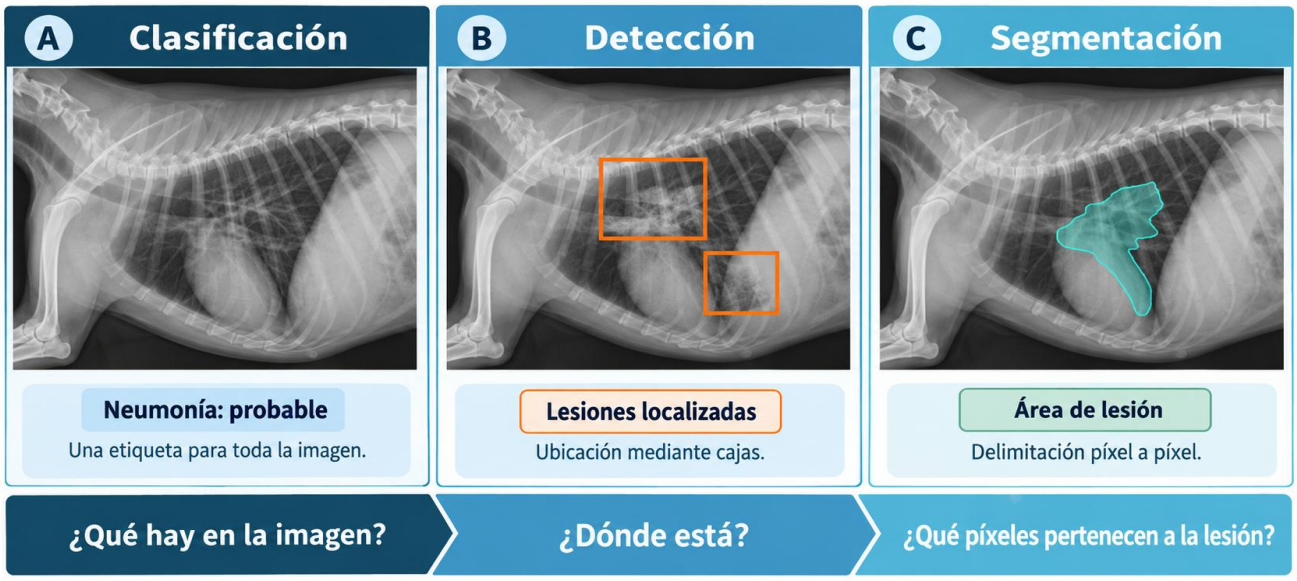
Tabla 9.2.

Tareas de visión computacional en Ciencias Veterinarias

Tarea	Salida	Ejemplo veterinario	Métricas prioritarias
Clasificación	Etiqueta o probabilidad por imagen/estudio.	Radiografía torácica normal frente a patrón pulmonar anormal.	Sensibilidad, especificidad, AUPRC, calibración.
Detección	Caja, clase y confianza por objeto.	Detección de garrapatas o cuerpos extraños.	Sensibilidad por objeto, mAP, falsos positivos por imagen.
Segmentación	Máscara por píxel.	Contorno de una herida o volumen tumoral.	Dice, IoU, distancia de superficie, error de área/volumen.
Cuantificación	Medición derivada.	Índice vertebral cardiaco o porcentaje de fibrosis.	Sesgo, límites de acuerdo, error absoluto, ICC.
Control de calidad	Categoría técnica o puntuación.	Rotación, exposición o colimación inadecuada.	Sensibilidad por error, acuerdo y tasa de repetición.

Figura 9.3.

Diferencias entre clasificación, detección y segmentación



9.5. Radiografía, ecografía, tomografía y resonancia magnética

La radiografía es una proyección bidimensional con superposición anatómica. Posicionamiento, fase respiratoria, exposición, colimación y procesamiento modifican la apariencia. Los archivos DICOM conservan profundidad de bits y metadatos; convertirlos tempranamente a JPEG puede perder rango dinámico y añadir compresión. Banzato et al. (2023) mostraron una aplicación de IA para identificar errores técnicos en radiografías torácicas caninas, mientras Shim et al. (2023) estudiaron diagnóstico profundo de enfermedades de la rodilla canina. Estos trabajos ilustran que la visión computacional puede actuar tanto sobre calidad como sobre hallazgos.

La ecografía depende intensamente del operador. Ganancia, profundidad, foco, transductor, presión y plano cambian el contenido. Una imagen fija seleccionada por un experto puede ocultar la dificultad real de adquirir el plano. Los estudios de video o cine-loop requieren partición por examen y, de ser posible, evaluación prospectiva de adquisición. Las etiquetas deben distinguir calidad insuficiente, estructura no visible y ausencia verdadera.

La tomografía computarizada produce volúmenes con espesores de corte, kernels de reconstrucción y contrastes variables. La partición por corte es incorrecta: cortes adyacentes comparten anatomía. Debe separarse por animal y estudio. Los modelos 3D aprovechan

continuidad, pero exigen memoria y cohortes mayores; estrategias 2.5D integran cortes vecinos. La estandarización de unidades Hounsfield y ventanas debe documentarse.

La resonancia magnética añade heterogeneidad de secuencias, intensidad no estandarizada, orientación y artefactos. La normalización se realiza dentro de cada estudio, pero puede borrar información si se aplica sin criterio. En columna canina, la gradación automática de degeneración discal es una tarea posible, aunque su utilidad depende de reproducir protocolos y categorías clínicas en poblaciones externas (Niemeyer et al., 2024).

La modalidad determina qué variaciones son ruido y cuáles contienen biología. Un protocolo no debe imponer la misma estrategia de normalización a todas las imágenes. La evaluación debe incluir condiciones difíciles: obesidad, animales muy pequeños, movimiento, implantes, enfermedad coexistente y estudios técnicamente subóptimos.

Tabla 9.3.

Particularidades de las modalidades de imagen

Modalidad	Estructura del dato	Fuentes críticas de variación	Validación recomendable
Radiografía	Proyecciones 2D; uno o varios planos.	Posicionamiento, exposición, colimación, procesado y marcadores.	Por animal, equipo, centro y periodo.
Ecografía	Imagen dinámica y dependiente del operador.	Plano, presión, ganancia, transductor y selección de fotograma.	Por examen y operador; incluir video si ese será el uso.
Tomografía	Volumen 3D con cortes correlacionados.	Espesor, kernel, contraste, campo de visión y movimiento.	Separación por estudio; prueba con protocolos distintos.
Resonancia	Múltiples secuencias con intensidad relativa.	Secuencia, bobina, campo magnético, orientación y artefactos.	Estratificar por secuencia/equipo; armonización documentada.
Histología	Portaobjetos completo con gigapíxeles y parches.	Tinción, escáner, laboratorio, grosor y selección tisular.	Separación por animal y laboratorio; evaluación a nivel de lámina/caso.

9.6. Imágenes histopatológicas y citológicas

La patología digital transforma láminas en imágenes de muy alta resolución. El flujo habitual divide el portaobjetos completo en teselas, excluye fondo y analiza regiones. Esto introduce decisiones sobre magnificación, tamaño de parche, solapamiento, balance de tejidos y agregación. La referencia diagnóstica puede derivarse de consenso de patólogos, inmunohistoquímica, resultado molecular o seguimiento, según la pregunta.

En un estudio multicéntrico, Salvi et al. (2021) evaluaron clasificación profunda de tumores cutáneos caninos de células redondas. El carácter multicéntrico es metodológicamente valioso porque incorpora variación de preparación y escaneo. Fragoso-Garcia et al. (2023) desarrollaron un sistema para siete tumores cutáneos caninos en imágenes H&E de portaobjetos completos; sus resultados apoyan factibilidad, pero el diseño sigue siendo prueba de concepto y no sustituye validación clínica prospectiva.

La citología presenta células superpuestas, fondo proteináceo, sangre, artefactos y diferencias de tinción. Puede plantearse detección celular, clasificación de campos o diagnóstico de la lámina. Las etiquetas por campo no deben heredarse automáticamente del diagnóstico del caso si el campo no contiene la lesión: esa asignación introduce ruido sistemático. Para recuento celular se deben validar detección, clasificación y agregación por separado.

El modelo puede aprender el laboratorio de origen. Variación de color y escáner debe auditarse con visualizaciones y pruebas cruzadas entre dominios. La normalización de tinción puede mejorar consistencia, pero también eliminar rasgos diagnósticos o crear imágenes artificiales. Debe ajustarse dentro del conjunto de entrenamiento y someterse a análisis de sensibilidad.

9.7. Reconocimiento visual de lesiones y parásitos

Las fotografías clínicas permiten estudiar heridas, alopecia, eritema, masas, ectoparásitos y lesiones podales. A diferencia de una imagen controlada, contienen pelo, reglas, manos, etiquetas, fondos, iluminación y distancia variables. Antes de entrenar conviene construir un protocolo de captura: escala métrica, carta de color cuando sea pertinente, distancia aproximada, vistas mínimas y reglas para ocultar información identificable.

Una lesión no equivale a un diagnóstico etiológico. Eritema, alopecia o ulceración son hallazgos compartidos por enfermedades distintas. Un modelo que clasifica fotografías debe definir si predice morfología, localización o diagnóstico confirmado. Para dermatología, la historia, distribución corporal, citología y pruebas complementarias suelen ser indispensables. El sistema visual puede apoyar triaje o medición, no reemplazar el proceso clínico.

En parasitología, la escala y el contexto importan. Huevos y ooquistes pueden compartir morfología; los objetos pequeños son sensibles a enfoque y resolución. Conviene reportar desempeño por taxón, carga parasitaria, preparación y microscopio. Cuando el propósito es estimar huevos por gramo, la validación debe incluir el error del procedimiento de laboratorio, no solo del detector.

Los fondos pueden crear atajos: una regla usada solo en lesiones graves, un color de mesa asociado con una clínica o un tipo de portaobjetos ligado a un parásito. La prueba de enmascaramiento de fondos, las oclusiones controladas y la validación en otro centro ayudan a revelar estos atajos.

9.8. Aumento de datos y transferencia de aprendizaje

El aumento de datos genera variaciones plausibles durante el entrenamiento: rotación leve, recorte, traslación, cambio moderado de brillo, ruido o deformación. No aumenta el número de animales ni corrige falta de representatividad. Cada transformación debe respetar la adquisición y la anatomía. Invertir horizontalmente una radiografía puede ser aceptable para ciertas tareas, pero puede invalidar lateralidad; rotar ampliamente una proyección estandarizada puede enseñar orientaciones imposibles.

En histología son razonables rotaciones y reflejos cuando la orientación no tiene significado. En ecografía, transformaciones geométricas deben preservar la relación con el haz y las marcas. En fotografía de heridas, cambios extremos de color pueden destruir señales de perfusión o inflamación. La intensidad del aumento es un hiperparámetro y debe evaluarse en validación, preferentemente mediante ablación.

La transferencia de aprendizaje inicializa una red con pesos aprendidos en otro conjunto. El preentrenamiento en imágenes naturales aporta filtros generales; el preentrenamiento médico o

veterinario puede acercar la representación al dominio. Se puede congelar parte del extractor y entrenar la cabeza, o ajustar toda la red con tasa menor. La selección debe considerar tamaño de muestra y distancia entre dominios.

Buschi et al. (2023) combinaron transferencia desde imágenes de heridas humanas con aprendizaje activo semisupervisado para segmentar heridas en animales de compañía. El ejemplo muestra una ventaja y una advertencia: transferir conocimiento reduce anotación, pero las diferencias de piel, pelo, pigmentación y contexto impiden asumir equivalencia entre especies. La evaluación final debe hacerse en animales y condiciones de captura que representen el objetivo.

Tabla 9.4.

Juicio de plausibilidad para aumento de datos

Transformación	Uso razonable	Precaución veterinaria
Rotación pequeña	Variación de posicionamiento o fotografía.	No convertir una proyección estandarizada en orientación imposible.
Reflejo	Histología o tareas sin lateralidad.	Evitar si derecha/izquierda o marcadores tienen significado.
Cambio de brillo/contraste	Variación moderada de exposición o iluminación.	No ocultar hallazgos sutiles ni simular adquisiciones irreales.
Recorte/zoom	Robustez a encuadre y escala.	No excluir contexto necesario ni alterar tamaño aparente sin control.
Deformación elástica	Segmentación de tejidos blandos, con moderación.	Puede crear anatomía no plausible y bordes falsos.
Mixup/CutMix	Regularización en clasificación.	Puede generar combinaciones sin correlato clínico; evaluar calibración.

9.9. Anotación de imágenes y concordancia entre evaluadores

La anotación define el objetivo que aprenderá el sistema. Debe existir un manual con criterios de inclusión, unidad de anotación, categorías, manejo de incertidumbre y reglas de borde. En radiografía, un estudio puede recibir etiquetas por hallazgo y diagnóstico; en segmentación, se debe acordar si se incluye costra, exudado, pelo o sombra. Las clases “indeterminada” y “calidad insuficiente” evitan forzar certezas falsas.

La experiencia de los evaluadores debe corresponder al estándar de referencia. El consenso puede construirse mediante lectura independiente y adjudicación ciega. Si un evaluador conoce el resultado del algoritmo o la etiqueta de otro, se introduce dependencia. Las etiquetas extraídas de informes son eficientes, pero reflejan omisiones y estilo de documentación; una muestra debe verificarse directamente en las imágenes.

La concordancia no debe reducirse a porcentaje. Para categorías nominales pueden emplearse kappa de Cohen o Fleiss; para ordinales, kappa ponderado; para medidas continuas, coeficiente de correlación intraclase y análisis de Bland-Altman. Kappa depende de prevalencia y distribución marginal. Conviene presentar matriz de desacuerdos y concordancia específica por clase.

El desempeño del modelo debe interpretarse frente a la incertidumbre humana. Si el borde de una lesión tiene Dice interevaluador de 0,80, exigir 0,98 al algoritmo puede ser incoherente; pero tampoco debe considerarse correcto porque coincide con un único anotador. Un conjunto de prueba con consenso o adjudicación aporta una referencia más sólida.

Tabla 9.5.

Plan mínimo de anotación

Elemento	Decisión que debe documentarse	Ejemplo
Unidad	Imagen, proyección, estudio, lámina, parche, lesión o animal.	Todas las proyecciones torácicas del mismo perro forman un estudio.
Referencia	Criterio clínico, histológico, molecular o de seguimiento.	Edema confirmado mediante interpretación radiológica y contexto clínico.
Evaluadores	Número, formación, independencia y cegamiento.	Dos radiólogos, lectura independiente y tercer adjudicador.
Incertidumbre	Categorías y procedimiento para casos dudosos.	Normal, anormal, indeterminado y no evaluable.
Borde	Regla de inclusión y resolución de anotación.	Incluir tejido ulcerado; excluir pelo y regla métrica.
Concordancia	Estadístico, intervalo y análisis por clase.	Kappa ponderado para gravedad; Dice entre máscaras.

9.10. Calidad de la imagen y variabilidad entre equipos

La calidad puede afectar la etiqueta y convertirse en señal espuria. Radiografías de urgencia suelen tener peor posicionamiento y mayor enfermedad; una red puede asociar rotación con gravedad. Antes del modelado se necesita auditoría de adquisición: fabricante, modelo, parámetros, protocolo, software de posprocesamiento, resolución, compresión y frecuencia de artefactos.

Eliminar todas las imágenes deficientes produce una prueba limpia, pero puede sobrestimar desempeño real. Es preferible definir un conjunto de desarrollo con criterios claros y una evaluación pragmática que incluya una proporción representativa de estudios subóptimos. Los casos no evaluables deben contabilizarse: abstenerse puede ser una salida segura si se mide cobertura y rendimiento condicionado.

La armonización busca reducir variación técnica sin borrar enfermedad. Incluye reescalamiento, normalización, remuestreo de vóxel y corrección de intensidad. Todo parámetro aprendido —media, desviación, normalizador de tinción— se estima en entrenamiento. La prueba nunca debe participar en la configuración.

Un análisis de robustez puede estratificar resultados por equipo, tamaño del animal, resolución y calidad; simular compresión o ruido; y usar validación leave-one-device-out. Cuando el rendimiento cae en un equipo específico, la respuesta no es ocultar el estrato, sino investigar protocolos, recalibrar o limitar el ámbito de uso.

9.11. Interpretabilidad mediante mapas de activación

Los mapas de activación de clase, como Grad-CAM, resaltan regiones asociadas con una predicción. Son útiles para auditoría visual: permiten detectar atención en marcadores, bordes o zonas fuera del paciente. No constituyen una explicación causal ni demuestran que la red use la misma evidencia que un especialista. Dos modelos con la misma predicción pueden producir mapas distintos.

La resolución del mapa suele ser menor que la imagen y la interpolación crea áreas difusas. El resultado cambia con la capa, la variante del método y la normalización. Por ello, un estudio debe

especificar implementación y evitar seleccionar solo ejemplos atractivos. Se recomienda evaluar series consecutivas, casos correctos e incorrectos y controles negativos.

La interpretación puede cuantificarse cuando existen regiones anotadas: proporción de activación dentro de la lesión, prueba de borrado, inserción progresiva o sensibilidad a oclusión. Sin embargo, una red puede usar contexto válido fuera de la lesión, y una superposición aparente no garantiza razonamiento correcto. Ghassemi et al. (2021) advirtieron que explicaciones pos hoc no sustituyen evidencia de seguridad, rendimiento y utilidad.

En veterinaria, los mapas son especialmente útiles para revelar marcas de lateralidad, collares, reglas, fondos y zonas de colimación. Deben incorporarse al control de calidad del desarrollo, no presentarse como certificación de confianza clínica.

9.12. Sesgo de espectro y generalización entre centros

El sesgo de espectro surge cuando los casos evaluados no cubren la variedad de presentación clínica. Comparar animales sanos muy claros con enfermos graves produce una tarea artificialmente fácil. En práctica, el modelo deberá distinguir enfermedad leve, comorbilidades, variantes anatómicas y estudios ambiguos. La inclusión debe basarse en la población objetivo, no en la disponibilidad de ejemplos perfectos.

La generalización entre centros falla por cambios de prevalencia, población, equipos, protocolos y decisiones clínicas. Incluso una red que conserva discriminación puede perder calibración. La validación externa debe estimar intervalos, curvas de calibración y utilidad en umbrales clínicos. Si se ajusta el modelo al centro externo, ese centro deja de ser una validación puramente externa y se convierte en recalibración o actualización.

Los subgrupos veterinarios incluyen especie, raza, conformación, tamaño, edad, sexo, estado reproductivo y contexto de referencia. Una raza puede estar asociada con un centro o una enfermedad. Los análisis deben estar preespecificados y acompañados de tamaños por estrato; comparar métricas inestables en grupos mínimos puede generar conclusiones falsas.

Las guías CLAIM 2024 y TRIPOD+AI promueven descripción transparente de datos, particiones, referencia, preprocesamiento, evaluación y aplicabilidad (Collins et al., 2024; Tejani et al., 2024). Para herramientas diagnósticas, STARD-AI amplía el énfasis en diseño de exactitud diagnóstica (Sounderajah et al., 2025). Aunque se originan en salud humana, sus principios son transferibles a investigación veterinaria si se adaptan a especies, propietarios, centros y flujos clínicos.

Figura 9.3.

Desarrollo y validación de un modelo de imágenes sin fuga de información

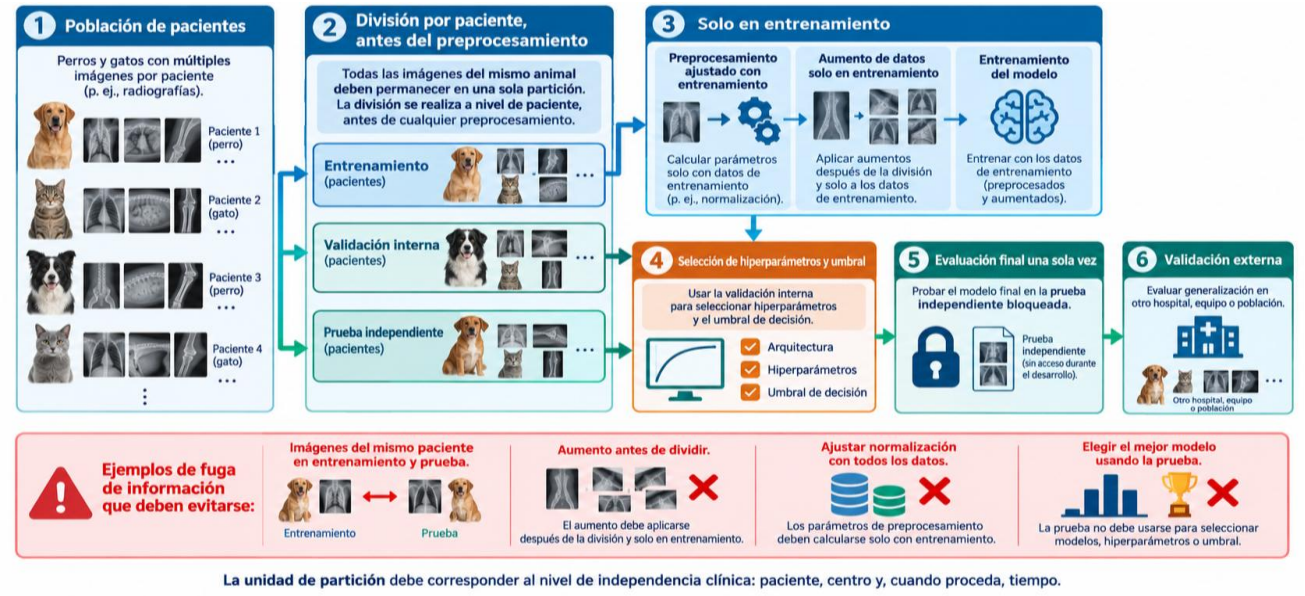


Tabla 9.6.

Amenazas a la generalización y pruebas de resistencia

Amenaza	Manifestación	Prueba o mitigación
Sesgo de espectro	Casos graves frente a controles sanos.	Incluir diferenciales, grados leves y casos indeterminados.
Fuga por paciente	Imágenes del mismo animal en entrenamiento y prueba.	Agrupar partición por animal y auditar duplicados.
Atajo por centro	Marcadores o procesamiento predicen la etiqueta.	Validación leave-one-center-out y enmascaramiento.
Cambio de prevalencia	Valores predictivos y calibración se deterioran.	Recalibración local y análisis de utilidad.

Cambio de equipo	Variación de contraste, resolución o reconstrucción.	Estratificar por dispositivo y realizar prueba externa.
Subgrupo escaso	Rendimiento desconocido en razas o tamaños infrecuentes.	Muestreo deliberado, intervalos y límites de uso.
Deriva temporal	Cambian protocolos, población o software.	Monitoreo posterior, conjuntos temporales y revalidación.

9.13. Caso didáctico: clasificación de radiografías torácicas caninas

Caso didáctico simulado. Los tamaños, prevalencias y resultados de esta sección son ficticios y se presentan exclusivamente para enseñar el diseño.

Problema científico. Un hospital veterinario desea evaluar si un modelo puede identificar estudios torácicos con patrón pulmonar alveolar que requieren revisión prioritaria. El propósito no es emitir diagnóstico etiológico ni reemplazar al radiólogo. La pregunta es: ¿un modelo entrenado con proyecciones torácicas caninas discrimina estudios con patrón alveolar y mantiene calibración en un hospital externo?

Hipótesis. El modelo alcanzará sensibilidad de al menos 0,90 en el umbral de triaje, con una especificidad compatible con la capacidad de revisión, y conservará rendimiento dentro de límites predefinidos en el centro externo. Población: perros sometidos a radiografía torácica por cualquier indicación durante tres años. Unidad de análisis: estudio por animal y fecha. Se excluyen estudios sin identificación, pero los técnicamente imperfectos se mantienen y se clasifican por calidad.

Referencia. Dos radiólogos veterinarios evalúan independientemente el estudio completo y el contexto clínico mínimo. Las categorías son patrón alveolar presente, ausente, indeterminado y no evaluable. Un tercer especialista adjudica desacuerdos. Se registra distribución y se calcula kappa; los indeterminados se analizan por separado y no se fuerzan a una clase.

Partición. Dos centros aportan desarrollo; el tercero se reserva como validación externa. Dentro del desarrollo, la división se realiza por animal y cronología. Todas las proyecciones y estudios cercanos del mismo episodio permanecen juntas. Antes de entrenar se buscan duplicados perceptuales, imágenes exportadas más de una vez y marcas que revelen el centro.

Modelo. Se usa una CNN preentrenada con entrada de alta resolución. Cada proyección genera una representación y un módulo agrega las vistas disponibles. El conjunto de entrenamiento recibe aumentos moderados compatibles con radiografía; la validación selecciona hiperparámetros y umbral. Se compara con regresión logística basada en mediciones y con una CNN más pequeña, manteniendo la misma partición.

Evaluación. Se reportan sensibilidad, especificidad, AUPRC, AUROC, valores predictivos con la prevalencia observada, curva de calibración, Brier score y matriz de confusión. Para triaje se calcula cuántos estudios serían priorizados por cada 100 y cuántos positivos se omitirían. Se estiman intervalos mediante bootstrap agrupado por animal. El rendimiento se estratifica por centro, peso, calidad, proyección y gravedad.

Resultados simulados. En la prueba interna, la sensibilidad podría ser 0,92 y la especificidad 0,76; en el centro externo, 0,84 y 0,72. La caída indicaría que la meta de sensibilidad externa no se alcanzó. Un mapa de activación revela que parte de los falsos positivos se concentra en tubos y artefactos. La conclusión científica no sería “modelo validado”, sino “prototipo prometedor que requiere ajuste del protocolo, ampliación multicéntrica y reevaluación prospectiva silenciosa”.

Tabla 9.7.

Protocolo resumido del caso torácico

Componente	Especificación didáctica
Uso previsto	Priorización de estudios para revisión, no diagnóstico autónomo.
Población	Serie consecutiva de perros con radiografía torácica por cualquier indicación.
Unidad	Estudio por animal y fecha; vistas agregadas.
Referencia	Dos radiólogos independientes y adjudicación.
Partición	Por animal; centro completo reservado para validación externa.
Comparadores	Regresión con variables definidas y CNN compacta.
Métrica primaria	Sensibilidad externa al umbral operativo preespecificado.
Métricas secundarias	Especificidad, AUPRC, calibración, carga de alertas y subgrupos.
Criterio de avance	Cumplir sensibilidad y límites de calibración sin degradación crítica por subgrupo.

9.14. Caso didáctico: segmentación de lesiones cutáneas

Caso didáctico simulado. La cohorte y los resultados se construyen con fines docentes; no corresponden a una investigación clínica real.

Problema científico. Se desea segmentar heridas cutáneas en perros y gatos para estimar área y cambio relativo durante el seguimiento. La pregunta es: ¿una red encoder-decoder produce mediciones suficientemente concordantes con expertos para seguimiento, usando fotografías tomadas en condiciones clínicas habituales?

Protocolo de captura. Cada imagen incluye escala métrica en el plano de la lesión, identificación codificada, distancia y ángulo aproximados, y una vista contextual. Se registra especie, región anatómica, pigmentación, presencia de pelo, exudado, costra, iluminación y dispositivo. Fotografías del mismo episodio permanecen en el mismo subconjunto y se conserva el orden temporal para evaluar seguimiento.

Anotación. Dos clínicos delinean independientemente el lecho visible con un manual que define costra, exudado y tejido no evaluable. Una tercera persona adjudica discrepancias mayores. Se calcula Dice interevaluador, distancia de superficie y diferencia de área. Las imágenes con borde oculto se etiquetan como parcialmente evaluables y no se confunden con máscaras completas.

Modelo y comparación. Se entrena una U-Net con encoder preentrenado y aumentos limitados. Se compara con segmentación semiautomática por color y con medición manual. El análisis primario se realiza en un conjunto externo capturado con dispositivos no usados en desarrollo. Dice es una métrica técnica; la métrica clínica principal es el error absoluto y porcentual del área corregida por escala.

Resultados simulados. El modelo obtiene Dice mediano de 0,87, pero el error de área aumenta en lesiones pequeñas y con pelo húmedo. En seguimientos semanales, la clasificación de reducción mayor al 30% concuerda en 82% de episodios. El análisis revela que la resolución insuficiente de la escala métrica explica parte del error. La recomendación sería usar control automático de calidad y revisión humana de bordes antes de calcular tendencia.

Interpretación. Una buena máscara no prueba etiología, profundidad ni infección. El sistema cuantifica superficie visible y debe integrarse con examen, dolor, tejido, exudado y evolución. La salida puede apoyar documentación longitudinal, siempre que se mantengan captura estandarizada, trazabilidad y posibilidad de corrección.

Tabla 9.8.

Matriz de evaluación del caso de segmentación

Nivel	Pregunta	Indicador
Píxel	¿La máscara se superpone con la referencia?	Dice, IoU y distancia de superficie.
Medición	¿El área calculada coincide con la medición experta?	Error absoluto, error porcentual y Bland-Altman.
Seguimiento	¿Detecta cambios clínicamente definidos?	Sensibilidad/especificidad para reducción o aumento.
Robustez	¿Funciona con dispositivos y condiciones nuevas?	Rendimiento por dispositivo, especie, pelo y tamaño.
Seguridad	¿Reconoce imágenes no evaluables?	Cobertura, tasa de abstención y errores después del filtro.
Flujo	¿Reduce tiempo sin empeorar decisiones?	Tiempo, correcciones humanas y concordancia final.

9.15. Síntesis, autoevaluación y actividad práctica

Síntesis

El aprendizaje profundo permite representar imágenes complejas sin especificar manualmente todas sus características, pero su validez se construye desde el diseño. La unidad correcta suele ser el animal o el estudio; la anotación define el objetivo; la modalidad condiciona el preprocesamiento; y la validación externa revela si el modelo aprendió anatomía o particularidades locales. La revisión de Xiao et al. (2025) muestra la diversidad creciente de aplicaciones veterinarias, junto con la heterogeneidad metodológica que todavía limita su traslación.

Clasificación, detección, segmentación y cuantificación responden preguntas diferentes. Sus métricas no son intercambiables y deben traducirse a consecuencias clínicas. La discriminación necesita complementarse con calibración, intervalos, análisis de subgrupos y evaluación de carga

de trabajo. Un mapa de activación ayuda a auditar, pero no convierte una asociación en razonamiento causal.

La utilidad clínica exige una secuencia: desarrollo retrospectivo, prueba externa, evaluación prospectiva silenciosa, estudio del trabajo humano con asistencia y vigilancia posterior. DECIDE-AI propone criterios útiles para la evaluación clínica temprana de sistemas de apoyo (Vasey et al., 2022). En diagnóstico por imágenes veterinario, el resultado más prudente suele ser una herramienta de apoyo con límites explícitos, no una decisión autónoma. Ninguna arquitectura compensa etiquetas ambiguas, cohortes sesgadas o fuga de información.

Autoevaluación

1. ¿Por qué una red neuronal puede alcanzar alta exactitud sin haber aprendido la lesión?

Respuesta orientativa: Porque puede explotar atajos correlacionados con la etiqueta, como equipo, marcador, fondo, colimación o calidad. La validación por centro, el enmascaramiento y el análisis de activación ayudan a detectarlos.

2. ¿Cuál debe ser la unidad de partición cuando un animal aporta varias proyecciones?

Respuesta orientativa: El animal o, según la pregunta, el episodio clínico completo. Todas las imágenes correlacionadas deben permanecer en el mismo subconjunto.

3. ¿Qué diferencia existe entre detección y segmentación?

Respuesta orientativa: La detección localiza objetos mediante cajas; la segmentación clasifica píxeles y permite delinear bordes, áreas o volúmenes.

4. ¿Por qué AUROC puede ser insuficiente en enfermedad infrecuente?

Respuesta orientativa: Puede parecer alta aunque existan muchos falsos positivos. La curva precisión-recall, valores predictivos y carga de alertas reflejan mejor el problema bajo baja prevalencia.

5. ¿Qué aporta la calibración?

Respuesta orientativa: Evalúa si probabilidades predichas corresponden a frecuencias observadas; es necesaria para interpretar riesgo y elegir umbrales.

6. ¿Por qué los parches histológicos no son réplicas biológicas?

Respuesta orientativa: Porque comparten lámina y animal; la independencia y el tamaño muestral se determinan en niveles superiores.

7. ¿Cuándo una transformación de aumento es inapropiada?

Respuesta orientativa: Cuando crea anatomía, orientación, color o adquisición sin plausibilidad o destruye información diagnóstica.

8. ¿Cómo debe manejarse el desacuerdo entre evaluadores?

Respuesta orientativa: Con lectura independiente, medición de concordancia, reglas de adjudicación y categorías de incertidumbre, no mediante una etiqueta forzada sin registro.

9. ¿Qué es el sesgo de espectro?

Respuesta orientativa: La distorsión que aparece cuando la muestra no incluye la diversidad y dificultad de casos del uso previsto.

10. ¿Por qué una validación aleatoria interna no demuestra generalización?

Respuesta orientativa: Porque comparte centro, equipos, protocolos y población; puede conservar las mismas señales espurias.

11. ¿Qué prueba es prioritaria para un modelo multicéntrico?

Respuesta orientativa: Una evaluación en centros no usados para desarrollo, con métricas, calibración y análisis de subgrupos preespecificados.

12. ¿Qué demuestra un mapa Grad-CAM sobre causalidad?

Respuesta orientativa: No demuestra causalidad. Solo muestra sensibilidad espacial aproximada de la predicción bajo una implementación particular.

13. ¿Qué debe reportarse cuando el modelo puede abstenerse?

Respuesta orientativa: Cobertura, motivos de abstención, rendimiento entre casos aceptados y distribución de los excluidos.

14. ¿Cuál es la limitación de transferir un modelo de heridas humanas a animales?

Respuesta orientativa: Existe cambio de dominio por piel, pelo, pigmentación, anatomía y contexto; se requiere ajuste y validación veterinaria.

15. ¿Qué secuencia de evaluación precede a la adopción clínica?

Respuesta orientativa: Desarrollo, validación interna, validación externa, evaluación prospectiva silenciosa, estudio con usuarios y monitoreo posterior.

Actividad práctica integradora

Título: Protocolo de validación de un sistema de visión computacional veterinaria.

Propósito. Elaborar, en equipos, un protocolo reproducible para una de estas tareas: clasificación de radiografías torácicas, detección de parásitos en microscopía, segmentación de heridas o clasificación de imágenes citológicas. No se requiere entrenar un modelo; el producto central es el diseño científico.

1. Defina el uso previsto, usuario, momento del flujo y acción posterior a la salida.
2. Formule pregunta, hipótesis y criterio mínimo de desempeño con justificación clínica.
3. Describa población, centros, periodo, criterios de inclusión y espectro de dificultad.
4. Identifique unidad de análisis y niveles de dependencia: animal, estudio, imagen, corte o parche.
5. Diseñe el estándar de referencia, número de evaluadores, cegamiento, adjudicación y medición de concordancia.

6. Proponga particiones de entrenamiento, validación y prueba que eviten fuga; reserve un dominio externo.
7. Especifique preprocesamiento y aumentos, justificando su plausibilidad por modalidad.
8. Seleccione un modelo sencillo de referencia y un modelo profundo; defina una comparación justa.
9. Elija métricas técnicas y clínicas, intervalos de confianza, umbral y análisis de calibración.
10. Planifique análisis de errores, subgrupos, mapas de activación y pruebas de atajos.
11. Redacte límites de uso, criterios de abstención y pasos para evaluación prospectiva.

Producto esperado. Protocolo de 4 a 6 páginas, diagrama del flujo de datos, diccionario de etiquetas, tabla de métricas y plan de análisis. Debe incluir una declaración explícita de que las imágenes del mismo animal no se distribuyen entre subconjuntos.

Lista de comprobación para un estudio de imágenes veterinarias

- El uso previsto y la población objetivo están definidos antes del algoritmo.
- La unidad de partición evita fuga entre imágenes relacionadas.
- El estándar de referencia, el desacuerdo y los casos indeterminados están documentados.
- Preprocesamiento y aumento son compatibles con la modalidad y se ajustan sin usar la prueba.
- Se informa distribución por clase, centro, equipo y subgrupos relevantes.
- Las métricas corresponden a la tarea e incluyen incertidumbre y calibración cuando aplica.
- Existe validación externa o se reconoce explícitamente su ausencia.
- Los mapas de activación se usan como auditoría, no como prueba causal.
- Se analizan errores, no evaluables, abstenciones y consecuencias clínicas.
- El informe sigue principios de CLAIM, TRIPOD+AI o STARD-AI adaptados al contexto veterinario.

CAPÍTULO 10

PROCESAMIENTO DEL LENGUAJE NATURAL E INTELIGENCIA ARTIFICIAL GENERATIVA

Introducción

Una fracción sustancial de la información veterinaria no reside en variables ordenadas, sino en narraciones: motivo de consulta, anamnesis, notas de evolución, informes de imagen, descripciones histopatológicas, recetas, comunicaciones con propietarios y comentarios de campo. Esos textos contienen signos, temporalidad, negaciones, incertidumbre diagnóstica y decisiones terapéuticas. El procesamiento del lenguaje natural (PLN) intenta convertir esa expresión humana en representaciones analizables sin reducirla a una simple búsqueda de palabras.

La inteligencia artificial generativa amplía el repertorio: además de clasificar o extraer información, produce texto, código, traducciones y resúmenes. Su fluidez puede ser útil, pero no equivale a verdad. Un modelo de lenguaje estima secuencias plausibles; no observa al animal, no verifica por sí solo una referencia y no asume responsabilidad profesional. La consecuencia metodológica es directa: la salida generada debe tratarse como un borrador o una predicción que requiere trazabilidad, contraste y validación humana.

En investigación veterinaria, el valor del PLN aparece cuando la tarea científica está bien delimitada. Extraer vómito, diarrea y pérdida de peso de historias clínicas felinas para construir una cohorte es distinto de pedir un diagnóstico abierto. Resumir 30 informes histopatológicos tampoco es equivalente a realizar una revisión sistemática. Cada uso exige un estándar de referencia, métricas apropiadas, análisis de error, protección de datos y una definición previa de qué decisiones podrán apoyarse en el sistema.

Objetivos de aprendizaje

1. Explicar cómo se representa computacionalmente el lenguaje científico y clínico.
2. Distinguir modelos discriminativos, generativos y sistemas con recuperación de información.
3. Diseñar tareas de extracción, clasificación y resumen sobre textos veterinarios.
4. Construir prompts científicos con contexto, restricciones, formato de salida y reglas de incertidumbre.
5. Auditar afirmaciones, referencias, código y análisis estadísticos generados.

6. Aplicar salvaguardas de confidencialidad y declarar de forma transparente el uso de IA generativa.
7. Evaluar modelos con métricas lingüísticas y, sobre todo, con criterios clínicos y epidemiológicos.

10.1. Fundamentos del procesamiento del lenguaje natural

El PLN reúne métodos lingüísticos, estadísticos y computacionales para analizar o generar lenguaje. Una tubería clásica separa segmentación, normalización, representación y modelado. La segmentación divide documentos en oraciones o unidades léxicas; la normalización armoniza mayúsculas, variantes y abreviaturas; la representación transforma texto en números; el modelo aprende una regla para una tarea. En veterinaria, cada fase enfrenta variación entre especies, servicios y profesionales.

Las tareas más frecuentes son reconocimiento de entidades (especie, signo, fármaco, dosis), detección de negación, extracción de relaciones, clasificación documental, búsqueda semántica, resumen y respuesta a preguntas. No todas necesitan aprendizaje profundo. Si se desea localizar una abreviatura estable y poco ambigua, una regla puede ser suficiente. Si “PU/PD” aparece con variantes, errores y contexto diverso, un modelo contextual puede ser más robusto, pero requerirá ejemplos anotados y validación externa.

La unidad de análisis debe definirse antes del algoritmo: token, oración, nota, consulta, animal, explotación o episodio clínico. Dividir aleatoriamente notas del mismo animal entre entrenamiento y prueba produce fuga de información. En datos longitudinales, la partición debe realizarse por animal y, si el propósito es prospectivo, también por tiempo o centro.

Tarea	Entrada veterinaria	Salida	Riesgo dominante
Extracción	Nota de consulta	Entidades y atributos	Omitir negación o temporalidad
Clasificación	Informe narrativo	Categoría o código	Desbalance y cambio de dominio
Resumen	Historia longitudinal	Síntesis clínica	Inventar u omitir hechos

Búsqueda semántica	Pregunta + corpus	Documentos relevantes	Confundir pertinencia con calidad
Generación	Instrucción y contexto	Texto o código	Plausibilidad sin verificación

10.2. Representación computacional del lenguaje científico

El texto puede representarse mediante conteos de palabras, n-gramas o TF-IDF. Estas matrices son dispersas, transparentes y competitivas en conjuntos modestos. Su limitación es que "no presenta fiebre" y "presenta fiebre" comparten casi los mismos términos; la polaridad depende del orden y del contexto. Los embeddings densos asignan vectores a palabras, oraciones o documentos, aproximando semejanzas semánticas.

La tokenización fragmenta el texto. En modelos modernos una palabra puede dividirse en subpalabras; esto reduce el problema de vocabulario desconocido, aunque términos como "hiperadrenocorticismo", nombres comerciales o abreviaturas locales pueden quedar mal representados. El investigador debe inspeccionar cómo se tokenizan términos críticos y no asumir que un modelo biomédico humano comprende automáticamente nomenclatura veterinaria.

El lenguaje clínico añade negación, incertidumbre y temporalidad. "Se descartó parvovirosis", "compatible con", "antecedente de convulsiones" y "convulsionó hoy" no son equivalentes. Una extracción válida debe asociar entidad, presencia, sujeto y tiempo. También debe diferenciar al animal de otros sujetos: "el propietario refiere diabetes" no describe al paciente.

Representación	**Ventaja**	**Limitación**	**Uso razonable**
Bolsa de palabras / TF-IDF	Simple e interpretable	Pierde orden y contexto	Línea base de clasificación
Embeddings estáticos	Captan similitud global	Una representación por palabra	Exploración semántica
Embeddings contextuales	Cambian según la oración	Mayor costo y opacidad	NER, clasificación, búsqueda
Representación estructurada	Facilita análisis y auditoría	Requiere ontología y mapeo	Cohortes, vigilancia, interoperabilidad

10.3. Transformers y modelos de lenguaje

El transformer sustituyó gran parte del procesamiento recurrente por mecanismos de atención que ponderan relaciones entre elementos de una secuencia y permiten entrenamiento paralelo (Vaswani et al., 2017). BERT introdujo preentrenamiento bidireccional útil para clasificación y extracción (Devlin et al., 2019). Los modelos autoregresivos, en cambio, aprenden a predecir el siguiente token y pueden generar texto extenso.

Un modelo fundacional se preentrena en grandes corpus y luego puede adaptarse mediante ajuste fino, aprendizaje con pocos ejemplos o instrucciones. La adaptación de dominio importa: PetBERT fue entrenado con registros veterinarios de primera opinión y mostró la utilidad de un lenguaje especializado para codificación sindrómica (Farrell et al., 2023). Boguslav et al. (2026) ajustaron trece modelos sobre 246 473 visitas codificadas y 7 739 códigos SNOMED-CT; el hallazgo apoya la adaptación, pero no elimina la necesidad de validación en otros hospitales, idiomas o especies.

La atención no es una explicación causal. Que una palabra reciba peso no demuestra que sea el fundamento clínico correcto. Asimismo, un modelo grande no posee una base de datos de hechos consultable de forma perfecta: comprime regularidades estadísticas y puede mezclar contenidos. La recuperación aumentada (RAG) agrega documentos recuperados al contexto; mejora trazabilidad si las fuentes son pertinentes, pero una recuperación deficiente o un documento débil trasladan el error a la respuesta.

10.4. Extracción de información de historias clínicas

La extracción convierte narraciones en campos analizables: signo, localización, duración, frecuencia, gravedad, negación, medicamento, dosis, vía y respuesta. La anotación humana debe apoyarse en un manual operativo con ejemplos limítrofes. Dos veterinarios pueden discrepar no por descuido, sino porque la nota es ambigua. Por eso conviene estimar concordancia y resolver diferencias antes de considerar las etiquetas como verdad de referencia.

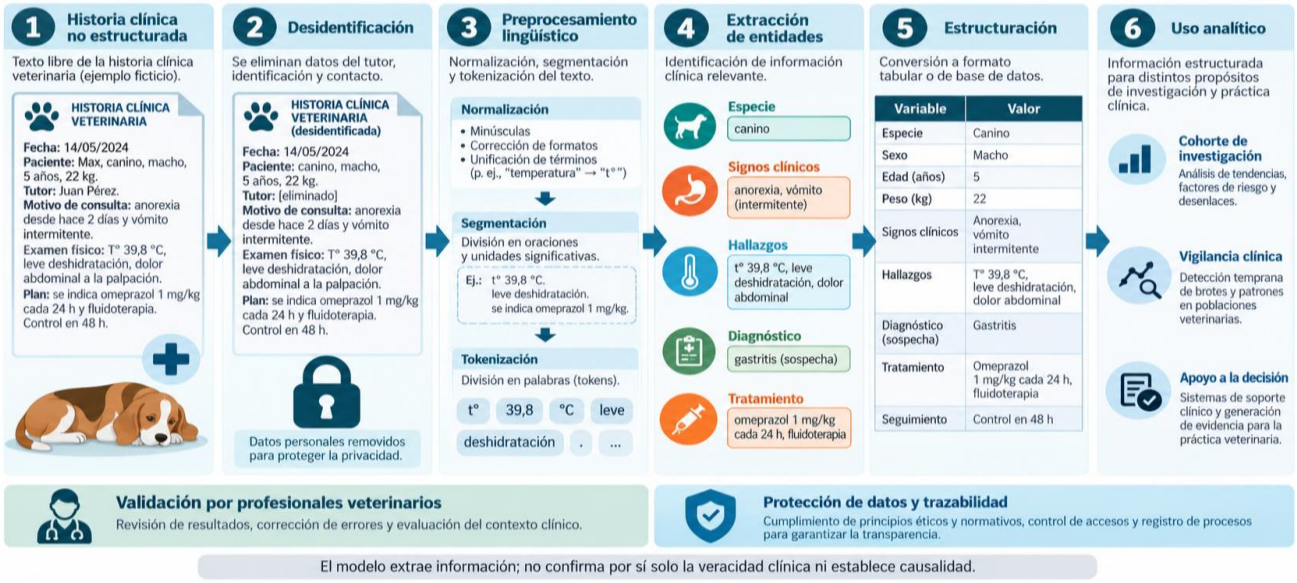
Wulcan et al. (2025) evaluaron la identificación de seis signos asociados con enteropatía crónica felina en 250 historias. El estudio informó alto rendimiento de GPT-4o frente a la opinión

mayoritaria y mostró que muchos errores coincidían con desacuerdo humano. La lección no es que el modelo sustituya la revisión, sino que la ambigüedad del texto debe incorporarse al análisis de error y al diseño del estándar de referencia.

La validación debe calcularse por entidad y por nota, informar sensibilidad, precisión, F1 y errores clínicamente relevantes. En vigilancia, perder casos tempranos puede ser más costoso que revisar falsos positivos; en farmacovigilancia, una falsa atribución de evento adverso también puede ser grave. El umbral óptimo depende de la decisión posterior.

Figura 10.1.

Procesamiento de una historia clínica narrative



10.5. Clasificación y resumen de textos veterinarios

Clasificar significa asignar una o varias etiquetas: síndrome gastrointestinal, sospecha infecciosa, prioridad de triaje o código diagnóstico. PetBERT automatizó categorías sindrómicas compatibles con ICD-11 para apoyar detección de brotes en registros de primera opinión (Farrell et al., 2023). Davies et al. (2024) muestran que la minería de texto para vigilancia exige comprender el lenguaje del corpus y entrenar detectores sobre anotaciones pertinentes.

El resumen extractivo selecciona fragmentos existentes; el abstractivo redacta una síntesis nueva. El segundo puede ser más legible, pero puede alterar dosis, fechas o relaciones causales. Para historias longitudinales se recomienda una salida estructurada con cronología, problemas activos, tratamientos, resultados y campos “no consta”. La fidelidad factual debe evaluarse oración por oración contra el registro original, no solo con métricas de similitud léxica.

Una evaluación útil combina: completitud, factualidad, relevancia, orden temporal, ausencia de contradicciones y tiempo ahorrado. ROUGE o BERTScore pueden complementar, pero no sustituyen la revisión veterinaria. Un resumen con alta similitud textual puede omitir una alergia o confundir 0,5 con 5 mg/kg.

10.6. ChatGPT y otros asistentes generativos

ChatGPT y otros asistentes conversacionales ofrecen una interfaz para formular preguntas, transformar texto, generar código y explorar alternativas. La marca o versión no define por sí sola la calidad: cambian el modelo, las herramientas disponibles, la fecha de conocimiento, la política de datos y el comportamiento de una actualización a otra. Todo protocolo debe registrar proveedor, modelo o versión declarada, fecha, configuración, prompt y archivos utilizados.

En educación veterinaria, Alonso Sousa et al. (2025) compararon nueve modelos en 250 preguntas de examen y observaron variación sustancial por modelo, dificultad y presencia de imágenes. El desempeño en preguntas cerradas no valida prescripción, diagnóstico ni razonamiento bajo incertidumbre. Un examen mide una parte del conocimiento; la práctica clínica añade examen físico, contexto del propietario, disponibilidad terapéutica y responsabilidad profesional.

Uso	Aporte posible	Condición de seguridad
Ideación	Generar preguntas o contraargumentos	Revisión del investigador
Redacción	Reorganizar o aclarar un borrador	Conservar significado y autoría humana
Código	Proponer funciones y pruebas	Ejecutar con datos simulados y validar
Clínica	Estructurar información	No usar como decisión autónoma
Bibliografía	Sugerir términos de búsqueda	Verificar cada registro en bases reales

10.7. Búsqueda orientativa y revisión de literatura

Un asistente puede ayudar a transformar una pregunta PICO/PECO en sinónimos, términos MeSH y palabras de título. También puede proponer criterios de inclusión o una matriz de extracción. No debe actuar como fuente ni como sustituto de PubMed, CAB Abstracts, Scopus, Web of Science o repositorios oficiales. Una lista generada puede incluir artículos inexistentes, DOI incorrectos o trabajos irrelevantes.

En una revisión, la búsqueda debe quedar reproducible: base, plataforma, fecha, estrategia completa, filtros, duplicados y proceso de selección. PRISMA 2020 orienta el reporte de revisiones sistemáticas (Page et al., 2021). La IA puede apoyar cribado priorizado, pero una muestra de exclusiones debe auditarse y la decisión final debe seguir criterios preespecificados.

Procedimiento mínimo: usar el modelo para ampliar vocabulario; ejecutar la estrategia en bases reales; exportar metadatos; deduplicar; verificar DOI y estado editorial; leer texto completo; vincular cada afirmación con la fuente primaria. “La IA dice que existe evidencia” nunca constituye evidencia.

10.8. Apoyo en programación y análisis estadístico

Los asistentes pueden explicar errores de R, proponer pruebas unitarias, documentar funciones o traducir una pregunta a pseudocódigo. El riesgo aparece cuando seleccionan automáticamente una prueba, inventan nombres de variables o interpretan $p < 0,05$ como relevancia clínica. El investigador debe proporcionar diseño, unidad experimental, estructura de dependencia, distribución, objetivo inferencial y formato de datos.

Antes de ejecutar código generado se revisan paquetes, versiones, rutas, transformaciones, semillas, manejo de faltantes y objetos de salida. Se ensaya con datos simulados cuyo resultado sea conocido. Después se contrasta con una implementación alternativa o cálculo manual. Ningún código compensa pseudorreplicación, fuga de información o un estándar de referencia defectuoso.

Ejemplo de instrucción robusta: “Escribe R para un modelo logístico mixto de mastitis subclínica, con vaca como intercepto aleatorio y hato como bloque; explica supuestos, no inventes variables,

incluye diagnóstico de convergencia y devuelve únicamente código comentado y lista de verificaciones". La precisión del encargo no garantiza corrección, pero hace auditables los supuestos.

10.9. Construcción de prompts científicos

Un prompt científico es una especificación de tarea. Debe declarar propósito, rol funcional, contexto, datos permitidos, unidad de análisis, criterios, formato, límites y conducta ante incertidumbre. Pedir "analiza estos datos" oculta decisiones. Pedir "identifica signos gastrointestinales explícitos, distingue presentes, negados e históricos, no infieras diagnóstico y devuelve JSON según este esquema" reduce la ambigüedad.

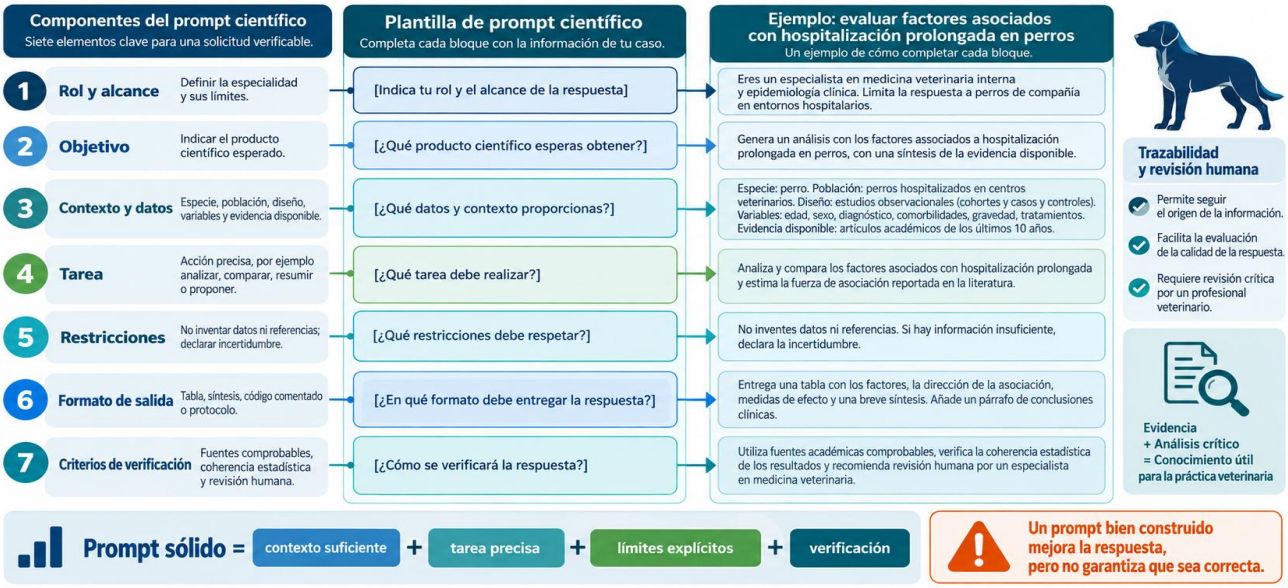
Componente	Pregunta de diseño	Ejemplo veterinario
Objetivo	¿Qué producto se necesita?	Extraer signos, no diagnosticar
Contexto	¿Especie, servicio, idioma?	Gatos de medicina interna
Definiciones	¿Qué cuenta como positivo?	Mención actual atribuida al paciente
Restricciones	¿Qué no debe inferirse?	No deducir diarrea por fármaco
Salida	¿Qué esquema exacto?	signo, estado, evidencia, fecha
Incertidumbre	¿Cómo reportar ambigüedad?	"incierto" + fragmento textual
Verificación	¿Qué debe comprobarse?	Citar oración fuente; no inventar

Prompt modelo: "Actúa como anotador de historias clínicas veterinarias. Para cada nota identifica únicamente vómito, diarrea, hiporexia y pérdida de peso. Clasifica estado como presente, negado, histórico o incierto; copia un fragmento breve como evidencia; no completes información ausente. Devuelve CSV con id_nota, signo, estado, temporalidad y evidencia. Si hay contradicción, crea dos filas y marca revisión_humana=Sí".

La evaluación de prompts debe hacerse sobre un conjunto congelado y separado del material usado para diseñarlos. Cambiar el prompt después de ver errores del conjunto de prueba equivale a ajustar sobre la prueba. Conviene conservar versiones, ejecutar repeticiones cuando el sistema sea estocástico y documentar temperatura u otros parámetros disponibles.

Figura 10.2.

Anatomía de un prompt científico veterinario



10.10. Verificación de afirmaciones y referencias

Cada afirmación generada debe clasificarse como verificable, interpretativa o recomendación. Las verificables exigen fuente primaria o documento oficial. Para una referencia se comprueban autores, año, título, revista, volumen, páginas o número de artículo y DOI; luego se confirma que el texto respalda exactamente la proposición. La existencia del artículo no valida una afirmación distinta.

- Buscar el título exacto y el DOI en la editorial o Crossref/PubMed.
- Abrir resumen y, cuando sea necesario, texto completo.
- Comprobar población, especie, diseño y desenlace.
- Distinguir preprint, artículo aceptado, corrección o retractación.
- Registrar qué afirmación sostiene cada fuente.
- Eliminar referencias no verificadas; nunca “corregir” un DOI por intuición.

Las respuestas con citas automáticas pueden producir “lavado de autoridad”: una afirmación parece científica por estar acompañada de una referencia, aunque esta no la respalde. La auditoría

debe ser a nivel de afirmación, no de párrafo. En asuntos clínicos se priorizan guías, consensos y estudios primarios pertinentes a especie y población.

10.11. Alucinaciones, sesgos y errores metodológicos

Una alucinación es contenido no sustentado por la entrada o por hechos verificables; puede ser intrínseca, cuando contradice la fuente, o extrínseca, cuando agrega información no demostrada (Ji et al., 2023). La fluidez agrava el riesgo porque el error adopta forma convincente. Pedir seguridad numérica no crea una probabilidad calibrada.

Los sesgos reflejan corpus, anotadores, idioma y contexto de despliegue. La literatura está concentrada en inglés y en centros con registros digitalizados; razas, sistemas productivos y condiciones tropicales pueden estar subrepresentados. Bender et al. (2021) advierten sobre la opacidad de corpus extensos, costos y reproducción de sesgos. En veterinaria, el traslado acrítico desde medicina humana puede confundir fisiología, dosis, toxicidad o prevalencias.

Error	Ejemplo	Control
Referencia falsa	DOI plausible inexistente	Verificación en editorial/PubMed
Negación omitida	“No vómito” → vómito presente	Prueba específica de negación
Sujeto incorrecto	Dato del propietario atribuido al animal	Etiqueta de sujeto
Causalidad espuria	Asociación narrada como efecto	Revisión del diseño y DAG
Fuga de datos	Notas del mismo animal en train/test	Partición por animal/centro
Automatización complaciente	Aceptar salida por su tono	Doble revisión y registro de cambios

10.12. Confidencialidad de la información clínica

Aunque el animal no sea titular de derechos de privacidad en el mismo sentido que una persona, su historia puede contener nombres, teléfonos, direcciones, ubicación de fincas, datos de pago y conversaciones del propietario. En Ecuador, la Ley Orgánica de Protección de Datos Personales regula el tratamiento de datos personales y exige principios de finalidad, minimización, seguridad y responsabilidad (Asamblea Nacional del Ecuador, 2021).

Figura 10.3.

Circuito de verificación de una respuesta generada por IA



No se deben pegar historias identificables en servicios externos sin base jurídica, consentimiento cuando corresponda, contrato institucional y evaluación de seguridad. La seudonimización sustituye identificadores, pero fechas raras, ubicación y combinación de atributos pueden reidentificar. La anonimización efectiva exige valorar el riesgo residual.

- Minimizar: enviar solo campos necesarios.
- Desidentificar antes de cualquier transferencia.
- Revisar retención, entrenamiento con datos y ubicación del proveedor.
- Usar entornos institucionales aprobados o despliegue local cuando proceda.
- Controlar acceso, cifrado, registros y eliminación.
- Incluir el flujo de IA en protocolo, consentimiento y evaluación ética.

10.13. Traducción y redacción científica asistidas

La traducción automática puede mejorar acceso, pero debe preservar especie, anatomía, unidades, intensidad de evidencia y modalidad. “May”, “likely” y “confirmed” no son intercambiables. Un

revisor bilingüe debe comparar título, resumen, métodos, tablas y términos clínicos. Los nombres taxonómicos, genes y fármacos se verifican en fuentes autorizadas.

En redacción, la IA puede detectar redundancias, proponer transiciones o adaptar claridad; no debe fabricar resultados ni borrar incertidumbre. Se recomienda trabajar sobre texto del autor, pedir conservación exacta de cifras y citas, y usar control de cambios. La responsabilidad intelectual, la selección de evidencia y la interpretación permanecen en los autores.

La homogeneización estilística también es un riesgo: puede borrar matices, convertir limitaciones en frases genéricas o introducir anglicismos. Una revisión final debe recuperar la voz científica, comprobar correspondencia entre objetivos, métodos y conclusiones, y evitar afirmaciones más fuertes que los datos.

10.14. Criterios para declarar el uso de IA generativa

El ICMJE (2026) exige revelar el uso de tecnologías asistidas por IA y rechaza su autoría porque no pueden asumir responsabilidad. También señala que los autores deben revisar exactitud, integridad, originalidad y atribución. WAME formula principios semejantes (Zielinski et al., 2023). La política concreta de la revista prevalece y debe consultarse antes del envío.

Dónde declarar	**Contenido mínimo**
Métodos	Herramienta, versión, fecha, tarea, prompts o protocolo, parámetros, datos y validación
Agradecimientos/declaración	Uso editorial o lingüístico según política
Carta de presentación	Alcance del uso y responsabilidad humana
Material suplementario	Prompts completos, salidas relevantes, código y registro de auditoría

Ejemplo: “Se utilizó [herramienta y versión], consultada el [fecha], para proponer expresiones regulares y depurar código R sobre datos simulados. El código fue revisado por los autores, probado con casos conocidos y ejecutado sobre la base final. La herramienta no se empleó para generar datos, seleccionar referencias ni interpretar resultados. Los autores asumen responsabilidad total”. Si la IA intervino en análisis o extracción, la descripción debe permitir reproducir y evaluar esa intervención.

10.15. Caso didáctico: extracción de signos clínicos desde textos narrativos

Caso didáctico simulado. Un hospital desea construir una cohorte retrospectiva de gatos con sospecha de enteropatía crónica. La pregunta es: ¿con qué exactitud puede extraerse de notas narrativas la presencia, negación y temporalidad de vómito, diarrea, hiporexia y pérdida de peso? La unidad observacional es la nota; la unidad de partición es el animal.

Diseño metodológico

- Población: consultas felinas de medicina interna entre 2023 y 2025.
- Muestra: 600 animales; una o más notas por animal.
- Referencia: anotación independiente por dos médicos veterinarios y adjudicación de desacuerdos.
- Partición: 60% entrenamiento/desarrollo, 20% validación y 20% prueba, sin compartir animales; prueba externa en otra clínica.
- Salida: signo, estado, temporalidad, fragmento de evidencia y bandera de ambigüedad.
- Métricas: sensibilidad, precisión, F1 por signo, exactitud balanceada, concordancia y análisis de errores.

Notas simuladas: A) “Desde hace tres semanas vomita alimento dos veces por semana; heces formadas. Ha perdido 0,6 kg”. B) “Niega vómitos. Diarrea resuelta hace un mes; apetito normal”. C) “Propietaria diabética; el gato come menos desde ayer”. El sistema debe reconocer negación, historia resuelta y sujeto. No debe convertir “heces formadas” en diarrea presente ni atribuir diabetes humana al gato.

Nota	Signo	Estado	Temporalidad	Evidencia
A	Vómito	Presente	3 semanas	“vomita alimento…”
A	Diarrea	Negado implícito/ausente	Actual	“heces formadas”
A	Pérdida de peso	Presente	No precisada	“Ha perdido 0,6 kg”
B	Vómito	Negado	Actual	“Niega vómitos”
B	Diarrea	Histórico/resuelto	Hace un mes	“Diarrea resuelta…”

C	Hiporexia	Presente	Desde ayer	“el gato come menos…”

Implementación reproducible en R

La siguiente línea base no pretende resolver todo el lenguaje; sirve para detectar fallos antes de comparar modelos complejos.

```
library(dplyr)
library(stringr)

extraer_signos <- function(texto) {
  tibble(
    vomito = str_detect(texto, regex("\b(vomit|emesis)", ignore_case = TRUE)) &
             !str_detect(texto, regex("(niega|sin)\s+vomit", ignore_case = TRUE)),
    diarrea = str_detect(texto, regex("\bdiarrea\b", ignore_case = TRUE)) &
              !str_detect(texto, regex("(niega|sin)\s+diarrea", ignore_case = TRUE)),
    hiporexia = str_detect(texto, regex("hiporexia|come menos|apetito disminuido", ignore_case =
TRUE)),
    perdida_peso = str_detect(texto, regex("perdi[oó]\s+[0-9,.]+\s*(kg|g)|p[eé]rdida de peso",
ignore_case = TRUE))
  )
}

# Probar primero con frases positivas, negativas, históricas y ambiguas.
# Luego comparar contra la anotación veterinaria y revisar falsos positivos/negativos.
```

La expresión regular fallará ante negaciones distantes, errores ortográficos y referencias históricas. Esa debilidad es informativa: establece una línea base y justifica, si el incremento es clínicamente relevante, un modelo contextual. La comparación debe usar el mismo conjunto de prueba y reportar intervalos de confianza.

Interpretación y límites

Un F1 global alto puede ocultar baja sensibilidad en pérdida de peso, el signo menos frecuente. Deben informarse resultados por clase, prevalencia y costo del error. La validación externa comprueba transferencia entre estilos de escritura. Antes del uso operativo se requiere vigilancia de deriva, revisión humana de casos inciertos y un mecanismo para corregir etiquetas.

10.16. Caso didáctico: evaluación crítica de una respuesta generada por IA

Caso didáctico simulado. Pregunta al asistente: “En un perro con dolor posoperatorio tras ovariohisterectomía, ¿la combinación meloxicam-tramadol es superior a meloxicam solo?”.

Respuesta hipotética: “Sí, múltiples ensayos demuestran una reducción de 50% del dolor; se recomienda tramadol 5 mg/kg cada 8 horas y la combinación es segura en todos los perros”.

Auditoría de la respuesta

Elemento	Juicio crítico	Acción
Conclusión categórica	No define población, escala ni momento	Reformular pregunta y buscar ensayos
“Múltiples ensayos”	Afirmación sin referencias verificables	Solicitar y comprobar cada fuente
50%	Precisión numérica sin contexto	Localizar resultado, IC y denominador
Dosis	Consejo clínico potencialmente riesgoso	Verificar guía/fármaco, paciente y normativa
“Segura en todos”	Generalización absoluta	Revisar contraindicaciones y eventos adversos

La pregunta correcta separa eficacia analgésica, relevancia clínica y seguridad. Debe especificar especie, procedimiento, escala validada, tiempo, comparador y diseño. La búsqueda debe privilegiar ensayos y revisiones veterinarias, evaluar sesgo y heterogeneidad, y no extrapolar analgesia humana. Si no existe evidencia suficiente, la respuesta apropiada es incertidumbre, no una recomendación inventada.

Prompt de reparación: “No emitas recomendación terapéutica. Formula una pregunta PICO sobre perros sometidos a ovariohisterectomía, meloxicam más tramadol frente a meloxicam, dolor medido con escala validada a 24 h. Propón términos de búsqueda. Marca toda referencia como ‘pendiente de verificación’ y no inventes dosis ni resultados”. Después, el investigador ejecuta la búsqueda real y analiza la evidencia.

10.17. Síntesis, autoevaluación y actividad práctica

Síntesis

El PLN permite estructurar texto veterinario, pero el significado clínico depende de negación, temporalidad, sujeto y contexto. Los transformers ofrecen representaciones contextuales y los

modelos generativos producen respuestas flexibles; ninguno convierte lenguaje plausible en evidencia. El desempeño debe evaluarse sobre tareas definidas, datos independientes y errores con significado biológico o clínico.

La IA generativa puede apoyar búsqueda orientativa, programación, traducción y redacción. Su uso responsable exige minimizar datos, verificar afirmaciones y referencias, conservar trazabilidad y declarar la intervención. La regla central es sencilla: automatizar una tarea no transfiere la responsabilidad científica.

Autoevaluación

1. ¿Por qué una búsqueda por palabras puede fallar ante negación y temporalidad?
2. ¿Qué diferencia existe entre un embedding estático y uno contextual?
3. ¿Por qué las notas del mismo animal no deben dividirse entre entrenamiento y prueba?
4. ¿Cuándo sería preferible una regla a un transformer?
5. ¿Qué dimensiones deben evaluarse en un resumen clínico?
6. ¿Por qué una puntuación alta en un examen no valida un modelo para decisiones clínicas?
7. ¿Cómo se verifica que una referencia respalda una afirmación?
8. ¿Qué diferencia hay entre seudonimización y anonimización?
9. ¿Qué elementos mínimos debe contener un prompt de extracción?
10. ¿Por qué el F1 global puede ocultar un problema clínico?
11. ¿Qué información debe incluir una declaración de uso de IA?
12. ¿Qué controles aplicaría al código R generado por un asistente?

Respuestas orientativas

1. Porque compartir vocabulario no determina si un signo está presente, negado, es histórico o pertenece a otro sujeto.
2. El estático asigna un vector estable a una palabra; el contextual cambia según la secuencia.
3. Porque el estilo y contenido repetido generan fuga y una estimación optimista.
4. Cuando la expresión es estable, la tarea es pequeña y la transparencia supera la ganancia esperada de complejidad.
5. Factualidad, completitud, relevancia, cronología, contradicciones y utilidad clínica.

6. Porque el entorno clínico incluye información incompleta, examen, consecuencias y distribución distinta.
7. Comprobando metadatos y leyendo la fuente primaria para vincularla con la proposición exacta.
8. La seudonimización conserva una clave o posibilidad de vinculación; la anonimización busca impedir identificación razonable.
9. Objetivo, contexto, definiciones, restricciones, esquema, manejo de incertidumbre y evidencia textual.
10. Porque las clases frecuentes dominan el promedio y pueden ocultar baja sensibilidad en un signo raro.
11. Herramienta/versión, fecha, tarea, datos, prompts, parámetros, validación y responsabilidad humana.
12. Inspección, pruebas con datos conocidos, verificación de paquetes, supuestos, semillas y comparación independiente.

Actividad práctica integradora

Forme equipos de tres personas. Construyan 40 notas clínicas simuladas de una especie, con cuatro signos, negaciones, antecedentes, errores ortográficos y menciones de otro sujeto. Definan un manual de anotación; dos integrantes etiquetan de forma independiente y el tercero adjudica. No utilicen datos clínicos reales identificables.

- Implementen una línea base por reglas y un segundo método contextual o generativo.
- Congelen 10 notas como prueba antes de ajustar prompts o reglas.
- Calculen precisión, sensibilidad y F1 por signo; documenten desacuerdos humanos.
- Clasifiquen cada error por negación, temporalidad, sujeto, abreviatura o inferencia indebida.
- Redacten una declaración de uso de IA y una matriz de riesgos de privacidad.
- Entreguen protocolo, manual, datos simulados, código, prompts versionados, resultados y reflexión crítica.

CAPÍTULO 11

INTELIGENCIA ARTIFICIAL EN MEDICINA CLÍNICA CIRUGÍA Y ESPECIALIDADES VETERINARIAS

Introducción

La medicina clínica veterinaria genera datos heterogéneos en momentos en los que las decisiones deben adoptarse con rapidez y bajo incertidumbre. La anamnesis, la exploración física, los resultados de laboratorio, las imágenes, las señales fisiológicas y la evolución temporal describen facetas distintas de un mismo paciente. Los métodos de inteligencia artificial pueden ordenar esa información, identificar patrones y estimar probabilidades; no convierten una observación incompleta en verdad clínica ni sustituyen el juicio profesional. Su valor depende de la calidad del problema científico, la definición del desenlace, la representatividad de los animales estudiados y la forma en que la predicción se inserta en el trabajo asistencial.

Este capítulo examina la IA aplicada a perros y gatos, con extensiones a otras especies cuando el razonamiento es transferible. La discusión distingue tres funciones: automatizar mediciones, estimar riesgos y apoyar decisiones. Se insiste en una diferencia central. Un algoritmo diagnóstico intenta reconocer una condición presente; uno pronóstico estima un resultado futuro; y un sistema prescriptivo propone una acción cuyos beneficios y daños deben evaluarse de manera causal. Una AUC alta no demuestra que usar el modelo mejore la supervivencia, reduzca complicaciones o alivie dolor.

La evidencia veterinaria disponible muestra avances concretos en análisis radiográfico, auscultación digital, oftalmología y evaluación facial del dolor. Sin embargo, muchos estudios son retrospectivos, proceden de pocos centros y evalúan el desempeño en conjuntos de datos similares a los utilizados para desarrollar el modelo. Por ello, el tránsito desde una prueba de concepto hasta una herramienta clínica exige validación externa, evaluación prospectiva, análisis de calibración, estudio de impacto, vigilancia posimplementación y mecanismos explícitos para rechazar entradas de mala calidad.

Objetivos de aprendizaje

1. Diferenciar sistemas diagnósticos, pronósticos, de monitorización y de apoyo a la decisión clínica.

2. Reconocer aplicaciones y límites de la IA en medicina interna, cirugía, anestesiología, dolor, oncología, cardiología, neurología, dermatología, oftalmología, rehabilitación y farmacología.
3. Seleccionar variables, algoritmos, métricas y estrategias de validación coherentes con una pregunta clínica veterinaria.
4. Interpretar discriminación, calibración, utilidad clínica y errores según sus consecuencias para el paciente.
5. Diseñar estudios multimodales sin fuga de información, pseudorreplicación ni particiones incorrectas por imagen o visita.
6. Formular protocolos de implementación que conserven la responsabilidad del médico veterinario y protejan la confidencialidad de los datos.

Tabla 11.1

Funciones clínicas de los sistemas de inteligencia artificial

Función	Pregunta	Salida	Validación prioritaria	Riesgo clínico
Medición	¿Cuánto mide o dónde está?	Valor, contorno o señal	Concordancia, error absoluto, IoU	Propagar una medición sesgada
Diagnóstico	¿Qué condición está presente?	Probabilidad o clase	Sensibilidad, especificidad, LR, calibración	Falso negativo o sobrediagnóstico
Pronóstico	¿Qué ocurrirá y cuándo?	Riesgo futuro o supervivencia	Calibración temporal, C-index, Brier	Intervención innecesaria o tardía
Monitorización	¿Está cambiando el estado?	Alerta o tendencia	Tiempo de detección, falsas alarmas	Fatiga de alertas
Decisión	¿Qué acción conviene?	Recomendación contextual	Impacto prospectivo y utilidad clínica	Automatización acrítica

11.1 Sistemas de apoyo a la decisión clínica

Un sistema de apoyo a la decisión clínica recibe datos del paciente y devuelve información diseñada para modificar una decisión concreta. Puede priorizar diferenciales, recordar una contraindicación, estimar el riesgo de deterioro, sugerir una prueba confirmatoria o advertir que una dosis calculada excede un intervalo previamente definido. La utilidad no reside en producir una lista extensa, sino en entregar una señal pertinente, trazable y oportuna dentro del flujo de trabajo.

El diseño debe comenzar con una decisión clínica observable. Por ejemplo, en un perro con vómito agudo, el objetivo no es “diagnosticar con IA”, sino estimar si la combinación de edad, duración de los signos, perfusión, lactato, electrolitos, ecografía focal y respuesta inicial justifica hospitalización inmediata. Esta formulación identifica población, momento de predicción, variables disponibles y acción asociada. También impide incluir información obtenida después de la decisión, una forma frecuente de fuga de datos.

Los sistemas basados en reglas codifican conocimiento explícito, mientras que los modelos estadísticos o de aprendizaje automático estiman relaciones desde datos. Ambos pueden combinarse. Una regla puede bloquear una recomendación farmacológica si hay insuficiencia renal grave; un modelo puede calcular el riesgo de lesión renal aguda; y el veterinario decide considerando diagnóstico, bienestar, preferencias del tutor y recursos. La presencia de un algoritmo no elimina el deber de comprobar identidad, especie, peso, unidades y contexto.

Figura 11.1.

Sistema veterinario de apoyo a la decisión clínica



La interfaz influye en el resultado tanto como el modelo. Una alerta sin umbral de acción, sin explicación de las variables relevantes o sin indicación de incertidumbre favorece el sesgo de automatización. Debe mostrarse el horizonte temporal, la población en la que se validó, el

significado de la probabilidad y las condiciones que invalidan la salida. Los usuarios necesitan registrar cuándo aceptan o rechazan la recomendación y por qué. Ese registro permite estudiar seguridad, deriva y utilidad real.

TRIPOD+AI orienta el reporte transparente de modelos predictivos desarrollados con regresión o aprendizaje automático, mientras DECIDE-AI se concentra en evaluaciones clínicas tempranas de sistemas de apoyo durante su uso real (Collins et al., 2024; Vasey et al., 2022). Aunque estas guías proceden de salud humana, sus principios metodológicos son aplicables a investigación veterinaria: describir participantes, flujo de datos, interacción humano-algoritmo, fallos, modificaciones y resultados clínicos.

11.2 Predicción diagnóstica y pronóstica

La predicción diagnóstica estima la probabilidad de una condición en el momento de la evaluación. La pronóstica se refiere a un desenlace posterior, como recaída, complicación, hospitalización o muerte. Esta diferencia determina el tiempo cero, la ventana de observación y los predictores legítimos. Si se pretende predecir hipotensión durante la anestesia, solo pueden usarse datos disponibles antes del episodio; incluir la dosis de vasopresor administrada después produciría un modelo artificialmente preciso y clínicamente inútil.

La discriminación responde si el modelo ordena correctamente a pacientes con mayor y menor riesgo. La calibración compara probabilidades predichas con frecuencias observadas. Un modelo puede distinguir bien y, aun así, sobreestimar el riesgo en una clínica con menor gravedad de casos. Por ello deben presentarse curvas de calibración, pendiente, intercepto y, cuando sea posible, puntuación de Brier. En desenlaces raros, la exactitud global es engañosa: clasificar a todos los animales como negativos puede producir más de 99 % de aciertos y ninguna utilidad.

La sensibilidad debe priorizarse cuando perder un caso genera daño grave, como no detectar deterioro respiratorio. La especificidad adquiere mayor peso cuando una falsa alarma conduce a un procedimiento invasivo o costoso. Los valores predictivos cambian con la prevalencia, por lo que no deben trasladarse sin ajuste entre un hospital de referencia y una consulta primaria. Las razones de verosimilitud y el beneficio neto ayudan a relacionar el resultado con decisiones clínicas.

La validación aleatoria interna es insuficiente cuando hay observaciones repetidas. Todas las imágenes, visitas y segmentos de señal del mismo animal deben permanecer en una sola partición. Para estudiar generalización se recomienda separar hospitales, periodos o regiones. Una validación temporal reproduce mejor la pregunta de implementación: desarrollar con pacientes anteriores y evaluar en casos posteriores sin modificar el modelo.

La evaluación responsable comprende rendimiento promedio y subgrupos clínicamente plausibles: especie, raza, conformación, edad, sexo, condición corporal, equipo, protocolo de adquisición y centro. Las diferencias no prueban discriminación algorítmica por sí mismas; señalan dónde faltan datos, cambia la medición o el modelo puede ser inseguro. El análisis debe planificarse, con intervalos de confianza, y no limitarse a comparar porcentajes aislados.

Tabla 11.2

Métricas según el tipo de tarea clínica

Tarea	Métricas principales	Interpretación clínica	Error que debe vigilarse
Clasificación diagnóstica	Sensibilidad, especificidad, AUC, LR	Capacidad de detectar o descartar	Umbral elegido después de ver el test
Riesgo pronóstico	AUC, calibración, Brier	Probabilidad individual y orden de riesgo	Buena AUC con probabilidades sesgadas
Tiempo hasta evento	C-index, Brier temporal, calibración	Orden y riesgo a horizontes definidos	Censura o tiempos ignorados
Segmentación	Dice, IoU, distancia de borde	Coincidencia anatómica	Alta media con fallos focales peligrosos
Monitorización	Tiempo de detección, alarmas por hora	Alerta útil antes del deterioro	Fatiga por falsos positivos

11.3 Medicina interna de perros y gatos

La medicina interna reúne problemas en los que los signos son inespecíficos y los datos se acumulan de forma secuencial. Un modelo puede combinar antecedentes, exploración física, hemograma, bioquímica, urianálisis y tendencias para estimar riesgo de enfermedad renal,

endocrina, gastrointestinal o infecciosa. La predicción debe complementar el razonamiento por síndromes y la confirmación diagnóstica. Una probabilidad alta de diabetes mellitus no reemplaza la demostración de hiperglucemia persistente, glucosuria y evaluación clínica apropiada.

Las historias clínicas electrónicas permiten estudiar grandes poblaciones, pero contienen abreviaturas, códigos variables, resultados repetidos y diagnósticos que reflejan prácticas locales. En un estudio multicéntrico, Szlosek et al. (2024) evaluaron un modelo de gradient boosting para clasificar visitas de bienestar en perros y gatos frente al consenso de veterinarios. El trabajo ilustra una aplicación útil para investigación de servicios preventivos, y también una limitación: una clasificación administrativa correcta no equivale a detectar enfermedad subclínica ni demuestra beneficio asistencial.

Los datos longitudinales son especialmente valiosos. Una creatinina dentro del intervalo de referencia puede adquirir relevancia si aumenta sostenidamente respecto del valor basal. Para modelar estas trayectorias se pueden resumir pendientes, variabilidad y velocidad de cambio, o usar modelos mixtos y métodos de series temporales. El corte debe realizarse por animal y el momento índice debe fijarse antes del desenlace. El algoritmo no debe aprender indirectamente el diagnóstico desde tratamientos, derivaciones o pruebas ordenadas como consecuencia de la sospecha clínica.

En urgencias, la IA puede apoyar triaje, pero los datos faltantes rara vez son aleatorios. Una prueba no solicitada indica que el clínico consideró innecesaria esa medición, que el tutor no autorizó el costo o que el paciente estaba demasiado inestable. Imputar sin representar el patrón de ausencia puede borrar información o introducir inequidad. Debe compararse el modelo con escalas simples y con el desempeño del veterinario en el mismo momento clínico.

Un uso prudente es priorizar revisiones, no emitir diagnósticos autónomos. Por ejemplo, un sistema puede señalar gatos geriátricos con pérdida de peso, aumento progresivo de creatinina y descenso de densidad urinaria para revisión profesional. La salida debe explicar que se trata de una alerta de patrón, identificar los datos faltantes y recomendar confirmación mediante un protocolo clínico predefinido.

11.4 Cirugía y seguimiento posoperatorio

En cirugía, la IA puede intervenir antes, durante y después del procedimiento. En la planificación preoperatoria, la segmentación de tomografía permite estimar volumen tumoral, relación con estructuras vasculares o tamaño de implantes. Durante la intervención, el análisis de video puede reconocer fases, instrumentos y eventos, aunque la evidencia veterinaria clínica todavía es limitada. En el seguimiento, fotografías seriadas, temperatura, actividad, ingesta y registros del tutor pueden contribuir a detectar infección, dehiscencia o recuperación más lenta de la esperada.

La automatización de una medición no elimina el error de adquisición. Fotografías de heridas tomadas con distinta iluminación, distancia o ángulo producen cambios aparentes. Un protocolo debe incluir escala métrica, orientación, distancia, iluminación y momento de captura. Si se calculan área, eritema o exudado, conviene comparar el algoritmo con evaluadores entrenados y cuantificar concordancia, no solo correlación. Dos métodos pueden correlacionarse y presentar un sesgo clínicamente inaceptable.

La predicción de complicaciones posoperatorias requiere definir con precisión el desenlace. “Complicación” puede mezclar infección superficial, reintervención, náusea o muerte, eventos con mecanismos y costos diferentes. Cuando el número de eventos es bajo, un resultado compuesto puede ser necesario, pero sus componentes deben informarse. También deben considerarse riesgos competitivos: un animal dado de alta pronto deja de aportar observaciones hospitalarias, aunque el evento pueda aparecer en casa.

Los dispositivos portátiles ofrecen actividad y sueño como marcadores de recuperación. La línea basal previa al procedimiento es más informativa que un umbral universal, pues el patrón depende de edad, raza, vivienda y enfermedad previa. Un algoritmo de anomalías puede detectar desviaciones respecto del propio paciente. La alerta debe conducir a una evaluación, no a la modificación automática de analgesia o antibióticos.

El estándar de referencia puede combinar examen clínico, cultivo cuando corresponda, evaluación ciega de la herida y desenlaces objetivos. Si la fotografía se toma después de que el veterinario ha limpiado o tratado la lesión, el modelo puede aprender señales del manejo y no de la complicación. Registrar el tiempo respecto de curaciones, fármacos y retiro de apósitos es parte del diseño.

11.5 Anestesiología y monitorización

La anestesia produce señales continuas de frecuencia cardiaca, presión arterial, oximetría, capnografía, temperatura y concentración de gases. El análisis automático puede filtrar artefactos, estimar tendencias y advertir combinaciones precursoras de hipotensión, hipoventilación o recuperación prolongada. El desafío no es reconocer un valor fuera de rango, tarea que resuelve una alarma convencional, sino interpretar secuencias, calidad de señal y contexto farmacológico.

La mortalidad asociada a anestesia es un desenlace grave y poco frecuente. Un análisis mundial de 55 022 anestesias caninas informó 0,69 % de mortalidad relacionada con anestesia y situó la mayoría de las muertes en el periodo posoperatorio (Redondo et al., 2024). Las cifras dependen de población, definición temporal y gravedad; no constituyen un umbral universal. El estudio refuerza que la recuperación necesita vigilancia y que cualquier modelo debe abarcar el periodo posterior a la extubación.

La clase ASA, edad, urgencia, comorbilidad, tipo de procedimiento y datos preoperatorios pueden alimentar un modelo de riesgo. Las variables intraoperatorias permiten actualizarlo, pero cambian la pregunta: un modelo preoperatorio apoya planificación y consentimiento; uno dinámico apoya vigilancia. Mezclarlos sin declarar el momento de predicción confunde interpretación. Las guías de AAHA recomiendan evaluar y monitorizar al paciente a lo largo del continuo anestésico, principio que debe conservarse al incorporar algoritmos (Grubb et al., 2020).

La calidad de señal merece una salida propia. Movimiento, perfusión periférica deficiente, electrodos desplazados y calibración incorrecta generan datos plausibles pero falsos. Un sistema seguro debe declarar “no evaluable” cuando la incertidumbre supera un límite. Forzar una predicción en cada segundo crea falsa confianza. La tasa de alarmas por hora, el tiempo de anticipación y las alarmas clínicamente accionables son métricas tan importantes como la AUC.

Para validar una alerta de hipotensión, la unidad de análisis es el animal o el episodio, no cada segundo del monitor. Miles de puntos del mismo paciente no equivalen a miles de casos independientes. Se requiere partición por animal, evaluación en equipos distintos y análisis de cómo la intervención del anestesista modifica la trayectoria. Si el clínico trata una caída temprana, el desenlace puede no ocurrir; este fenómeno complica el etiquetado retrospectivo.

11.6 Evaluación automatizada del dolor

El dolor es una experiencia multidimensional que se infiere mediante comportamiento, postura, expresión facial, interacción, respuesta a palpación y cambios fisiológicos. Ninguna señal aislada representa el dolor de forma completa. Los modelos automatizados pueden estandarizar observaciones y ampliar la frecuencia de medición, pero deben anclarse en escalas validadas y conservar la evaluación clínica. Las guías AAHA y WSAVA recomiendan valoración sistemática y manejo multimodal del dolor en perros y gatos (Gruen et al., 2022; Monteiro et al., 2023).

Las expresiones faciales permiten tareas de detección de puntos anatómicos, clasificación y estimación de puntuaciones. Steagall et al. (2023) desarrollaron modelos profundos automatizados para predecir dolor mediante la Feline Grimace Scale con posible uso en teléfonos. El resultado es relevante, aunque no autoriza extrapolar el sistema a todas las razas, edades, tipos de dolor o condiciones de iluminación. El rendimiento debe estudiarse en gatos braquicéfalos, con enfermedad ocular, sedación o expresiones atípicas.

El video aporta dinámica corporal: levantarse, proteger una región, cambios de marcha, inquietud o disminución de interacción. La fusión con acelerometría, frecuencia cardiaca y escalas observacionales puede mejorar la estabilidad, pero también aumenta el riesgo de aprender variables espurias. El modelo podría asociar jaula, collar isabelino o presencia del personal con dolor porque esos elementos coinciden con el posoperatorio.

La etiqueta de referencia no debe depender de un único observador. Es preferible usar dos o más evaluadores ciegos, una escala validada, reglas para resolver discrepancias y mediciones antes y después de analgesia. La respuesta a un analgésico puede aportar evidencia, aunque tampoco constituye una prueba perfecta. Debe cuantificarse concordancia interevaluador y reconocer el error de la etiqueta durante el entrenamiento.

El umbral clínico debe vincularse a una acción segura. Una alerta puede solicitar reevaluación por el equipo; no debe administrar una dosis sin comprobar fármaco, intervalo, función renal, interacción y estado hemodinámico. Los falsos negativos prolongan sufrimiento. Los falsos positivos pueden inducir sedación, efectos adversos o enmascarar otras causas de conducta. La selección del umbral debe reflejar ambos daños.

11.7 Oncología veterinaria

La oncología combina imágenes, citología, histopatología, inmunohistoquímica, datos moleculares, tratamiento y supervivencia. La IA puede detectar figuras mitóticas, segmentar núcleos, cuantificar marcadores o estimar pronóstico. La contribución más inmediata suele ser reducir variabilidad en tareas repetitivas y proporcionar mediciones reproducibles para revisión del patólogo. La clasificación final conserva la integración morfológica y clínica.

En mastocitoma cutáneo canino, la binucleación y multinucleación forman parte de criterios asociados con comportamiento tumoral. Bertram et al. (2021) publicaron un conjunto de imágenes completas con anotaciones de células bi y multinucleadas y evaluaron detección profunda. El estudio muestra el valor de bases abiertas y del etiquetado asistido por expertos, pero también que una métrica de detección moderada no debe convertirse directamente en una decisión terapéutica.

Los modelos pronósticos pueden integrar grado histológico, índice mitótico, tamaño, localización, estadio, márgenes, mutaciones y tratamiento. Si el objetivo es supervivencia, Kaplan-Meier y Cox siguen siendo herramientas apropiadas para describir y modelar tiempo hasta evento; random survival forests o boosting pueden captar no linealidad si el tamaño muestral y el número de eventos lo permiten. La comparación debe usar las mismas particiones, horizontes y métricas, además de calibración.

El tratamiento crea confusión por indicación: pacientes con peor pronóstico reciben terapias más intensas. Un algoritmo retrospectivo puede asociar tratamiento con mortalidad y concluir erróneamente que la terapia causa daño. Predecir no equivale a estimar efecto terapéutico. Para decidir entre cirugía, quimioterapia, radioterapia u observación se necesitan ensayos, métodos causales o evidencia clínica pertinente, no solo asociaciones predictivas.

La validación externa debe considerar laboratorio, escáner, tinción, especie, raza y distribución de tumores. Los modelos histológicos son sensibles a coloración y preparación. La normalización digital ayuda, pero no reemplaza muestras de múltiples centros. Los mapas de atención pueden señalar regiones que influyen en la salida; no demuestran que el razonamiento sea biológicamente correcto.

11.8 Cardiología y neurología

En cardiología, las señales digitales facilitan clasificación de electrocardiogramas, fonocardiogramas y ecocardiografía. McDonald et al. (2024) estudiaron 756 perros y ajustaron un algoritmo para detectar y graduar soplos a partir de estetoscopios electrónicos. La sensibilidad para detectar cualquier soplo fue 87,9 % y la especificidad 81,7 %. Para distinguir estadios preclínicos B1 y B2 de enfermedad mixomatosa mitral, el área bajo la curva fue 0,861. Estas cifras apoyan el cribado, no la sustitución del ecocardiograma cuando está indicado.

Kim et al. (2023) desarrollaron modelos de aprendizaje automático con historias clínicas para predecir insuficiencia cardiaca en perros con enfermedad mixomatosa de la válvula mitral. Este tipo de sistema puede priorizar seguimiento, siempre que se evalúen calibración y rendimiento en hospitales ajenos al desarrollo. Los datos faltantes, la variación entre ecocardiografistas y la selección de pacientes derivados condicionan la generalización.

En neurología, la IA puede segmentar lesiones en resonancia, clasificar electroencefalogramas, cuantificar marcha y reconocer crisis en video. La referencia diagnóstica suele ser compleja: epilepsia, síncope, discinesia y episodios conductuales pueden parecer similares. Un conjunto etiquetado solo con notas clínicas transmite la incertidumbre del diagnóstico. El consenso de neurólogos, seguimiento y pruebas complementarias fortalecen el estándar de referencia.

La cuantificación de marcha mediante video o sensores puede apoyar el seguimiento de enfermedad medular, rehabilitación o efectos de tratamiento. Longitud de zancada, simetría, velocidad y variabilidad deben validarse contra sistemas cinemáticos o plataformas de fuerza cuando se afirme equivalencia. El piso, la correa, el guía y el ritmo inducido alteran la medición; el protocolo de adquisición forma parte del modelo.

Los sistemas de monitorización domiciliaria abren una oportunidad y un riesgo. Permiten observar eventos raros fuera de la consulta, pero generan videos sensibles del hogar y grandes volúmenes de falsos positivos. Se requieren consentimiento, minimización de datos, cifrado, tiempos de retención y un plan sobre quién revisa las alertas y en qué plazo.

11.9 Dermatología y oftalmología

Las lesiones dermatológicas pueden fotografiarse con equipos accesibles, lo que favorece proyectos de visión computacional. Sin embargo, eritema, alopecia, costras y pigmentación no son diagnósticos etiológicos. Una red puede clasificar apariencia, pero la confirmación de pioderma, dermatofitosis, demodicosis o hipersensibilidad requiere historia, distribución, citología, raspado, cultivo u otras pruebas. La salida más segura puede ser describir la lesión y priorizar diferenciales, no asignar una enfermedad única.

La variación por color de pelaje y piel, recorte, flash, cámara y tratamiento previo debe quedar representada. Una base descargada de internet puede contener fondos, marcas de agua o estilos fotográficos asociados a cada etiqueta. La división por imagen permitiría que fotografías del mismo caso aparezcan en entrenamiento y prueba. Se debe dividir por animal, conservar un conjunto externo y publicar ejemplos de fallos.

En oftalmología, Kim et al. (2022) aplicaron detección de objetos para evaluar estabilidad de la película lagrimal precorneal en perros. Aunque informaron una precisión media elevada sobre imágenes derivadas de videos de un número pequeño de animales, el tamaño efectivo es el número de perros, no los miles de fotogramas. El trabajo ilustra por qué el desempeño por imagen debe interpretarse junto con independencia, diversidad clínica y validación externa.

La IA también puede segmentar úlceras, opacidades, vasos o estructuras retinianas. La adquisición debe estandarizarse y documentar midriasis, sedación, iluminación y equipo. Enfermedades que alteran transparencia o anatomía pueden causar exclusiones sistemáticas, justo en pacientes complejos. Por ello se necesita reportar proporción de imágenes no evaluables y razones de rechazo.

En ambas especialidades, el seguimiento cuantitativo puede ser más factible que el diagnóstico autónomo. Comparar área de una lesión o progresión de opacidad dentro del mismo paciente reduce parte de la variabilidad entre individuos. Aun así, se debe estimar el error mínimo detectable para distinguir cambio real de fluctuación por fotografía.

11.10 Medicina regenerativa y rehabilitación

La medicina regenerativa veterinaria estudia productos como plasma rico en plaquetas, concentrados celulares y células estromales mesenquimales. La IA puede caracterizar imágenes, predecir calidad de preparación, identificar perfiles de respuesta y relacionar biomarcadores con evolución. No debe usarse para encubrir heterogeneidad biológica ni para justificar indicaciones sin evidencia. Fuente celular, procesamiento, dosis, viabilidad, vía, momento y enfermedad definen intervenciones distintas.

En rehabilitación, visión computacional, plataformas de fuerza y sensores inerciales permiten cuantificar apoyo, simetría, amplitud articular y actividad. Un sistema puede detectar recuperación después de cirugía ortopédica o variación en osteoartritis. Las mediciones necesitan validación analítica y clínica. Una diferencia estadística en simetría no implica una mejoría percibida por el animal o el tutor.

Los modelos de respuesta al tratamiento son vulnerables a muestras pequeñas y múltiples variables. Si se exploran decenas de citocinas, características de imagen y parámetros de marcha en pocos animales, el sobreajuste es probable. Es preferible predefinir un desenlace primario, limitar predictores con fundamento biológico, usar reducción de dimensionalidad dentro de la validación cruzada y confirmar los hallazgos en una cohorte independiente.

La rehabilitación produce medidas repetidas. Un modelo mixto puede estimar evolución media y variabilidad individual; un algoritmo de predicción puede estimar qué pacientes requieren seguimiento más intenso. No son objetivos equivalentes. La unidad experimental, el efecto del terapeuta y la adherencia domiciliaria deben registrarse. Si ambos miembros de un perro se analizan, su dependencia no puede ignorarse.

La aplicación clínica más responsable es adaptar frecuencia de reevaluación o detectar estancamiento, no prometer regeneración. Los resultados deben integrar dolor, función, calidad de vida, eventos adversos y medidas objetivas. Cuando la evidencia de una intervención es incierta, un modelo complejo no eleva su nivel de certeza.

11.11 Farmacología y medicina personalizada

La medicina personalizada busca ajustar fármaco, dosis e intervalo a características del paciente. En veterinaria, especie, raza, edad, composición corporal, función renal y hepática, estado crítico, interacciones y variantes genéticas pueden modificar exposición y respuesta. Los modelos farmacocinéticos poblacionales y fisiológicamente basados mantienen una interpretación mecanística; el aprendizaje automático puede complementar la predicción cuando existen datos suficientes.

Un sistema de dosis debe distinguir predicción de concentración, probabilidad de eficacia y riesgo de toxicidad. La concentración estimada no es un resultado clínico. En antimicrobianos, la selección incorpora microorganismo, concentración mínima inhibitoria, sitio de infección, farmacocinética y farmacodinamia. Automatizar una recomendación sin estos elementos puede reforzar uso inapropiado y resistencia.

La variación genética tiene consecuencias conocidas en algunas razas, pero usar raza como sustituto perfecto de genotipo es impreciso. Cuando una variante es relevante y existe prueba disponible, el dato genético directo es preferible. La base de entrenamiento debe incluir poblaciones diversas y declarar cómo se definieron raza y mezcla. El algoritmo no debe extrapolar dosis entre especies sin evidencia.

El aprendizaje por refuerzo se propone para ajustar infusiones o tratamientos secuenciales. Su uso clínico exige extrema cautela porque aprende una política a partir de recompensas definidas por el investigador. Si la recompensa prioriza estabilidad hemodinámica inmediata, puede ignorar recuperación, dolor o toxicidad. La evaluación inicial debe realizarse fuera de línea, con simulación y límites farmacológicos, antes de cualquier estudio prospectivo.

La responsabilidad permanece en el prescriptor. La interfaz debe mostrar datos usados, fecha de laboratorio, supuestos, intervalo plausible y condiciones que obligan a revisar la recomendación. Los sistemas requieren control de versiones, porque una actualización puede modificar dosis o alertas. Cada recomendación debe quedar auditada.

11.12 Integración de datos clínicos multimodales

Los datos multimodales combinan fuentes que describen niveles distintos del paciente: texto clínico, variables tabulares, imágenes, audio, video, genómica y series temporales. La fusión temprana concatena características antes del modelo; la fusión tardía combina predicciones; y los modelos conjuntos aprenden representaciones compartidas. No existe una arquitectura universalmente superior. La elección depende de tamaño muestral, disponibilidad, sincronización y mecanismo clínico.

La integración requiere identidad y tiempo consistentes. Un resultado de laboratorio debe vincularse al episodio correcto; una radiografía tomada después del deterioro no puede predecir el estado previo; y un video de dolor necesita una ventana respecto de cirugía y analgesia. Los errores de vinculación pueden ser más dañinos que la elección del algoritmo. Un diccionario de datos y una línea temporal por paciente son requisitos básicos.

Las modalidades faltantes son parte de la práctica. Un modelo que exige resonancia, genómica y video puede excluir a la mayoría de pacientes. Se puede entrenar con máscaras de ausencia, modelos por combinación o rutas alternativas, pero deben evaluarse patrones reales. La imputación de una imagen o señal completa puede crear información ficticia y no debe presentarse como observación del paciente.

El beneficio de añadir una modalidad se demuestra comparando modelos anidados. Si edad, ASA y procedimiento logran buena calibración para riesgo anestésico, incorporar video facial y cientos de variables del monitor solo se justifica si mejora utilidad clínica, estabilidad y oportunidad, no por una diferencia mínima de AUC. También deben evaluarse costo, tiempo, privacidad y posibilidad de fallo.

Una explicación multimodal debe indicar qué fuentes sostuvieron la predicción y cuáles faltaron. Los mapas de calor, valores SHAP o atención son aproximaciones del comportamiento del modelo; no explicaciones causales. La revisión clínica puede revelar dependencia de artefactos, como tubos, etiquetas o marcas de equipo. La prueba de ablación, en la que se retira una modalidad, ayuda a examinar dependencia y redundancia.

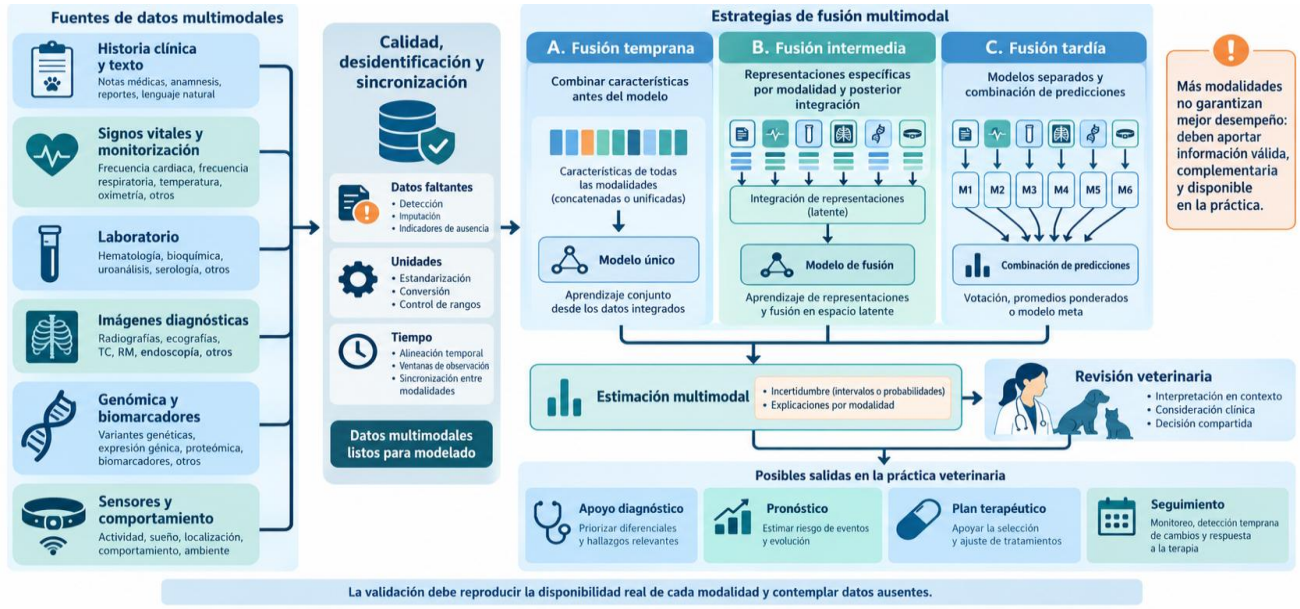
Tabla 11.3

Control metodológico de un estudio clínico multimodal

Componente	Decisión que debe documentarse	Fallo frecuente
Población	Criterios, centros, periodo y flujo de inclusión	Solo casos completos o de referencia
Tiempo índice	Momento exacto de la predicción	Variables posteriores al desenlace
Identidad	Vínculo de visitas, imágenes y señales por animal	Dividir fotogramas del mismo paciente
Referencia	Evaluadores, pruebas y resolución de desacuerdo	Etiqueta tomada de una nota incierta
Ausencia	Patrón y manejo por modalidad	Imputación que oculta acceso desigual
Validación	Interna, temporal, geográfica y prospectiva	Un único test similar al entrenamiento
Impacto	Acción, beneficio, daño y carga de alertas	Suponer utilidad por una AUC alta

Figura 11.2.

Estrategias de integración de datos multimodales



11.13 Caso didáctico predicción del riesgo anestésico

Caso didáctico simulado. Los datos, resultados y métricas de este apartado son ficticios y se presentan únicamente para enseñar diseño y análisis. No deben utilizarse como tasas reales ni como reglas de decisión clínica.

Planteamiento científico

Un hospital veterinario desea estimar, antes de la inducción, el riesgo de un evento anestésico mayor en perros sometidos a cirugía. El objetivo operativo es identificar pacientes que podrían beneficiarse de revisión por anestesiólogo, monitorización invasiva o recuperación en cuidados intensivos. El sistema no elige el protocolo ni autoriza la cirugía.

Pregunta: en perros que recibirán anestesia general, ¿qué capacidad tienen los datos preoperatorios disponibles para predecir un evento mayor desde la inducción hasta 24 horas después de la extubación? Hipótesis: un modelo con edad, condición corporal, clase ASA, urgencia, tipo de cirugía, hematocrito, glucosa, creatinina, albúmina y comorbilidades tendrá mejor calibración y beneficio neto que clasificar solo por ASA.

Figura 11.3.

Línea temporal del riesgo anestésico y dolor posoperatorio

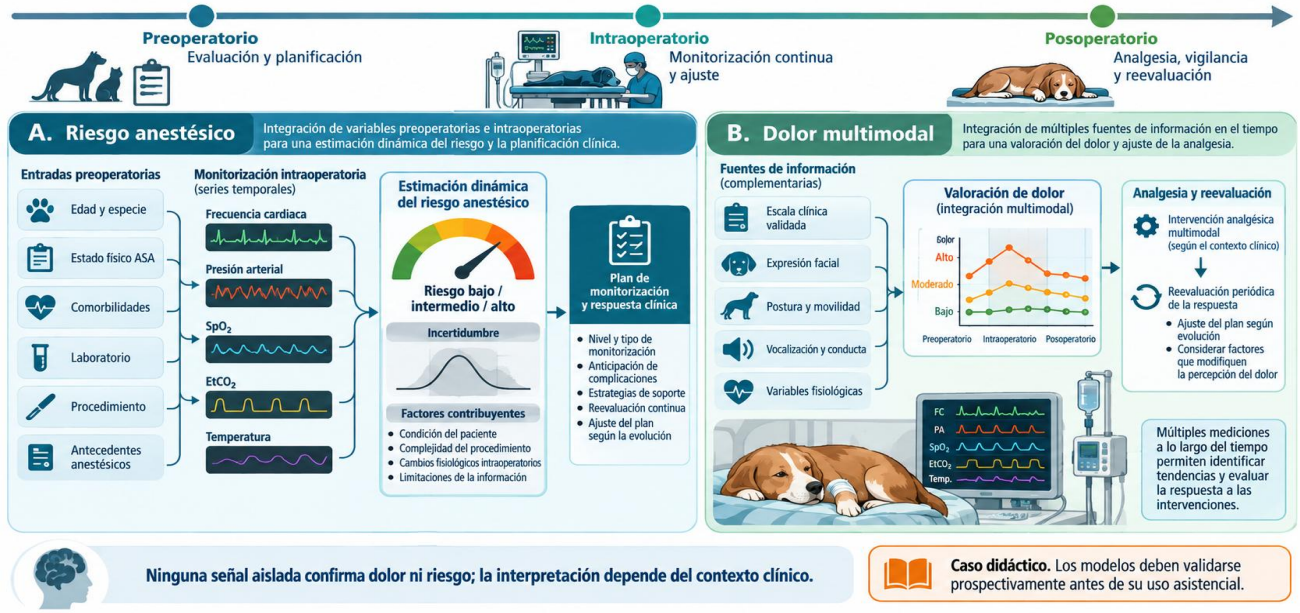


Tabla 11.4

Protocolo del caso simulado de riesgo anestésico

Elemento	Especificación
Población	Perros anestesiados en tres hospitales durante cuatro años
Unidad de análisis	Primer procedimiento por perro en el periodo
Tiempo índice	Cierre de la evaluación preanestésica y antes de premedicar
Desenlace	Paro, ventilación no planificada, vasopresor persistente, reintubación o muerte relacionada en 24 h
Predictores	Solo variables disponibles en el tiempo índice
Comparadores	ASA; regresión logística penalizada; gradient boosting
Validación	Desarrollo en hospitales A y B; validación externa en C y validación temporal
Métricas	AUC, sensibilidad, especificidad, calibración, Brier y beneficio neto
Uso previsto	Alerta para revisión y plan de vigilancia reforzada

Construcción y validación

La cohorte simulada contiene 3 600 perros y 108 eventos, prevalencia del 3 %. El conjunto de desarrollo incluye 2 700 animales y el hospital externo 900. Los valores faltantes se imputan dentro de cada partición de validación para evitar fuga. Las variables continuas conservan su escala y se

modelan con relaciones no lineales cuando existe fundamento; no se convierten arbitrariamente en normal o anormal.

Se usa validación cruzada agrupada por hospital y periodo para ajustar penalización e hiperparámetros. El umbral se define en desarrollo con prioridad a sensibilidad, porque omitir un paciente de alto riesgo puede ser grave. Antes de fijarlo se calcula cuántas alertas puede revisar el servicio. La utilidad se compara con las estrategias de revisar a todos, no revisar a ninguno y revisar solo ASA III a V.

Tabla 11.5

Resultados simulados en la validación externa

Modelo	AUC	Sensibilidad	Especificidad	Brier	Calibración
ASA	0,70	0,68	0,63	0,031	Riesgo no individualizado
Logística penalizada	0,79	0,82	0,66	0,027	Intercepto -0,08; pendiente 0,94
Gradient boosting	0,82	0,85	0,68	0,026	Intercepto -0,21; pendiente 0,81

Interpretación crítica

El boosting discrimina algo mejor, pero sobrestima los extremos y presenta pendiente de calibración menor que uno. La regresión ofrece probabilidades más estables y puede ser preferible si su beneficio neto es similar. La decisión no debe basarse en la AUC. Se requiere recalibración local y una prueba prospectiva silenciosa, durante la cual se calculan riesgos sin mostrarlos al equipo, para medir deriva y carga de alertas.

El modelo no demuestra que la monitorización reforzada reduzca eventos. Esa hipótesis exige un estudio de impacto, por ejemplo una implementación escalonada por servicio. Deben compararse complicaciones, retrasos, consumo de recursos y posibles daños. Si el modelo envía más pacientes a cuidados intensivos sin mejorar desenlaces, su rendimiento estadístico no justifica el uso.

Limitaciones del caso: el desenlace compuesto mezcla eventos de gravedad distinta; los hospitales pueden codificar ASA de forma diferente; y excluir procedimientos repetidos reduce aplicabilidad a animales sometidos a anestesias sucesivas. Un cambio en monitores o protocolos puede alterar calibración. La auditoría debe revisar subgrupos y cada evento adverso, no solo promedios.

11.14 Caso didáctico evaluación del dolor posoperatorio mediante datos multimodales

Caso didáctico simulado. Los datos y resultados que siguen son ficticios. El propósito es mostrar cómo diseñar una investigación multimodal sin presentar la salida del algoritmo como sustituto de la valoración clínica.

Pregunta diseño y referencia

Se estudian perros sometidos a ovariohisterectomía electiva. La pregunta es si video facial y corporal, acelerometría y variables fisiológicas permiten identificar episodios que requieren reevaluación analgésica durante las primeras 24 horas. La unidad experimental es el perro; la unidad observacional es la ventana de diez minutos. Todas las ventanas de un mismo animal permanecen en la misma partición.

La referencia se obtiene mediante la Glasgow Composite Measure Pain Scale Short Form aplicada independientemente por dos evaluadores entrenados y ciegos a la salida del modelo. Las discrepancias cercanas al umbral se resuelven mediante un tercer evaluador. Se registran hora, fármacos, sedación, presencia del personal y tiempo desde cirugía. Se incluye una evaluación antes y después de analgesia de rescate cuando está clínicamente indicada.

Tabla 11.6

Modalidades del caso simulado de dolor posoperatorio

Modalidad	Características	Control de calidad	Riesgo de confusión
Video facial	Orejas, ojos, tensión facial, orientación	Visibilidad y ángulo	Braquicefalia o enfermedad ocular
Video corporal	Postura, cambios de posición, protección	Cuerpo completo visible	Jaula, vendaje o personal

Acelerometría	Actividad, inmovilidad, transiciones	Sincronización y fijación	Temperamento y sedación
Fisiología	FC, FR, temperatura y tendencia	Artefactos y momento	Estrés, fiebre o fármacos
Escala clínica	Puntuación y decisión de reevaluación	Concordancia entre evaluadores	Subjetividad y conocimiento del caso

Modelos y análisis

Se reclutan de manera simulada 240 perros en dos centros. El centro A se usa para desarrollo y el B para validación externa. Se comparan un modelo clínico simple, modelos unimodales y una fusión tardía. La selección de ventanas evita que el número de episodios de cada perro domine el entrenamiento. Las métricas se calculan por ventana y por animal; esta última evaluación responde si el sistema identifica al menos un episodio relevante sin generar alarmas excesivas.

Tabla 11.7

Resultados simulados del modelo de dolor

Modelo	AUC por ventana	Sensibilidad	Alarmas falsas por perro	Observación
Clínico simple	0,72	0,70	0,9	Usa tiempo, FC y FR
Video	0,80	0,79	1,2	Falla con rostro oculto
Acelerometría	0,75	0,73	1,0	Confunde sedación e inmovilidad
Fusión multimodal	0,85	0,86	1,1	Mejor desempeño, mayor complejidad

Interpretación y seguridad

La fusión mejora la discriminación, pero una media de 1,1 alarmas falsas por perro puede sobrecargar al equipo. Se exploran umbrales según turno y capacidad de respuesta sin modificar el conjunto externo. El modelo debe emitir “datos insuficientes” si el rostro no es visible y no inferir dolor desde una señal ausente.

El análisis de errores identifica falsos negativos en perros quietos con puntuaciones clínicas altas y falsos positivos durante manipulación. Esto orienta nuevas variables y reglas de adquisición. No se corrigen los errores mediante exclusión retrospectiva de los casos difíciles, porque son parte de la población de uso. Se comparan razas braquicéfalas y no braquicéfalas con intervalos de confianza, reconociendo que el estudio puede no tener potencia para demostrar equivalencia.

La salida propuesta es “reevaluar dolor en los próximos diez minutos”, acompañada de probabilidad, modalidades disponibles y calidad. El veterinario confirma la puntuación, examina al paciente y decide. La evaluación de impacto debe medir tiempo hasta rescate, evolución del dolor, sedación, eventos adversos y satisfacción del equipo. El objetivo es reducir dolor no reconocido sin inducir medicación automática.

11.15 Síntesis autoevaluación y actividad práctica

Síntesis

La IA clínica veterinaria aporta valor cuando automatiza una medición bien definida, estima un riesgo relevante o presenta información que mejora una decisión. El desempeño depende de la población, el tiempo de predicción, la referencia y la calidad del dato. La unidad de análisis debe ser coherente con el paciente: imágenes, fotogramas y segundos repetidos no pueden repartirse entre entrenamiento y prueba como observaciones independientes.

La evaluación exige más que exactitud o AUC. Se necesitan sensibilidad y especificidad en el umbral de uso, calibración, incertidumbre, utilidad clínica, carga de alertas y análisis de fallos. La validación externa y prospectiva precede a la implementación. Después del despliegue se vigilan deriva, subgrupos, cambios de equipos y consecuencias no previstas.

En cirugía, anestesia, dolor y especialidades, el sistema debe apoyar al profesional y declarar cuándo no puede evaluar. La predicción no prueba que una intervención sea eficaz; una recomendación terapéutica exige evidencia causal y consideración de bienestar, preferencias del tutor, recursos y riesgos. La responsabilidad clínica no se transfiere al algoritmo.

Preguntas de autoevaluación

1. ¿Qué diferencia metodológica existe entre un modelo diagnóstico y uno pronóstico?
2. ¿Por qué una AUC alta no demuestra que usar el algoritmo mejore los desenlaces?
3. ¿Qué fuga de información aparecería si se usa un vasopresor administrado después del inicio de hipotensión para predecirla?
4. ¿Por qué todos los fotogramas de un animal deben permanecer en una misma partición?
5. ¿Qué aporta la calibración que no aporta la sensibilidad?
6. ¿Cómo cambia el valor predictivo positivo entre consulta primaria y hospital de referencia?
7. ¿Qué condiciones pueden confundir un sistema de evaluación facial del dolor?
8. ¿Por qué un modelo que asocia tratamiento oncológico con mortalidad no estima necesariamente el efecto del tratamiento?
9. ¿Qué diferencia existe entre validar una medición automática y validar una decisión clínica?
10. ¿Cuándo podría una regresión logística ser preferible a gradient boosting?
11. ¿Qué métricas usaría para evaluar una alerta continua además de la AUC?
12. ¿Qué información debería mostrar una interfaz clínica antes de que el veterinario acepte la salida?
13. ¿Cómo evaluaría un modelo multimodal cuando falta una de las modalidades?
14. ¿Qué significa que un modelo emita una salida no evaluable?
15. ¿Qué diseño permitiría estudiar si la implementación reduce complicaciones y no solo si predice bien?

Respuestas orientativas

1. El diagnóstico estima una condición presente; el pronóstico estima un evento futuro y requiere un tiempo cero y horizonte definidos.
2. La AUC mide ordenamiento entre casos y no casos. No mide calibración, daños, carga, adopción ni efecto de una intervención basada en la salida.
3. El vasopresor es consecuencia del episodio o de su sospecha; incluirlo permite al modelo leer información posterior al momento de predicción.
4. Las observaciones del mismo animal están correlacionadas. Separarlas produce contaminación entre entrenamiento y prueba y sobreestima rendimiento.
5. La calibración determina si una probabilidad, por ejemplo 20 %, corresponde aproximadamente a una frecuencia observada de 20 %.
6. Cambia con la prevalencia y selección de casos; suele ser mayor en poblaciones con más enfermedad, aun con sensibilidad y especificidad similares.
7. Conformación facial, enfermedad ocular, sedación, iluminación, ángulo, estrés, visibilidad y tipo de dolor.
8. Existe confusión por indicación: los pacientes más graves pueden recibir terapias más intensas y tener peor pronóstico por su enfermedad basal.
9. La primera compara la salida con una referencia; la segunda demuestra que usarla cambia conducta y mejora beneficios sin daños desproporcionados.

10. Cuando el tamaño muestral es limitado, se necesita interpretabilidad y calibración estable, y la relación puede representarse adecuadamente.
11. Tiempo de anticipación, alarmas falsas por hora o paciente, proporción de alertas accionables, fallos de señal y tiempo de respuesta.
12. Población de validación, horizonte, probabilidad, calidad de entrada, datos faltantes, incertidumbre, variables influyentes y límites de uso.
13. Con casos incompletos reales, comparando rutas por disponibilidad y reportando rendimiento para cada patrón de ausencia.
14. Que la entrada o incertidumbre no permite una estimación segura; es una conducta de seguridad, no un fallo que deba ocultarse.
15. Un estudio prospectivo de impacto, idealmente controlado o escalonado, con desenlaces clínicos, recursos, carga y eventos adversos.

Actividad práctica

Diseñe un protocolo para un sistema de alerta temprana de deterioro posoperatorio en perros. Puede elegir cirugía abdominal, ortopédica u oncológica. El producto final debe ocupar entre cuatro y seis páginas y contener los apartados siguientes:

- Problema clínico, decisión que se desea apoyar y usuario responsable.
- Población, criterios de inclusión, unidad de análisis, tiempo índice y horizonte.
- Desenlace con definición operativa y proceso de adjudicación.
- Predictores disponibles en el momento de uso y justificación biológica o clínica.
- Diagrama temporal que demuestre ausencia de fuga de información.
- Modelo de referencia sencillo y algoritmo alternativo.
- Plan de partición por animal, validación temporal y validación externa.
- Métricas de discriminación, calibración, utilidad y carga de alertas.
- Análisis de subgrupos, manejo de datos faltantes y regla de salida no evaluable.
- Plan de implementación prospectiva, vigilancia, consentimiento y protección de datos.

Criterios de evaluación: coherencia entre pregunta y diseño, claridad del tiempo índice, prevención de pseudorreplicación, pertinencia de métricas, interpretación clínica, viabilidad y análisis de riesgos. El protocolo no debe seleccionar un algoritmo antes de definir población, desenlace y decisión.

Conclusiones del capítulo

La aplicación clínica de inteligencia artificial en veterinaria debe juzgarse por la calidad de la evidencia y por sus efectos sobre pacientes, profesionales y tutores. Los resultados más útiles no siempre proceden del modelo más complejo. Una medición reproducible, una probabilidad bien calibrada o una alerta temprana con baja carga pueden aportar más que una clasificación opaca de alta exactitud aparente.

La investigación rigurosa conserva el método clínico: define la pregunta, recoge datos pertinentes, reconoce incertidumbre, compara alternativas y prueba la utilidad en condiciones reales. Ningún algoritmo corrige etiquetas ambiguas, selección sesgada o seguimiento incompleto. La IA puede ampliar la observación y apoyar decisiones; la responsabilidad de interpretar, comunicar y actuar permanece en el equipo veterinario.

CAPÍTULO 12

GANADERÍA DE PRECISIÓN PRODUCCIÓN REPRODUCCIÓN Y BIENESTAR ANIMAL

Introducción

La ganadería de precisión transforma mediciones dispersas en información útil para decisiones sobre salud, producción, reproducción, alimentación y bienestar. Su materia prima no es el algoritmo, sino una cadena de observación que comienza en el animal o en su ambiente: un acelerómetro registra movimiento, un medidor de leche cuantifica producción y conductividad, una cámara capta postura, y una estación meteorológica describe el microclima. Esos datos solo adquieren sentido cuando se vinculan con una pregunta veterinaria y con una acción definida.

El valor de estos sistemas depende de tres condiciones. La primera es metrológica: el sensor debe medir de manera estable aquello que declara medir. La segunda es biológica: la señal registrada debe guardar una relación reproducible con el estado fisiológico, productivo o conductual de interés. La tercera es operativa: la alerta debe llegar a tiempo, ser comprensible y conducir a una revisión o intervención proporcionada. Un dispositivo puede ser técnicamente preciso y, aun así, resultar inútil si genera demasiadas falsas alarmas o si el personal no dispone de un protocolo para responder.

Este capítulo estudia sistemas bovinos, porcinos, avícolas y acuícolas, con atención a las condiciones de producción de América Latina. El análisis separa monitorización, inferencia, predicción y control. También examina el riesgo de confundir eficiencia productiva con bienestar. Una vaca que mantiene la producción puede estar enferma; un lote uniforme no demuestra ausencia de dolor; un algoritmo que detecta inmovilidad no explica por sí solo si esta corresponde a descanso, enfermedad, restricción ambiental o miedo. La ganadería de precisión debe complementar la inspección clínica y el conocimiento del cuidador, no desplazarlo.

Objetivos de aprendizaje

1. Explicar la arquitectura de un sistema de ganadería de precisión desde la captura de datos hasta la decisión en campo.
2. Seleccionar sensores y variables de acuerdo con problemas de producción, reproducción, nutrición, salud y bienestar.
3. Distinguir detección, diagnóstico, predicción temprana y control automatizado, evitando atribuir causalidad a asociaciones predictivas.

4. Diseñar particiones de entrenamiento, validación y prueba que respeten animal, corral, lote, granja y tiempo.
5. Interpretar sensibilidad, especificidad, valor predictivo, calibración, anticipación de la alerta y carga de falsas alarmas en términos veterinarios.
6. Evaluar riesgos de sesgo, fallas de sensores, dependencia tecnológica, privacidad, ciberseguridad y deterioro de las habilidades de observación.
7. Proponer investigaciones reproducibles y actividades de validación externa en sistemas pecuarios reales.

Tabla 12.1

Cadena funcional de un sistema de ganadería de precisión

Etapa	Pregunta de control	Ejemplo veterinario	Falla frecuente
Adquisición	¿Qué se mide y con qué frecuencia?	Conductividad eléctrica por ordeño	Sensor sucio o descalibrado
Transmisión	¿Los datos llegan completos y a tiempo?	Collar que envía actividad por red inalámbrica	Pérdida de paquetes o zonas sin cobertura
Integración	¿La señal se asocia al animal correcto?	Unir ordeño, rumia y paridad por identificador	Identificadores duplicados o cambios de arete
Análisis	¿El método responde a la pregunta?	Probabilidad de mastitis en 24 horas	Fuga de información o sobreajuste
Entrega	¿Quién recibe la salida y cómo la interpreta?	Lista priorizada de vacas para revisión	Alerta sin explicación ni umbral de acción
Acción	¿Qué protocolo se activa?	Examen de ubre y prueba confirmatoria	Tratamiento automático sin verificación
Auditoría	¿El sistema conserva su rendimiento?	Recalibración trimestral por granja	Deriva no detectada del sensor o población

12.1 Fundamentos de la ganadería de precisión

La ganadería de precisión puede definirse como la gestión de animales y procesos productivos mediante mediciones continuas o repetidas, análisis computacional y retroalimentación orientada

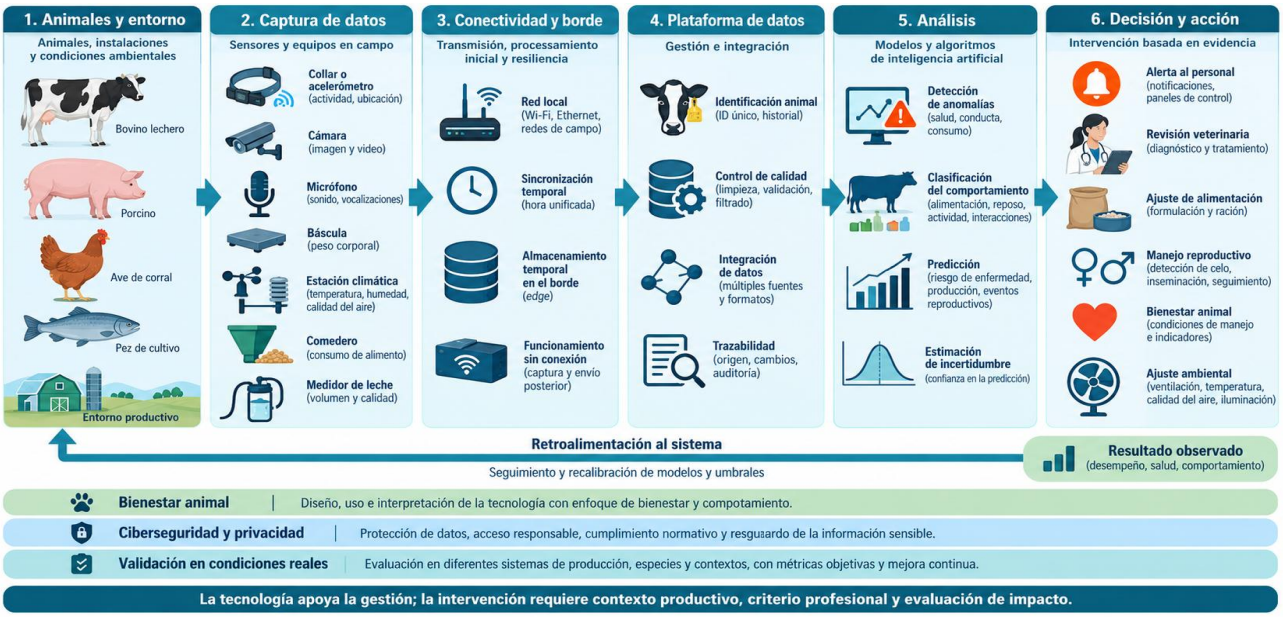
a decisiones. La unidad de gestión puede ser el individuo, el corral, el lote, el estanque o la granja. Berckmans (2017) planteó sus bases como una aplicación de la ingeniería de procesos a sistemas biológicos; trabajos recientes amplían el enfoque hacia integración de datos, aprendizaje automático y representación digital del animal (Curti et al., 2023; Kaur et al., 2023).

Un sistema completo opera como un ciclo. Primero observa; después compara la señal con una línea basal o un modelo; estima un estado o un riesgo; comunica una salida; registra la respuesta humana o automática; y utiliza el desenlace para evaluar si la decisión fue correcta. Esta retroalimentación diferencia un proyecto operativo de una simple colección de datos. La calidad del ciclo se deteriora si cualquiera de sus enlaces falla.

Conviene distinguir cuatro productos analíticos. La descripción resume lo ocurrido, por ejemplo, minutos diarios de rumia. La detección identifica un evento presente, como tos o cojera. La predicción estima un evento futuro, por ejemplo, mastitis clínica en las siguientes 24 horas. El control modifica el proceso, como ajustar ventilación o suministro de alimento. Cada producto necesita un estándar de referencia y un diseño de validación distinto.

Figura 12.1.

Ciclo operativo de la ganadería de precisión



La escala también cambia la interpretación. Un descenso de consumo en una vaca puede justificar examen individual; el mismo patrón simultáneo en un lote sugiere cambios de dieta, temperatura, acceso al comedero o falla del sistema de registro. Los modelos jerárquicos y los efectos aleatorios permiten separar parte de la variación atribuible al animal, al corral y a la granja. Un algoritmo entrenado sin reconocer esta estructura puede confundir identidad de granja con enfermedad.

La ganadería de precisión no convierte automáticamente datos observacionales en evidencia causal. Si las vacas de alta producción reciben más vigilancia y tratamientos, un modelo puede aprender que la intensidad de monitorización predice enfermedad. La asociación puede ser útil para anticipar casos, pero no demuestra que la monitorización cause la enfermedad. Las intervenciones deben evaluarse mediante diseños experimentales o cuasiexperimentales que comparen resultados clínicos, económicos y de bienestar.

12.2 Sensores dispositivos portátiles e internet de las cosas

Los sensores portátiles incluyen collares, aretes, podómetros, bolos intrarruminales y dispositivos fijados a extremidades o cola. Registran aceleración, orientación, localización, temperatura, presión o proximidad. Los sensores ambientales miden temperatura, humedad, concentración de gases, luminosidad, ruido, velocidad del aire y calidad del agua. Otros equipos están integrados al proceso: medidores de leche, balanzas, comederos electrónicos, bebederos instrumentados, cámaras y micrófonos.

La internet de las cosas conecta estos dispositivos con pasarelas, servidores y aplicaciones. La arquitectura puede procesar datos en el borde, cerca del sensor, o enviarlos a la nube. El procesamiento local reduce latencia y dependencia de conectividad, mientras el almacenamiento remoto facilita integración entre fuentes y actualización de modelos. La elección debe considerar cobertura, costo de transmisión, autonomía de batería, volumen de datos y necesidad de respuesta inmediata.

La frecuencia de muestreo se define por el fenómeno. Para localización de ganado extensivo puede bastar un registro cada varios minutos; para reconocer pasos, sacudidas o episodios breves se requieren frecuencias mayores. Aumentar la frecuencia no siempre mejora el sistema: eleva

consumo energético, almacenamiento y redundancia. El diseño debe justificar la resolución temporal mínima que conserve la señal biológica relevante.

La validación del sensor precede a la del algoritmo. Deben estimarse repetibilidad, concordancia con un método de referencia, proporción de datos perdidos, estabilidad entre dispositivos y tolerancia a polvo, humedad, golpes y cambios de posición. Para conducta, el estándar suele ser etología observacional codificada por evaluadores entrenados. Para temperatura, puede compararse con instrumentos clínicos bajo un protocolo explícito. Un sensor impreciso impone un techo al rendimiento del modelo.

Los dispositivos portátiles también pueden alterar la conducta o producir lesiones por presión, roce o peso. El protocolo debe incluir adaptación, revisión del ajuste, pérdida de dispositivos y retiro ante signos de daño. Curti et al. (2023) subrayan que adquisición, transferencia, almacenamiento, análisis y entrega forman una cadena inseparable. La ciberseguridad se incorpora a esa cadena: autenticación, control de acceso, respaldo, registro de cambios y un plan de operación manual si el sistema deja de funcionar.

Figura 12.2.

Arquitectura de sensores, borde y nube

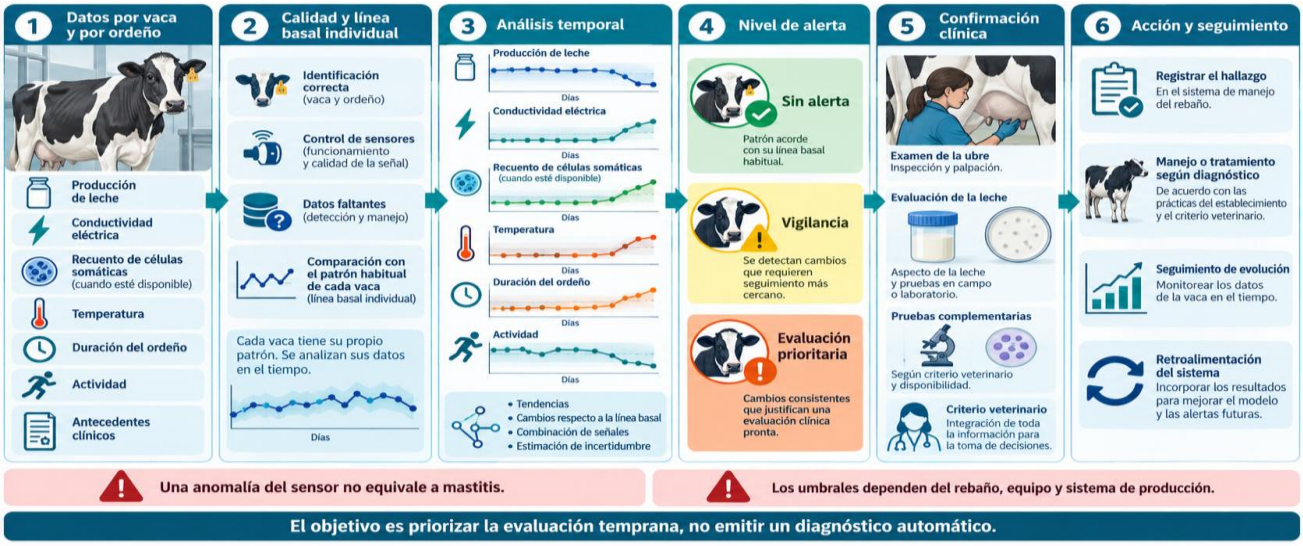


Tabla 12.2

Sensores frecuentes y criterios de validación

Tecnología	Variable primaria	Aplicaciones	Referencia de validación	Riesgo principal
Acelerómetro	Aceleración y orientación	Actividad, rumia, celo, parto, conducta	Observación directa sincronizada	Desplazamiento o etiqueta conductual ambigua
Bolo intrarruminal	Temperatura y movimiento	Fiebre, ingestión de agua, actividad	Temperatura de referencia y eventos registrados	Deriva, retención y efecto de consumo de agua
Medidor de leche	Volumen, flujo, conductividad, componentes	Producción, calidad y alertas de mastitis	Equipo calibrado y laboratorio	Ensuciamiento y diferencias entre equipos
Cámara RGB o profundidad	Imagen o nube de puntos	Identidad, postura, peso, conducta	Anotación ciega y medición física	Oclusión, luz, polvo y cambio de dominio
Cámara térmica	Temperatura superficial	Inflamación, estrés térmico y enfermedad	Temperatura clínica y condiciones ambientales	Emisividad, distancia y humedad
Micrófono	Señal acústica	Tos, vocalización y consumo	Registro humano sincronizado	Ruido de maquinaria y ecos
RFID o UWB	Identidad, proximidad o posición	Acceso, interacciones y uso del espacio	Mapa conocido y observación	Zonas muertas y colisiones de lectura
Sonda de agua	Oxígeno, pH, temperatura y turbidez	Acuicultura y control ambiental	Calibración química de laboratorio	Bioincrustación y deriva

12.3 Producción y calidad de leche

Los sistemas automáticos de ordeño generan series temporales de producción por vaca, duración, flujo, intervalo entre ordeños, ordeños incompletos y rechazos. Cuando se integran con conductividad eléctrica, color, temperatura, recuento celular o espectros infrarrojos, permiten estudiar simultáneamente cantidad y calidad. El patrón individual respecto a la línea basal suele aportar más que un umbral idéntico para todo el hato.

La producción de leche cambia con días en leche, número de parto, gestación, estación, dieta, enfermedad y frecuencia de ordeño. Por ello, una caída absoluta no debe interpretarse sin contexto. Un modelo estadístico puede estimar la curva esperada e identificar residuos anómalos; un modelo de aprendizaje automático puede integrar relaciones no lineales. Si el objetivo es explicar efectos de dieta o tratamiento, un modelo mixto conserva ventajas por su estructura inferencial y la posibilidad de modelar medidas repetidas.

La calidad incluye grasa, proteína, lactosa, sólidos, células somáticas, residuos, estabilidad y propiedades tecnológicas. Los analizadores en línea y la espectroscopia permiten predicciones rápidas, pero necesitan calibración contra laboratorio. La transferencia entre razas, estaciones, dietas y equipos no se presume. Una ecuación ajustada en vacas Holstein de alta producción puede mostrar sesgo en hatos cruzados o sistemas de pastoreo.

La investigación debe informar error absoluto y sesgo, no solo correlación. Dos métodos pueden correlacionarse fuertemente y diferir sistemáticamente. Para decisiones de pago o descarte de leche, el error cerca del punto de corte es más importante que el promedio global. La concordancia, los gráficos de diferencias y el análisis por intervalos de concentración complementan métricas como RMSE y R^2.

La utilidad se evalúa en el proceso. Una alerta sobre deterioro de calidad puede llevar a revisar refrigeración, rutina de ordeño, limpieza, salud de ubre o alimentación. Si el sistema señala al animal pero no diferencia una falla general del equipo, puede desviar la investigación. Los paneles deben mostrar simultáneamente tendencia individual, comparación de lote y estado del sensor.

12.4 Predicción de mastitis y enfermedades metabólicas

La mastitis ofrece un escenario clásico de alerta temprana porque altera producción, conductividad, temperatura, recuento de células somáticas, actividad y rumia. Esas señales no son específicas: el calor, el estrés, otras enfermedades y fallas de ordeño también las modifican. Un modelo debe tratar la predicción como priorización para examen y confirmación, no como sustituto de la evaluación clínica, bacteriología o pruebas de campo.

Tian et al. (2024) estudiaron algoritmos de aprendizaje automático basados en producción de leche, tiempo de rumia y conductividad eléctrica. Luo et al. (2023) desarrollaron un modelo de riesgo con información de vacas en lactación. Ambos trabajos ilustran el potencial de integrar variables rutinarias, pero no eliminan la necesidad de validación externa, definición temporal estricta del caso y evaluación del costo de las falsas alarmas.

Las enfermedades metabólicas del periparto plantean desafíos similares. Cetosis, hipocalcemia, desplazamiento de abomaso y metritis comparten factores de riesgo y signos tempranos. Pueden utilizarse cambios en ingestión, rumia, actividad, leche, condición corporal y antecedentes. El estándar de referencia debe incluir pruebas clínicas o de laboratorio tomadas sin conocimiento de la predicción, y el horizonte de interés debe fijarse antes del análisis.

La fuga temporal es un riesgo serio. Si una variable se registra después de que el personal sospecha la enfermedad o después del tratamiento, el modelo aprende consecuencias de la decisión, no señales tempranas. La ventana predictora debe terminar antes del momento de alerta y antes de cualquier acción clínica. Todas las observaciones de una vaca deben permanecer en la misma partición cuando se evalúa generalización a animales nuevos.

La prevalencia modifica el valor predictivo positivo. Incluso un modelo con buena sensibilidad y especificidad puede producir numerosas falsas alarmas cuando la enfermedad es poco frecuente. Por ello deben informarse precisión, valor predictivo negativo, alertas por 100 vacas-día, tiempo de anticipación, calibración y beneficio neto. El umbral se selecciona según capacidad de revisión y consecuencias de omitir un caso.

12.5 Reproducción y detección automatizada del celo

La detección del celo combina cambios en locomoción, tiempo de reposo, monta, proximidad, rumia, ingestión y temperatura. Acelerómetros, podómetros, cámaras y sistemas de localización permiten vigilancia continua. La señal varía con piso, clima, jerarquía social, salud podal, producción, paridad y manejo; en vacas con baja expresión conductual, la sensibilidad puede disminuir.

El evento de referencia debe definirse. La observación de conducta, la concentración de progesterona, la ecografía ovárica y la ovulación describen momentos relacionados pero no idénticos. Un estudio que etiqueta celo según inseminación registrada incorpora decisiones humanas y puede perder eventos no detectados. La etiqueta adecuada depende de si el objetivo es identificar conducta estral, estimar ovulación o recomendar el momento de inseminación.

La evaluación debe considerar episodio y animal. Dividir segmentos del mismo celo entre entrenamiento y prueba infla el rendimiento. Además de sensibilidad por episodio, conviene informar falsas alertas por vaca y por periodo, distribución del tiempo entre alerta y ovulación, y resultados reproductivos posteriores. La tasa de concepción no puede atribuirse únicamente al algoritmo porque depende de semen, inseminador, momento, salud uterina y fertilidad del animal.

En porcinos, los sistemas pueden observar inmovilidad ante presión dorsal, interacción con el verraco y actividad. En pequeños rumiantes, los collares y marcadores de monta pueden complementarse con manejo estacional. Curti et al. (2023) resumen aplicaciones de sensores y aprendizaje automático en reproducción, pero también señalan problemas de transferencia y aceptación. Un sistema debe configurarse con las condiciones locales antes de guiar decisiones.

La automatización reproductiva no debe reducir la vigilancia de problemas clínicos. Una disminución de actividad puede ocultar cojera; un supuesto anestro puede corresponder a enfermedad uterina, balance energético negativo o falla del dispositivo. La salida más segura es una lista priorizada con evidencia temporal y nivel de incertidumbre.

12.6 Nutrición y alimentación de precisión

La alimentación de precisión busca aproximar el aporte de nutrientes a los requerimientos de animales o grupos en periodos cortos. Integra peso, crecimiento, producción, consumo, composición de dieta y ambiente. El objetivo puede ser mejorar conversión, reducir desperdicio, limitar excreción de nitrógeno y fósforo o sostener salud. La optimización económica debe mantenerse dentro de restricciones biológicas y de bienestar.

Los comederos electrónicos registran identidad, hora, duración y cantidad. En bovinos lecheros, los sistemas pueden dosificar suplementos según producción o etapa de lactancia. En cerdos, la

mezcla dinámica de dietas permite ajustar lisina y otros nutrientes a curvas de crecimiento. Los registros de consumo presentan censura y errores cuando varios animales acceden, el alimento se derrama o cambia la calibración.

La predicción de consumo voluntario es difícil porque responde a clima, salud, jerarquía, palatabilidad y cambios de dieta. Los modelos deben incorporar retardos y diferenciar variación normal de una caída clínicamente relevante. Un GAM o un modelo mixto puede representar curvas no lineales y repetición por animal; un bosque aleatorio puede ser útil cuando existen interacciones complejas, siempre que la validación respete granja y tiempo.

Un sistema que minimiza costo sin restricciones puede formular dietas inadecuadas o generar cambios excesivos. Las reglas deben fijar mínimos y máximos, estabilidad de la ración, disponibilidad real de ingredientes y límites de transición. La recomendación algorítmica debe ser revisada por un profesional competente y quedar registrada.

La evaluación no termina en la ganancia de peso. Debe incluir variación entre animales, salud digestiva, lesiones, comportamiento alimentario, mortalidad, calidad del producto, excreción y costo por unidad producida. Una mejora promedio puede coexistir con una cola de animales perjudicados; por ello se analizan cuantiles y subgrupos, no solo la media.

12.7 Crecimiento y eficiencia productiva

El crecimiento puede medirse con balanzas de paso, plataformas, cámaras RGB, profundidad y reconstrucción tridimensional. Estas tecnologías reducen manejo y permiten series frecuentes. Las estimaciones visuales usan longitud, altura, perímetro y volumen corporal; su rendimiento depende de pose, suciedad, iluminación, cobertura del cuerpo y representatividad del conjunto de entrenamiento.

La eficiencia alimenticia combina consumo y ganancia. El índice de conversión es intuitivo, pero puede confundir diferencias de mantenimiento y composición corporal. El consumo residual de alimento ajusta consumo esperado según producción y tamaño, aunque su interpretación depende del modelo. Para selección o manejo se requieren mediciones confiables y un periodo suficiente que represente la variabilidad diaria.

Los algoritmos pueden predecir peso futuro, edad a sacrificio o riesgo de retraso. La partición debe hacerse por animal y, si se pretende uso en otras granjas, reservar granjas completas. El error se expresa en kilogramos y porcentaje, junto con sesgo por rango de peso, sexo, raza y sistema. Un RMSE aceptable para clasificación por grupos puede ser insuficiente para dosificar medicamentos.

La mejora productiva no equivale a eficiencia ambiental o bienestar. Animales de crecimiento rápido pueden presentar problemas locomotores o cardiometabólicos en algunos sistemas. La función objetivo debe incluir restricciones de salud y bienestar. Los análisis de frontera y optimización multiobjetivo hacen visible el intercambio entre ganancia, consumo, emisiones, mortalidad y calidad.

En investigación, la unidad experimental depende de la intervención. Si una dieta se asigna al corral y todos los animales comparten comedero, el corral es la unidad experimental aunque el peso sea individual. Tratar cada animal como réplica produce pseudorreplicación. Los sensores aumentan el número de observaciones, no el número de unidades independientes.

12.8 Avicultura y porcicultura inteligentes

La avicultura de precisión monitoriza lotes numerosos mediante cámaras, micrófonos, sensores ambientales, balanzas y sistemas de alimentación. Las variables incluyen distribución espacial, actividad, marcha, vocalización, tos, consumo, peso, temperatura y humedad. La densidad y la oclusión dificultan el seguimiento individual, por lo que muchas aplicaciones trabajan con indicadores de grupo.

Ojo et al. (2022) revisaron aplicaciones de internet de las cosas y aprendizaje automático en salud y bienestar avícola. Yang et al. (2024) analizaron sistemas de visión computacional para producción y bienestar, destacando dificultades de polvo, iluminación, densidad y necesidad de datos anotados. Esas limitaciones explican por qué un modelo preciso en videos experimentales puede degradarse en galpones comerciales.

En porcicultura, cámaras superiores y laterales permiten estimar peso, postura, conducta de alimentación, agresión, aplastamiento y parto. Micrófonos pueden detectar tos; cámaras térmicas exploran cambios superficiales; RFID vincula consumo y peso al individuo. Gómez et al. (2021)

encontraron que una proporción reducida de tecnologías comerciales revisadas contaba con validación externa, una advertencia contra asumir equivalencia entre disponibilidad comercial y evidencia.

La bioseguridad favorece sensores sin contacto, pero la infraestructura también puede introducir sesgos. Una cámara observa mejor la zona central que los rincones; un micrófono cercano a ventiladores registra un perfil distinto; el cambio de cama altera el fondo. El protocolo debe documentar ubicación, mantenimiento y condiciones ambientales, y la validación debe abarcar ciclos productivos completos.

La salida debe diseñarse para la escala de intervención. En aves, una alerta de distribución anómala puede activar revisión de temperatura, ventilación o agua. En cerdos, un aumento de agresiones puede llevar a revisar mezcla de grupos, espacio y acceso a recursos. La clasificación automatizada es el inicio de una investigación causal, no la explicación final.

12.9 Acuicultura de precisión

La acuicultura de precisión integra sensores de oxígeno disuelto, temperatura, pH, salinidad, turbidez, amonio y flujo con cámaras, sonar, hidrófonos y sistemas automáticos de alimentación. El ambiente acuático cambia con rapidez y afecta simultáneamente metabolismo, apetito, enfermedad y comportamiento. La detección debe considerar estas dependencias antes de atribuir una señal al animal.

La visión computacional puede estimar biomasa, contar peces, seguir trayectorias, evaluar distribución y reconocer respuestas alimentarias. Yang et al. (2021) describieron desafíos persistentes: oclusión, deformación corporal, variación de escala, reflejos, turbidez y bioincrustación de lentes. El sonar y la acústica ofrecen alternativas cuando la visibilidad es limitada, pero requieren calibración y experiencia de interpretación.

La alimentación automática busca detener o ajustar el suministro cuando disminuye la respuesta. Un exceso aumenta costo y deteriora el agua; una restricción sostenida compromete crecimiento y bienestar. El control debe incorporar especie, talla, temperatura, oxígeno y hora. Los ensayos

comparativos deben medir alimento no consumido, crecimiento, uniformidad, conversión, mortalidad, calidad de agua y lesiones.

Las redes IoT permiten alertas tempranas, pero la bioincrustación y deriva de sondas son amenazas centrales. Deben definirse calendarios de limpieza y calibración, redundancia para variables críticas y reglas de plausibilidad. Un valor de oxígeno incompatible con sensores vecinos o con la dinámica esperada debe activar comprobación del instrumento antes de una intervención brusca.

Antonucci y Costa (2020) sintetizaron innovaciones de ingeniería que continúan siendo base del área. Aunque esta referencia antecede al periodo prioritario, se conserva por su carácter fundacional. La investigación reciente debe avanzar hacia validación comercial, análisis de bienestar y evaluación energética del propio sistema de monitorización.

12.10 Evaluación automatizada del comportamiento

El comportamiento es una variable temporal. Reconocer una postura aislada no equivale a reconocer una secuencia con inicio, duración y contexto. Los proyectos deben construir un etograma operacional: cada categoría necesita una definición observable, reglas de inclusión y exclusión, duración mínima y tratamiento de conductas simultáneas.

Las etiquetas provienen de observadores y contienen error. Se recomienda entrenamiento previo, muestreo de videos, evaluación de concordancia y adjudicación de discrepancias. Para categorías nominales puede usarse kappa; para duraciones o proporciones, coeficientes de concordancia o modelos adecuados. La concordancia alta no garantiza validez si ambos observadores aplican una definición deficiente.

Los acelerómetros clasifican ventanas de señal mediante características de tiempo y frecuencia o redes profundas. Las cámaras añaden postura, interacción y espacio. Los métodos multifuente pueden mejorar discriminación, pero aumentan la complejidad y la probabilidad de datos faltantes. Se debe comparar la fusión con modelos más simples para demostrar ganancia real.

La división aleatoria por fotogramas casi siempre es incorrecta porque cuadros adyacentes son casi idénticos. La prueba debe reservar animales, episodios, días o granjas según el uso previsto. En

conductas raras, la exactitud global oculta fallas; se informan sensibilidad, precisión y F1 por clase, matriz de confusión y métricas macro.

Hao et al. (2023) propusieron una red de dos flujos para reconocer conducta porcina, combinando información espacial y temporal. Este tipo de arquitectura puede captar movimiento, pero su utilidad práctica depende de generalización, costo computacional y claridad del repertorio conductual. Un modelo que distingue categorías visuales no reemplaza la interpretación etológica.

12.11 Detección de dolor estrés y alteraciones del bienestar

El bienestar comprende estado físico, comportamiento, afectos y posibilidad de realizar conductas relevantes. Ningún sensor aislado resume esa condición. La inmovilidad puede reflejar descanso o dolor; una vocalización puede corresponder a hambre, separación o interacción social; el aumento de temperatura superficial depende del ambiente. La evaluación automatizada necesita indicadores convergentes y examen del contexto.

Los sistemas pueden vigilar cojera, lesiones, tos, cambios de consumo, alteración de postura, uso del espacio, termorregulación, agresión y conductas estereotipadas. La detección debe conectarse con protocolos de inspección y respuesta. Una alerta de dolor no autoriza tratamiento sin comprobar identidad, causa, historia y contraindicaciones.

Gómez et al. (2021) mostraron que la validación externa de tecnologías para bienestar porcino era escasa. Tuyttens et al. (2022) describieron amenazas directas e indirectas: fallas técnicas, predicciones inexactas, indicadores poco significativos, dependencia del sistema y pérdida de habilidades de cuidado. Schillings et al. (2021) añadieron cuestiones de autonomía, privacidad, responsabilidad y posible intensificación.

La evaluación debe incluir falsos negativos porque omitir sufrimiento tiene un costo ético. También importan falsos positivos: revisiones repetidas pueden aumentar manejo y estrés. La selección del umbral debe considerar ambas consecuencias y la capacidad real de respuesta. Conviene registrar el tiempo desde la alerta hasta la inspección y desde la inspección hasta la intervención.

La tecnología puede mejorar el bienestar si amplía la observación y favorece respuestas tempranas. Puede empeorarlo si legitima densidades o escalas que reducen el contacto de calidad, si adapta el alojamiento a la cámara en lugar del animal o si convierte indicadores productivos en sustitutos de bienestar. Papakonstantinou et al. (2024) proponen considerar conjuntamente aplicaciones, cambio climático y bienestar; esta integración debe ser explícita en objetivos y auditorías.

12.12 Visión computacional para identificación y seguimiento animal

La identificación puede usar patrones de pelaje, cara, hocico, iris, conformación o marcas artificiales. El seguimiento asocia detecciones del mismo individuo a través de fotogramas. La tarea se complica por oclusión, individuos visualmente similares, suciedad, crecimiento y salidas del campo de visión. RFID puede aportar identidad complementaria y reducir errores de asociación.

El desarrollo incluye detección, segmentación, extracción de características, reidentificación y seguimiento. Las métricas cambian según la tarea: precisión media para detección, intersección sobre unión para segmentación y medidas de identidad para seguimiento. En aplicaciones de bienestar, un intercambio de identidades puede asignar una alerta al animal equivocado, de modo que la exactitud de identidad debe evaluarse de manera separada.

Las cámaras de profundidad permiten estimar volumen, postura y distancia con menor dependencia del color, pero pueden verse afectadas por reflectancia, polvo o rango. La termografía aporta patrones térmicos sin contacto, condicionados por emisividad, ángulo y clima. La fusión debe justificarse por una mejora reproducible y no por la disponibilidad de más sensores.

Los conjuntos de datos deben documentar especie, raza, edad, sexo, granja, estación, equipo, iluminación, densidad y protocolo de anotación. La diversidad no se resuelve acumulando imágenes del mismo lugar. La validez externa requiere centros y condiciones nuevas. Los modelos deben incluir una salida de baja confianza o rechazo cuando la imagen está fuera del dominio conocido.

La videovigilancia también captura trabajadores y procesos. Se necesitan reglas de acceso, retención, finalidad y comunicación al personal. Los datos no deben reutilizarse para vigilancia

laboral o fines distintos sin una base ética y normativa. La gobernanza forma parte del diseño, no es una tarea posterior.

12.13 Impacto ambiental y sostenibilidad

La sostenibilidad debe evaluarse por unidad funcional y dentro de límites definidos. Emisiones por animal, por kilogramo de leche o carne y por hectárea responden preguntas diferentes. Un aumento de producción puede reducir intensidad por unidad de producto y aumentar emisiones totales. Las conclusiones deben informar ambos niveles cuando sean relevantes.

La ganadería de precisión puede mejorar ajuste de dieta, eficiencia alimenticia, uso de agua, manejo de estiércol, detección de fugas y adaptación al estrés térmico. Sensores de consumo, peso y clima facilitan acciones específicas. Sin embargo, el efecto ambiental no se presume: debe medirse mediante inventarios, análisis de ciclo de vida o ensayos que comparen el sistema con una referencia.

Las tecnologías también consumen energía, baterías, equipos y servicios de datos. Su huella suele ser menor que la de procesos pecuarios principales, pero debe incluirse cuando se comparan alternativas. La obsolescencia, interoperabilidad limitada y residuos electrónicos pueden aumentar costos ambientales y económicos.

La optimización multiobjetivo permite explorar dietas o manejos que equilibren margen, emisiones, producción y bienestar. No existe una solución única cuando los objetivos entran en conflicto. Las restricciones deben ser acordadas antes de optimizar; de otro modo, el algoritmo privilegiará la variable que domine la función de pérdida.

El IPCC (2023) destaca la necesidad de reducciones amplias de emisiones y adaptación. En este marco, los datos de precisión pueden apoyar medición y seguimiento, pero no sustituyen cambios de manejo validados. La transparencia exige documentar supuestos, calidad de los factores de emisión y grado de incertidumbre.

Tabla 12.3

Métricas y significado operativo

Métrica	Qué responde	Interpretación en campo	Precaución
Sensibilidad	¿Qué proporción de eventos reales detecta?	Casos de mastitis o celo identificados	Depende de la definición y oportunidad del evento
Especificidad	¿Qué proporción de no eventos descarta?	Animales sanos sin alerta	Puede ocultar muchas alertas si la prevalencia es baja
Valor predictivo positivo	¿Qué proporción de alertas se confirma?	Carga de revisiones útiles	Cambia con prevalencia y selección de población
F1 por clase	¿Equilibra precisión y sensibilidad?	Comparación de conductas raras	No incorpora calibración ni costo clínico
AUROC	¿Ordena casos por riesgo?	Discriminación general	Puede parecer alta con fuerte desbalance
AUPRC	¿Rinde en la clase positiva?	Más informativa para eventos raros	Su línea basal depende de prevalencia
Calibración	¿La probabilidad coincide con frecuencia?	Una predicción de 0,20 ocurre cerca de 20 %	Debe revisarse en cada granja y periodo
Anticipación	¿Con cuánto tiempo avisa?	Horas disponibles para examinar	Una alerta tardía puede ser exacta pero inútil
Alertas por 100 animales día	¿Cuál es la carga operativa?	Trabajo esperado del personal	Requiere periodo de observación comparable

12.14 Caso didáctico alerta temprana de mastitis bovina

Naturaleza del caso. Caso didáctico simulado. Los valores, tamaños y resultados que siguen se construyen para enseñar el proceso metodológico; no corresponden a un hato real ni deben utilizarse como umbrales clínicos.

Problema científico. Un sistema lechero dispone de medidores automáticos y collares, pero la revisión de todas las señales consume tiempo. Se desea priorizar vacas con riesgo de presentar mastitis clínica en las 24 horas siguientes, sin iniciar tratamiento automático.

Pregunta e hipótesis. ¿La combinación de cambios respecto a la línea basal individual en conductividad, producción y rumia mejora la alerta frente a una regla basada únicamente en

conductividad? Se plantea que un modelo multivariable temporal tendrá mayor sensibilidad a una carga de falsas alarmas comparable.

Población y unidades. Se simulan 1 200 lactaciones pertenecientes a 820 vacas de cuatro granjas, seguidas durante 18 meses. La unidad biológica es la vaca; la unidad de predicción es vaca-día; la granja constituye un nivel de agrupación. Los episodios separados por menos de 14 días se consideran parte del mismo proceso clínico.

Desenlace. El evento positivo es un diagnóstico clínico realizado por personal veterinario mediante signos de inflamación de ubre o alteración visible de leche, con registro de fecha y hora. El modelo utiliza exclusivamente información terminada 24 horas antes del diagnóstico. Las vacas tratadas preventivamente antes del desenlace se analizan según un protocolo preespecificado, porque el tratamiento modifica la observación del evento.

Tabla 12.4

Variables del caso simulado de mastitis

Dominio	Variable	Construcción temporal	Razonamiento
Ordeño	Conductividad	Cambio porcentual frente a mediana de 14 días	Captar alteración individual y reducir diferencias entre equipos
Producción	Leche diaria	Pendiente de 3 días y desviación de línea basal	Detectar descenso progresivo
Conducta	Rumia	Minutos diarios y cambio de 48 horas	Representar alteración sistémica temprana
Conducta	Actividad	Índice diario estandarizado por vaca	Distinguir cambio respecto al patrón habitual
Animal	Días en leche	Valor al día de predicción	Controlar etapa de lactación
Animal	Paridad	Primípara o multípara	Representar diferencias de riesgo
Historia	Mastitis previa	Evento anterior a la ventana	Incorporar susceptibilidad sin fuga temporal
Ambiente	Índice térmico	Máximo de 24 horas	Ajustar cambios conductuales por calor

Diseño del análisis. Las granjas 1 a 3 se usan para desarrollo y la granja 4 se reserva como prueba externa. Dentro del desarrollo se emplea validación cruzada agrupada por vaca y bloqueada por tiempo. Se comparan regresión logística penalizada, random forest y gradient boosting con una regla simple de conductividad. El ajuste de hiperparámetros se realiza solo dentro de los datos de desarrollo. La imputación, el escalamiento y la selección de variables se estiman nuevamente en cada partición.

Desbalance. La prevalencia simulada de vaca-día positiva es 1,8 %. No se aplica sobremuestreo antes de dividir. Si se usa ponderación de clases, esta se ajusta dentro de cada pliegue. La evaluación principal se mantiene en la prevalencia natural para que valores predictivos y carga de alertas sean interpretables.

Resultados simulados. En la granja externa, el modelo de gradient boosting alcanza sensibilidad de 0,78, especificidad de 0,91, valor predictivo positivo de 0,14 y valor predictivo negativo de 0,996 al umbral operativo. Produce 9,9 alertas por 100 vacas-día. La regla de conductividad alcanza sensibilidad de 0,56 y 8,7 alertas por 100 vacas-día. Estos números muestran que una mejora de sensibilidad puede coexistir con un valor predictivo positivo modesto cuando el evento es raro.

Interpretación. De 100 alertas del modelo, unas 14 se confirmarían como episodios según la simulación. Esto no invalida la herramienta si la revisión es barata, no invasiva y permite recuperar casos adicionales; sí impide usar la alerta como diagnóstico o indicación de antibiótico. El beneficio debe compararse con tiempo de inspección, cultivo, pérdidas evitadas y riesgo de sobretratamiento.

Calibración y umbral. Se examina gráfico de calibración por granja y paridad. Si el modelo sobreestima riesgo, se recalibra el intercepto o se utiliza calibración isotónica con datos locales separados. El umbral se elige con una restricción de máximo diez alertas por 100 vacas-día y una sensibilidad mínima acordada. No se selecciona por exactitud global.

Implementación. Cada mañana se genera una lista con riesgo, tendencia de conductividad, cambio de producción, rumia y estado del sensor. El personal confirma identidad, revisa la ubre, examina

leche y aplica pruebas establecidas. La decisión clínica permanece en el equipo veterinario. Se registra causa de falsa alarma para mejorar el sistema.

Limitaciones. El estándar clínico puede omitir cuadros leves; las granjas difieren en protocolos; la conectividad genera faltantes; y los datos simulados no representan distribuciones reales. Antes de una implementación se necesita estudio prospectivo silencioso, validación de impacto y vigilancia de deriva.

12.15 Caso didáctico clasificación del comportamiento en porcinos

Naturaleza del caso. Caso didáctico simulado. Su propósito es mostrar cómo diseñar y validar un clasificador de video sin confundir fotogramas con réplicas independientes.

Problema y objetivo. Una granja busca cuantificar descanso, alimentación, exploración, interacción social y agresión para estudiar bienestar después de una mezcla de grupos. El objetivo es clasificar ventanas de video de cinco segundos y producir presupuestos conductuales por corral y hora.

Población y muestreo. Se simulan 24 corrales de 12 cerdos, registrados durante seis semanas con cámaras cenitales. Doce corrales pertenecen a una nave y doce a otra. Se seleccionan días, horas y zonas de manera estratificada para evitar que el conjunto esté dominado por descanso nocturno. El corral es la unidad experimental para comparar manejos; el episodio es la unidad de clasificación.

Etograma y referencia. Dos observadores ciegos anotan 3 600 ventanas. Cada conducta cuenta con definición operacional y regla de prioridad cuando dos eventos coexisten. Se calcula kappa por clase; las discrepancias se resuelven por consenso. Las ventanas borrosas, ocluidas o sin visibilidad suficiente reciben la etiqueta no evaluable y se conservan para analizar la tasa de rechazo.

Tabla 12.5

Etograma operacional del caso porcino simulado

Clase	Definición operacional	Confusión previsible	Acción analítica

Descanso	Cuerpo apoyado sin locomoción durante la mayor parte de la ventana	Inmovilidad por enfermedad	Combinar duración y contexto; revisión de episodios prolongados
Alimentación	Cabeza dentro del comedero con permanencia compatible con consumo	Exploración del comedero	Validar con registro del comedero cuando exista
Exploración	Hociqueo dirigido a piso, objetos o estructura sin alimentación	Juego o búsqueda de alimento	Definición y ejemplos para anotadores
Interacción social	Contacto dirigido no agresivo con otro cerdo	Inicio de agresión	Usar secuencia temporal y postura
Agresión	Empuje, mordida, golpe o persecución con respuesta del receptor	Juego vigoroso	Priorizar sensibilidad y revisión humana
No evaluable	Oclusión o imagen insuficiente para una etiqueta confiable	Forzar una categoría visible	Permitir rechazo y medir su frecuencia

Partición y modelos. Se reservan seis corrales completos de la segunda nave para prueba externa. Los demás corrales se dividen por corral y semana; ningún fotograma adyacente cruza particiones. Se compara un clasificador de características visuales con una red que integra flujo temporal. El aumento de datos se aplica solo al entrenamiento.

Resultados simulados. La exactitud externa es 0,84 y el F1 macro 0,71. El F1 de descanso es 0,93, mientras el de agresión es 0,52 debido a pocos episodios y confusión con interacción social. La tasa de ventanas no evaluables es 8 %. La exactitud por fotograma, calculada incorrectamente con división aleatoria, habría sido 0,96; la diferencia ilustra la fuga por similitud temporal.

Interpretación y uso. El sistema es adecuado para estimar tiempo de descanso y alimentación a escala de corral, pero todavía no para intervenir automáticamente ante agresión. Los episodios con alta probabilidad de agresión se envían a revisión. Las comparaciones entre tratamientos usan proporciones diarias agregadas por corral y un modelo mixto o análisis compatible con la asignación experimental.

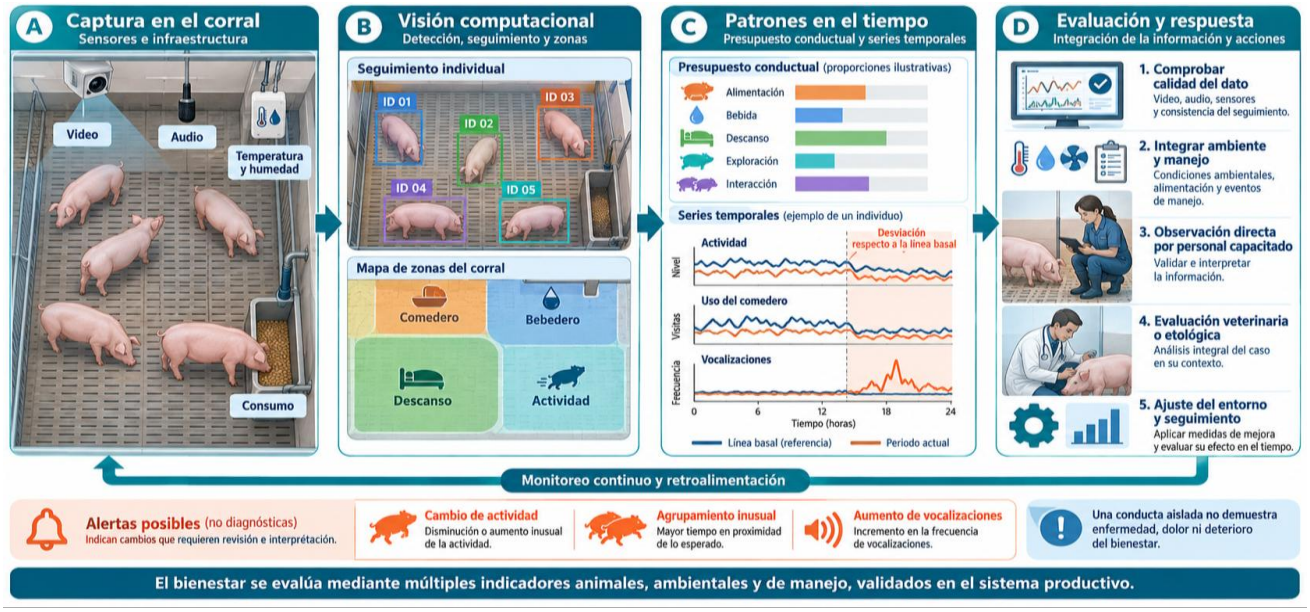
Validación de impacto. Un ensayo posterior puede comparar manejo habitual frente a alertas revisadas por cuidadores. Los desenlaces incluirían lesiones, intervenciones, mortalidad, variación

de peso, tiempo de revisión y falsos avisos. La hipótesis de beneficio requiere este ensayo; el rendimiento del clasificador por sí solo no demuestra mejora del bienestar.

Limitaciones. Los cambios de cámara, densidad, iluminación y color del piso pueden reducir rendimiento. La clasificación no captura estados afectivos directamente. Además, la cámara puede observar trabajadores, por lo que el protocolo debe limitar acceso y retención.

Figura 12.3.

Ejemplos de alertas en producción animal



12.16 Síntesis autoevaluación y actividad práctica

Síntesis

La ganadería de precisión une biología, veterinaria, ingeniería y análisis de datos. Los sistemas útiles parten de una decisión concreta y conservan trazabilidad entre señal, modelo, alerta, revisión y desenlace. La calidad del sensor, la definición del evento, el respeto por la estructura jerárquica y la validación externa importan tanto como el algoritmo.

En producción y reproducción, la mayor frecuencia de medición permite observar curvas y cambios individuales, pero también aumenta autocorrelación y riesgo de pseudorreplicación. En

bienestar, los indicadores deben representar al animal y no limitarse a rendimiento. Una alerta responsable prioriza observación, informa incertidumbre y admite no saber cuando la señal es insuficiente.

La implementación exige una prueba prospectiva, protocolos de respuesta, evaluación económica, gobernanza de datos y monitorización de deriva. Un sistema puede mostrar buena discriminación y fracasar por mala calibración, exceso de alertas o falta de integración con el trabajo. La evidencia final es la mejora demostrada de decisiones y resultados, no la complejidad del modelo.

Preguntas de autoevaluación

1. Explique la diferencia entre monitorización, detección, predicción y control con un ejemplo pecuario para cada función.
2. ¿Por qué la validación de un sensor debe preceder a la validación del algoritmo que utiliza sus datos?
3. Una alerta de mastitis tiene sensibilidad de 85 % y valor predictivo positivo de 12 %. ¿Qué significa cada cifra para el trabajo diario?
4. ¿Qué fuga de información se produce si los días de una misma vaca aparecen en entrenamiento y prueba?
5. Describa tres factores que pueden reducir la transferencia de un detector de celo entre granjas.
6. ¿Cuándo sería preferible un modelo mixto frente a un algoritmo de boosting en un estudio de producción de leche?
7. ¿Por qué el corral puede ser la unidad experimental aunque se registren cientos de animales?
8. Proponga un estándar de referencia para validar alimentación, agresión y cojera.
9. Explique por qué exactitud global es insuficiente para una conducta rara como agresión.
10. ¿Qué riesgos de bienestar pueden surgir por dependencia excesiva de sistemas automatizados?
11. ¿Cómo evaluaría si una tecnología reduce realmente la huella ambiental y no solo la intensidad por unidad de producto?
12. Diseñe un protocolo de respuesta ante una alerta de caída de consumo en cerdos.

Actividad práctica

Título. Diseño y evaluación de una alerta de riesgo con datos pecuarios longitudinales.

Propósito. Construir un flujo reproducible que respete la estructura por animal y traduzca las métricas a una decisión veterinaria. Puede usarse una base institucional anonimizada o un conjunto simulado.

Datos mínimos. Cada fila corresponde a un animal-día e incluye identificador de granja, animal, fecha, producción o consumo, actividad, una variable ambiental y un desenlace binario que ocurra al día siguiente. Si se usan datos reales, deben aplicarse autorización, minimización y anonimización.

1. Defina la pregunta, el horizonte de predicción, el evento, la unidad de análisis y la acción que seguiría a la alerta.
2. Elabore un diccionario con unidad, rango plausible, frecuencia, origen y mecanismo probable de faltantes.
3. Construya variables usando solo información disponible antes del momento de predicción.
4. Reserve una granja o un bloque temporal para prueba. Dentro del desarrollo, agrupe la validación por animal.
5. Compare una regresión logística con un árbol o bosque aleatorio. Ajuste preprocesamiento únicamente en entrenamiento.
6. Informe matriz de confusión, sensibilidad, especificidad, valor predictivo positivo, F1, AUPRC, calibración y alertas por 100 animales-día.
7. Seleccione un umbral a partir de una restricción operativa. Justifique el costo relativo de falsa alarma y evento omitido.
8. Redacte un protocolo de respuesta y un plan de monitorización de deriva.

Código orientativo en R. El ejemplo asume que la variable grupo_prueba identifica una granja reservada. Debe adaptarse al nombre y estructura de la base. La partición por animal se mantiene para evitar que la historia del mismo individuo aparezca en desarrollo y evaluación interna.

```
library(tidymodels)
library(dplyr)

desarrollo <- datos |> filter(grupo_prueba == 0)
prueba_ext <- datos |> filter(grupo_prueba == 1)

set.seed(1204)
folds <- group_vfold_cv(desarrollo, group = id_animal, v = 5)

receta <- recipe(evento_manana ~ produccion + actividad + ambiente + paridad,
                 data = desarrollo) |>
  step_impute_median(all_numeric_predictors()) |>
  step_normalize(all_numeric_predictors())

modelo <- logistic_reg(penalty = tune(), mixture = 1) |>
  set_engine("glmnet") |>
  set_mode("classification")

flujo <- workflow() |> add_recipe(receta) |> add_model(modelo)
rejilla <- grid_regular(penalty(range = c(-5, 0)), levels = 20)

ajuste <- tune_grid(flujo, resamples = folds, grid = rejilla,
                    metrics = metric_set(roc_auc, pr_auc, sens, spec))

mejor <- select_best(ajuste, metric = "pr_auc")
final <- finalize_workflow(flujo, mejor) |> fit(data = desarrollo)
pred <- bind_cols(prueba_ext,
                  predict(final, prueba_ext, type = "prob"),
                  predict(final, prueba_ext, type = "class"))
```

Producto esperado. Informe de cuatro a seis páginas con diagrama del flujo, diccionario abreviado, justificación de la partición, tabla de métricas, gráfico de calibración, análisis de errores y protocolo de respuesta. Debe diferenciar hallazgos predictivos de afirmaciones causales.

CAPÍTULO 13

EPIDEMIOLOGÍA SALUD PÚBLICA VETERINARIA Y ENFOQUE ONE HEALTH

Introducción

La epidemiología veterinaria estudia la frecuencia, distribución y determinantes de los eventos relacionados con la salud en poblaciones animales. Su campo de acción abarca desde un brote en una granja hasta la circulación transfronteriza de patógenos, la contaminación de alimentos y la transferencia de genes de resistencia entre animales, seres humanos y ambientes. La inteligencia artificial amplía la capacidad para explorar estos sistemas porque permite combinar registros clínicos, notificaciones sanitarias, movimientos de animales, secuencias genómicas, imágenes satelitales y variables meteorológicas. Ese potencial no elimina las reglas epidemiológicas. Una predicción precisa puede carecer de validez si el muestreo es sesgado, la definición de caso cambia entre territorios o la validación mezcla información del mismo brote en entrenamiento y prueba.

El enfoque One Health ofrece el marco para integrar esas fuentes. La definición propuesta por el One Health High-Level Expert Panel lo concibe como un enfoque integrado y unificador que busca equilibrar y optimizar de manera sostenible la salud de las personas, los animales y los ecosistemas (Adisasmito et al., 2022). La integración, por tanto, no consiste en reunir bases de datos bajo una misma etiqueta. Exige preguntas compartidas, definiciones compatibles, gobernanza, interpretación conjunta y mecanismos que conviertan la evidencia en vigilancia, prevención y respuesta. El Plan de Acción Conjunto One Health 2022-2026 organizó esta agenda en torno a capacidades, zoonosis, enfermedades desatendidas y transmitidas por vectores, inocuidad, resistencia a los antimicrobianos y ambiente (FAO et al., 2022).

Este capítulo examina la inteligencia artificial como componente de un sistema epidemiológico, no como sustituto de la vigilancia de campo. Se revisan sus usos en detección temprana, análisis espacial, pronóstico de brotes, redes de contacto, resistencia antimicrobiana, trazabilidad y vigilancia molecular. Los dos casos finales son simulados y muestran cómo formular problemas predictivos sin confundir clasificación con causalidad ni rendimiento interno con utilidad sanitaria.

Objetivos de aprendizaje

Al finalizar el capítulo, el lector podrá:

1. Explicar la función de la inteligencia artificial dentro del ciclo de vigilancia epidemiológica veterinaria.
2. Diferenciar detección de señales, estimación de riesgo, pronóstico, atribución causal y apoyo a decisiones.
3. Diseñar conjuntos de datos que integren información animal, humana y ambiental sin crear fugas de información.
4. Seleccionar métodos espaciales, temporales, de redes o genómicos según la pregunta científica.
5. Interpretar discriminación, calibración, oportunidad, utilidad y validez externa en modelos epidemiológicos.
6. Reconocer sesgos de notificación, muestreo, cobertura geográfica, verificación y cambio de dominio.
7. Proponer una estrategia de comunicación del riesgo que conserve la incertidumbre y evite alarmas injustificadas.

13.1 Inteligencia artificial en epidemiología veterinaria

La inteligencia artificial puede intervenir en distintas etapas del razonamiento epidemiológico. En la captura de datos, automatiza la extracción de diagnósticos desde historias clínicas o informes de laboratorio. Durante la vigilancia, identifica conglomerados inusuales y prioriza registros para revisión. En la investigación analítica, modela relaciones no lineales entre exposición, huésped y ambiente. En la respuesta, compara escenarios y ayuda a asignar inspecciones, muestreos o vacunas. Cada función requiere una pregunta, una unidad de análisis y un horizonte temporal explícitos.

Conviene separar cuatro propósitos. La descripción resume lo observado: incidencia por cantón, mortalidad semanal o distribución de serotipos. La inferencia estima asociaciones y su incertidumbre. La predicción calcula la probabilidad de un evento futuro o no observado. La inferencia causal pregunta qué cambiaría si se aplicara una intervención, como restringir movimientos o aumentar la vacunación. Un random forest puede clasificar predios con riesgo elevado y, aun así, no demostrar que la variable con mayor importancia sea una causa modificable. Para evaluar efectos se necesitan diseños y supuestos causales adecuados, además del algoritmo.

La unidad de análisis puede ser un animal, rebaño, granja, lote, establecimiento de faenamiento, municipio, celda espacial, semana epidemiológica, aislamiento bacteriano o nodo de una red. La unidad debe coincidir con la decisión. Si la autoridad asigna brigadas a parroquias, un modelo individual puede requerir agregación y recalibración. Si el objetivo es decidir el aislamiento de un animal, una predicción regional resulta demasiado general.

Los modelos más útiles suelen combinar conocimientos epidemiológicos y métodos computacionales. La regresión logística, los modelos de Poisson o binomial negativa y los modelos jerárquicos mantienen ventajas cuando interesan asociaciones interpretables, sobredispersión, estructura por rebaño o estimación de efectos. Árboles, random forest y boosting aportan flexibilidad ante interacciones y no linealidad. Las redes neuronales resultan justificables en imágenes, texto, señales de sensores o secuencias de gran dimensión. La revisión de Pillai et al. (2022) muestra la variedad de modelos aplicados a patógenos zoonóticos, pero también la heterogeneidad de datos y validaciones. La complejidad debe responder al problema y al volumen efectivo de información, no al atractivo del algoritmo.

Martínez-López et al. (2025) plantean la epidemiología veterinaria de precisión como uso de datos multinivel para comprender y gestionar la salud de poblaciones animales. El concepto es valioso si se conserva la perspectiva poblacional. Una alerta producida por sensores en una granja intensiva puede mejorar la detección local, pero no representa automáticamente a productores pequeños, sistemas pastoriles o fauna silvestre con menor digitalización.

Tabla 13.1

Pregunta epidemiológica y enfoque analítico

Pregunta	Unidad habitual	Enfoque apropiado	Producto	Riesgo de interpretación
¿Qué ocurrió y dónde?	Caso, predio, territorio	Descriptiva, SIG, tasas estandarizadas	Mapa o boletín	Confundir conteos con riesgo
¿Qué factores se asocian con el evento?	Animal o rebaño	Regresión, modelos mixtos, GAM	Razones de riesgo u odds	Interpretar asociación como causa

¿Dónde aparecerán casos?	Celda o área y periodo	Modelos espaciotemporales, boosting	Superficie de riesgo	Extrapolar fuera del dominio
¿Cuándo aumentará la incidencia?	Semana o mes	Series temporales, modelos dinámicos	Pronóstico con intervalo	Ignorar retrasos de notificación
¿Qué intervención reduciría la transmisión?	Red, población o territorio	Modelos causales y simulación	Escenarios comparables	Suponer parámetros no identificados

13.2 Vigilancia y detección temprana de enfermedades

La vigilancia es la recopilación, análisis, interpretación y difusión continuos de información para orientar acciones sanitarias. Puede ser pasiva, activa, centinela, basada en eventos, sindrómica, de laboratorio o participativa. La inteligencia artificial aporta sobre todo en volumen, velocidad y priorización: filtra texto, detecta anomalías, combina señales y ordena unidades para investigación. La confirmación continúa dependiendo de definiciones de caso, investigación de campo y pruebas diagnósticas.

En vigilancia pasiva, los datos representan contactos con el sistema y decisiones de notificación. Un aumento de reportes puede reflejar mayor transmisión, una campaña de sensibilización, cambios en el acceso al laboratorio o una nueva obligación normativa. El algoritmo aprende el proceso de observación junto con la enfermedad. Por ello deben registrarse variables de cobertura: número de veterinarios, pruebas realizadas, oportunidad de reporte, disponibilidad de reactivos y población bajo vigilancia.

La detección de anomalías compara lo observado con un patrón esperado. Métodos como CUSUM, EWMA, Farrington, isolation forest o autoencoders pueden señalar excesos, pero cada uno responde a estructuras diferentes. En eventos raros, un umbral muy sensible genera numerosas falsas alarmas. El rendimiento debe medirse con sensibilidad de brote, tasa de falsas alertas, retraso de detección y carga de revisión. La exactitud global es poco informativa cuando casi todas las semanas carecen de brote.

El procesamiento de lenguaje natural puede extraer especie, síndrome, ubicación y fecha desde notas clínicas, reportes de necropsia o noticias. Debe evaluarse por entidad y no solo por

documento. Un error en la negación cambia el significado de “sin signos neurológicos”; otro en la geolocalización puede desplazar el foco. Los sistemas basados en eventos requieren revisión humana, trazabilidad de la fuente y una política para rumores, duplicados y desinformación.

Una arquitectura razonable contiene cinco capas: adquisición, control de calidad, detección, verificación y respuesta. La salida del modelo no debe publicarse como caso confirmado. Se convierte primero en señal; luego un equipo contrasta antecedentes, laboratorio, vínculo epidemiológico y consistencia espacio-temporal. WOAH mantiene WAHIS como plataforma oficial de información sobre enfermedades animales. Las fuentes informales pueden complementar la oportunidad, pero no reemplazan la notificación oficial.

Figura 13.1.

Arquitectura de vigilancia epidemiológica asistida por IA

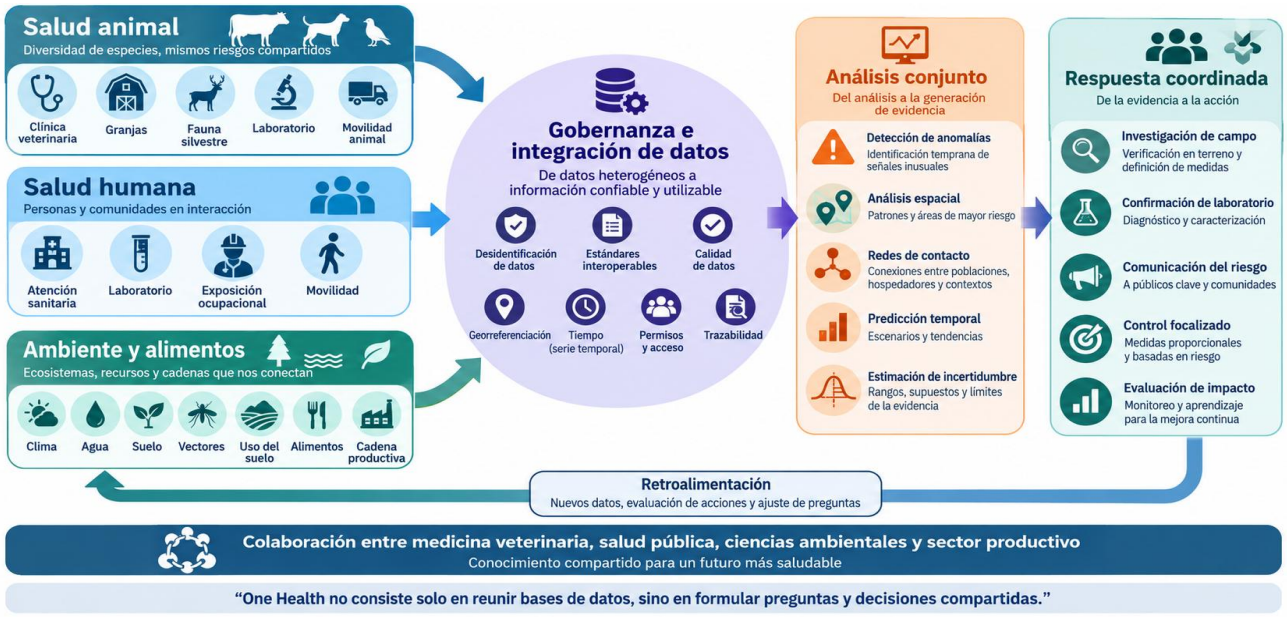


Tabla 13.2

Indicadores para evaluar un sistema de alerta

Dimensión	Indicador	Interpretación epidemiológica
Sensibilidad	Proporción de brotes detectados	Capacidad de no omitir eventos verdaderos

Valor predictivo positivo	Alertas confirmadas entre alertas investigadas	Eficiencia del uso de recursos
Oportunidad	Días desde inicio probable hasta alerta	Tiempo ganado para intervenir
Especificidad temporal	Semanas sin brote sin alerta	Control de falsas alarmas
Cobertura	Población o territorio con datos utilizables	Representatividad del sistema
Estabilidad	Disponibilidad y continuidad del flujo	Confiabilidad operativa
Aceptabilidad	Participación de notificadores	Viabilidad institucional

13.3 Modelos de riesgo y distribución espacial

Un modelo espacial busca explicar o predecir cómo cambia un evento según la ubicación. El primer requisito es definir el denominador. Mapear casos de rabia bovina sin considerar la población bovina, el esfuerzo de muestreo y el acceso a diagnóstico produce un mapa de notificación, no de incidencia. Cuando el denominador es incierto, puede ser preferible modelar presencia, positividad o intensidad de notificación con una declaración explícita de sus límites.

Los predictores pueden incluir elevación, temperatura, precipitación, cobertura vegetal, uso del suelo, densidad animal, proximidad a mercados, vías, mataderos, fronteras, cuerpos de agua y hábitat de vectores o reservorios. Las capas deben corresponder al periodo del desenlace. Utilizar un mapa actual de uso del suelo para explicar brotes ocurridos diez años antes introduce error temporal. También debe documentarse resolución, sistema de coordenadas, fuente y proporción de valores faltantes.

La autocorrelación espacial viola la independencia convencional: unidades cercanas tienden a parecerse. Los modelos autorregresivos, gaussianos, bayesianos jerárquicos y GAM con suavizadores espaciales permiten representar esta estructura. Los algoritmos de árboles pueden capturar patrones no lineales, pero no corrigen por sí mismos la dependencia. Una división aleatoria de registros próximos coloca observaciones casi idénticas en entrenamiento y prueba y sobreestima el rendimiento.

La validación espacial agrupa observaciones por bloques geográficos o deja regiones completas fuera del entrenamiento. Si el uso previsto es transferir un modelo entre provincias, la prueba debe imitar esa transferencia. Además de AUROC o AUPRC, deben evaluarse calibración regional, sensibilidad en zonas de baja cobertura y estabilidad de la clasificación de riesgo. Un mapa con colores continuos puede transmitir una falsa precisión; las clases deben corresponder a decisiones y acompañarse de incertidumbre.

La distribución potencial de una enfermedad no equivale a riesgo de brote. La idoneidad ambiental puede ser alta, pero la introducción depende de movilidad, bioseguridad, inmunidad y presencia del patógeno. Los modelos de nicho son útiles para orientar muestreo de vectores o reservorios, siempre que se distinga hábitat adecuado, probabilidad de presencia y riesgo sanitario.

13.4 Predicción de brotes

Predecir un brote exige definir qué constituye un brote, con qué anticipación debe detectarse y qué acción seguirá a la alerta. “Próximo brote” puede significar al menos un caso confirmado en 30 días, superar un umbral endémico o detectar transmisión en un territorio previamente libre. La etiqueta debe fijarse antes del modelado y justificarse sanitariamente.

Las variables históricas incluyen incidencia rezagada, estacionalidad, resultados de laboratorio, coberturas vacunales, movimientos, densidad de hospedadores, índices de vegetación y clima. El uso de rezagos protege el orden temporal. Una media calculada con datos posteriores a la fecha de predicción crea fuga de información. También ocurre cuando se imputan datos con toda la serie antes de separar entrenamiento y prueba o cuando brotes relacionados aparecen en ambos conjuntos.

Los modelos de referencia son esenciales. Un pronóstico estacional, la incidencia de la semana anterior o una regresión con pocos predictores pueden competir con métodos complejos. ARIMA, modelos de estado, regresión de Poisson, GAM y modelos mecanísticos describen estructuras epidemiológicas reconocibles. Gradient boosting, random forest y redes recurrentes pueden mejorar el rendimiento si existe suficiente historia y estabilidad del sistema. La comparación debe usar el mismo horizonte, ventanas temporales y datos disponibles.

La validación retrospectiva debe avanzar en el tiempo. En una evaluación con origen rodante, el modelo se entrena hasta una fecha y predice el periodo siguiente; luego la ventana avanza. Este procedimiento reproduce el uso prospectivo y muestra degradación por cambios en variantes, vacunación, clima o políticas. Una validación externa temporal o geográfica aporta más evidencia que repetir particiones aleatorias.

La probabilidad prevista debe calibrarse. Si cien áreas reciben un riesgo cercano a 0,20, aproximadamente veinte deberían presentar el evento bajo condiciones comparables. La discriminación ordena riesgos; la calibración determina si las probabilidades son utilizables para planificación. La decisión final necesita un umbral vinculado con costos: investigar una falsa alerta, perder un brote, movilizar personal y aplicar restricciones.

13.5 Zoonosis y enfermedades emergentes

Las zoonosis surgen de interacciones entre patógeno, hospedadores, vectores, prácticas productivas, movilidad humana, uso del suelo y clima. La IA puede ayudar a priorizar virus, especies o interfaces para vigilancia, pero no puede resolver la escasez de observación en fauna silvestre ni estimar por sí sola una tasa de salto interespecífico. Grange et al. (2021) desarrollaron un marco para ordenar virus de fauna según su potencial de transmisión a personas. Estos esquemas son instrumentos de priorización; su puntuación no representa una probabilidad clínica individual.

Los datos de emergencia presentan sesgos fuertes. Se secuencian más patógenos donde existen laboratorios y financiamiento. Las especies carismáticas o próximas a centros de investigación reciben mayor atención. Un modelo puede aprender que la actividad científica predice descubrimientos y confundirla con biología de emergencia. Entre los predictores deben considerarse esfuerzo de muestreo, número de publicaciones y accesibilidad, o realizar análisis de sensibilidad.

El cambio climático modifica áreas de distribución y oportunidades de contacto. Carlson et al. (2022) proyectaron que los desplazamientos geográficos de mamíferos pueden aumentar encuentros interespecíficos y oportunidades de intercambio viral. La conclusión no permite atribuir un brote concreto al clima sin análisis adicional. Sí justifica integrar escenarios climáticos, biodiversidad y vigilancia de interfaces en la planificación a largo plazo.

En animales de compañía, refugios y mercados, los sistemas pueden combinar consultas por síndrome, mortalidad, compras de medicamentos y resultados de laboratorio. En ganadería, se integran abortos, caídas de producción, mortalidad, vectores y movimientos. En fauna, las señales incluyen mortalidad inusual, rescates, observación comunitaria y secuencias. Una plataforma One Health debe conservar la procedencia de cada dato y su nivel de confirmación.

La ética de la predicción zoonótica incluye consecuencias económicas y sociales. Clasificar una zona como "foco probable" puede afectar comercio, turismo o percepción de una especie. La comunicación debe evitar estigmatizar comunidades o fauna y explicar que riesgo, presencia y transmisión confirmada no son equivalentes.

13.6 Resistencia a los antimicrobianos

La resistencia a los antimicrobianos es un problema One Health porque bacterias, genes, antimicrobianos y residuos circulan entre personas, animales, alimentos, aguas y suelos. El análisis puede centrarse en fenotipos de susceptibilidad, genes de resistencia, consumo, uso por especie, prácticas de prescripción, residuos ambientales o vínculos genómicos. Integrar sectores no significa sumar porcentajes obtenidos con poblaciones, antibióticos y puntos de corte diferentes.

La vigilancia fenotípica requiere métodos estandarizados, control de calidad y criterios interpretativos. Los perfiles pueden representarse como sensible, intermedio y resistente o mediante concentraciones mínimas inhibitorias. Antes de entrenar un modelo se debe decidir cómo se tratarán categorías intermedias, aislamientos repetidos, paneles incompletos y cambios de estándares. La fecha del aislamiento, origen de la muestra, especie animal, edad, producción, tratamiento previo y unidad epidemiológica son variables indispensables.

El aprendizaje automático se utiliza para predecir resistencia desde genomas, espectros, metagenomas o variables epidemiológicas. En secuencias, los predictores pueden ser genes, mutaciones, k-mers o representaciones aprendidas. La separación debe impedir que clones casi idénticos aparezcan en entrenamiento y prueba. Una partición aleatoria por aislamiento puede medir reconocimiento de linajes conocidos, no generalización a linajes o regiones nuevos.

El desempeño debe informarse por combinación bacteria-antimicrobiano. En clínica y vigilancia, los errores no tienen el mismo costo: clasificar un aislado resistente como susceptible puede orientar un tratamiento ineficaz; clasificar uno susceptible como resistente puede promover el uso de fármacos de mayor importancia. Se requieren sensibilidad para resistencia, especificidad, valores predictivos, calibración y análisis según prevalencia. La guía cuadripartita para vigilancia integrada de resistencia y uso recomienda objetivos explícitos, armonización, intercambio responsable e interpretación conjunta entre sectores (FAO et al., 2025).

Los modelos no sustituyen el antibiograma cuando la decisión exige confirmación. Pueden priorizar muestras, anticipar tendencias o apoyar vigilancia en contextos validados. El informe mundial GLASS ilustra la necesidad de sistemas comparables y de información sobre cobertura de pruebas, no solo porcentajes de resistencia (WHO, 2022). La estimación de la carga global de resistencia bacteriana también muestra la magnitud sanitaria del problema y la incertidumbre asociada a regiones con datos limitados (Murray et al., 2022).

Tabla 13.3

Niveles de análisis de resistencia antimicrobiana

Nivel	**Datos**	**Pregunta**	**Método posible**	**Limitación principal**
Aislamiento	CMI, disco, especie bacteriana	¿Es resistente?	Regresión o clasificación	Error de laboratorio y puntos de corte
Genómico	Genes, mutaciones, k-mers	¿Qué mecanismo explica el fenotipo?	Reglas, boosting, redes	Dependencia de linajes conocidos
Granja	Uso, bioseguridad, producción	¿Qué unidades requieren intervención?	Modelos mixtos o jerárquicos	Confusión y datos de uso incompletos
Cadena alimentaria	Animal, canal, producto, planta	¿Dónde aparece o se amplifica el riesgo?	Trazabilidad y redes	Muestreo no comparable entre etapas
One Health	Humanos, animales, alimentos, ambiente	¿Qué patrones se comparten entre sectores?	Integración multifuente y filogenia	Armonización, privacidad y atribución

13.7 Inocuidad y trazabilidad de alimentos

La inocuidad de alimentos de origen animal depende de controles a lo largo de la cadena. Los datos proceden de granjas, transporte, plantas de faenamiento, centros de acopio, laboratorios, cadena de frío, puntos de venta y notificaciones clínicas. La inteligencia artificial puede detectar combinaciones de temperatura y tiempo asociadas con fallos, clasificar imágenes de inspección, priorizar lotes para muestreo y apoyar la investigación retrospectiva de brotes.

La trazabilidad responde a dos preguntas: de dónde provino un producto y hacia dónde fue distribuido. Un identificador persistente vincula animal o lote, establecimiento, fecha, transformación y destino. Si los registros no comparten identificadores o si la captura se interrumpe, ningún algoritmo reconstruye con certeza la cadena. La tecnología de registro distribuido puede aumentar inmutabilidad, pero no garantiza que el dato inicial sea verdadero. La verificación, auditoría y gobernanza siguen siendo necesarias.

En análisis de riesgo microbiano, la predicción puede estimar probabilidad de contaminación o crecimiento según especie, proceso, temperatura, pH y tiempo. Deben conservarse las etapas del marco de riesgo: identificación del peligro, caracterización, evaluación de exposición y caracterización del riesgo. Un clasificador que distingue lotes positivos no equivale a una evaluación cuantitativa del riesgo de enfermedad en consumidores.

Las alertas de inocuidad requieren alta sensibilidad y explicación operativa. Variables importantes pueden identificar correlaciones con planta, turno o proveedor, pero una inspección debe confirmar condiciones. El uso punitivo de modelos opacos puede incentivar subregistro. Un sistema de aprendizaje sanitario necesita reglas claras sobre acceso, corrección de errores y separación entre mejora de procesos y sanción.

FAO ha señalado que automatización, inteligencia artificial, grandes datos y tecnologías de trazabilidad pueden apoyar la gestión de la inocuidad, siempre dentro de sistemas de control alimentario (FAO, 2023). En productos cárnicos, lácteos, huevos, pesca y acuicultura, el valor depende de interoperabilidad, mantenimiento de sensores y capacidad de actuar cuando el sistema detecta una desviación.

13.8 Integración de datos animales humanos y ambientales

La integración One Health empieza con una pregunta común. Para investigar resistencia a cefalosporinas, por ejemplo, pueden vincularse aislamientos de animales, alimentos, pacientes y aguas residuales, además de uso de antimicrobianos y metadatos de tiempo y lugar. La mera coexistencia de bases no demuestra transmisión. Se necesita una estructura analítica que defina interfaces, ventanas temporales y mecanismos plausibles.

Existen tres niveles. La integración conceptual acuerda objetivos y definiciones. La integración técnica armoniza formatos, identificadores, vocabularios y metadatos. La integración analítica combina fuentes en un modelo, una red o una síntesis. Un proyecto puede alcanzar un nivel sin completar los otros. También puede ser preferible un análisis federado que mantenga datos sensibles en cada institución y comparta parámetros o resultados agregados.

Los modelos de datos deben registrar procedencia, fecha de actualización, permisos y transformaciones. La misma etiqueta “positivo” puede significar PCR, cultivo, serología o diagnóstico clínico. Antes de fusionar, se crea un diccionario que incluya definición, unidad, dominio válido, método de medición y sistema responsable. Los códigos geográficos deben mantener una tabla histórica porque límites administrativos y nombres cambian.

La vinculación probabilística puede unir registros sin identificador común mediante fecha, lugar, especie y establecimiento. Esa unión incorpora error. Conviene conservar una puntuación de confianza y repetir el análisis con distintos umbrales. Una falsa vinculación puede crear aparentes cadenas de transmisión; una vinculación demasiado estricta elimina relaciones reales.

La gobernanza requiere finalidad legítima, minimización, seguridad, roles de acceso y mecanismos de rendición de cuentas. Los datos de granjas pueden revelar producción, comercio o incumplimientos; los humanos contienen información personal. El beneficio colectivo no elimina el deber de proteger a productores, trabajadores y pacientes. La inteligencia One Health propuesta por la Cuadripartita se orienta a transformar datos de varios sectores en evaluación conjunta y acción coordinada, no a centralizar indiscriminadamente toda la información (FAO et al., 2023).

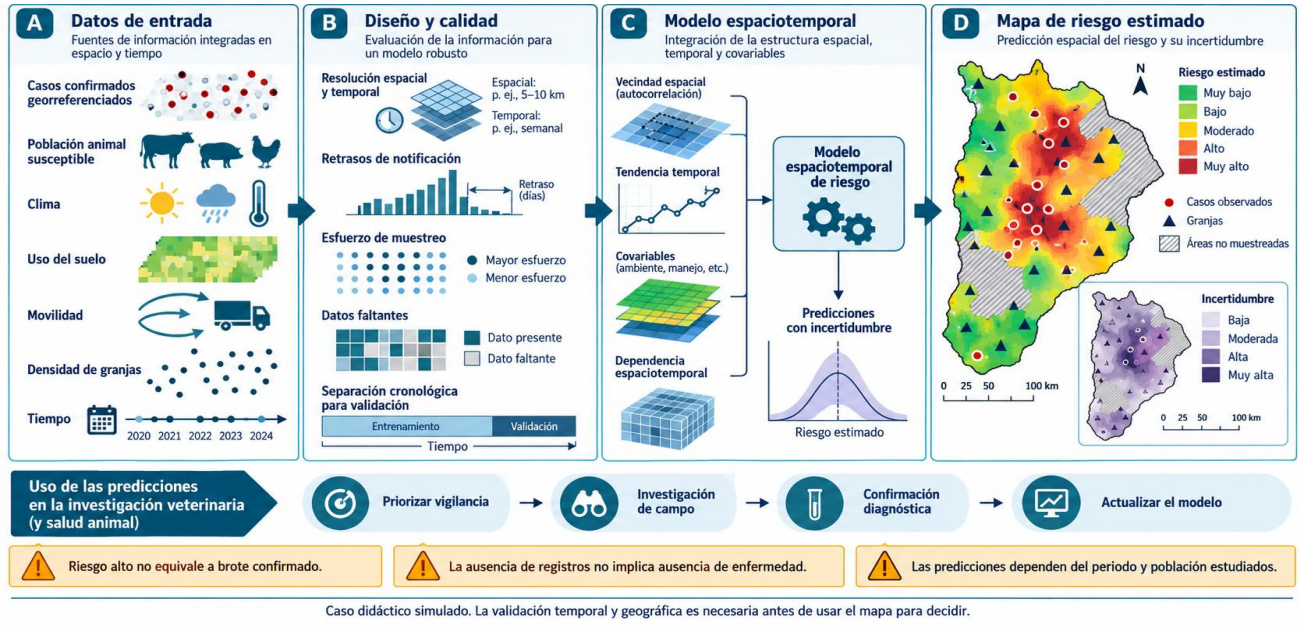
Tabla 13.4

Arquitectura mínima de integración One Health

Componente	Requisito	Ejemplo veterinario
Pregunta	Desenlace, población, tiempo y decisión	Riesgo semanal de leptospirosis por cantón
Semántica	Definiciones y códigos comunes	Especie, prueba y criterio de positividad
Tiempo	Fechas y rezagos comparables	Lluvia de semanas previas al diagnóstico
Espacio	Coordenadas y escalas compatibles	Predio, parroquia y cuenca hidrográfica
Procedencia	Fuente y transformación documentadas	Laboratorio oficial o clínica privada
Gobernanza	Acceso, finalidad y responsabilidad	Datos de predios seudonimizados
Acción	Umbral y protocolo de respuesta	Muestreo, vacunación o inspección dirigida

Figura 13.2.

Integración de datos bajo el enfoque One Health



13.9 Datos climáticos movilidad y sistemas productivos

El clima influye en vectores, supervivencia ambiental, disponibilidad de agua, migración de fauna y prácticas productivas. Temperatura, precipitación, humedad, inundación, sequía, índices de vegetación y anomalías oceánicas pueden incorporarse como exposiciones rezagadas. La relación rara vez es lineal. Un GAM o un modelo de árboles puede representar umbrales, pero la forma debe validarse y contrastarse con conocimiento biológico.

La resolución importa. Una estación meteorológica distante puede representar mal un valle; una celda satelital puede promediar microclimas. El error de exposición puede disminuir asociaciones o producir patrones artificiales. Las capas climáticas deben alinearse con coordenadas, periodo de incubación y ecología del vector. Probar numerosos rezagos sin corrección favorece hallazgos oportunistas.

La movilidad conecta poblaciones. Guías de movimiento, mercados, rutas, ferias, trashumancia, transporte de alimento, vehículos de servicio y desplazamiento de fauna crean vías diferentes. En sistemas informales, los registros administrativos capturan solo una parte. Encuestas, telefonía, sensores o modelos gravitacionales pueden completar información, pero toda inferencia debe distinguir movimiento observado de movimiento estimado.

Los sistemas productivos modifican el riesgo. Densidad, integración vertical, ciclo productivo, mezcla de edades, bioseguridad y dependencia de mercados determinan contactos y detección. Un modelo entrenado en granjas tecnificadas no debe aplicarse sin validación a producción familiar. La variable “tipo de granja” puede ocultar diferencias de vigilancia y no solo de biología.

Para pronóstico operacional se recomienda un conjunto de variables disponible antes de la fecha de predicción. Los productos climáticos revisados retrospectivamente pueden ser más precisos que los disponibles en tiempo real; entrenar con ellos genera una ventaja que desaparecerá en producción. La evaluación debe reconstruir la información que realmente habría existido.

13.10 Modelos de redes de contacto

Una red representa unidades epidemiológicas como nodos y relaciones como aristas. Los nodos pueden ser granjas, mercados, mataderos, animales, personas o lugares. Las aristas pueden indicar movimiento, proximidad, contacto compartido, comercio o similitud genética. La definición debe corresponder a una vía plausible de transmisión.

Las medidas de grado, intermediación, cercanía, comunidades y componentes ayudan a identificar nodos conectados, puentes y subredes. Un mercado con alta intermediación puede conectar regiones aunque no tenga el mayor volumen. Estas medidas no son propiedades biológicas fijas: dependen de la ventana temporal, dirección, peso y cobertura de la red.

Los modelos epidémicos sobre redes simulan estados susceptibles, infectados y recuperados, tiempos de movimiento y transmisión dentro y entre rebaños. Pueden comparar vigilancia dirigida a nodos centrales, restricciones de movimiento o vacunación. Los resultados dependen de parámetros y de movimientos no observados. Se necesitan análisis de sensibilidad y escenarios, no una única curva presentada como pronóstico cierto.

La validación debe respetar el tiempo. Una centralidad calculada con movimientos posteriores al brote no puede emplearse para una alerta previa. Asimismo, la red de entrenamiento y la de evaluación pueden compartir los mismos brotes. La partición por periodo o región reduce contaminación y aproxima el uso real.

Los algoritmos de grafos y las graph neural networks permiten aprender representaciones de nodos y enlaces. Son útiles cuando existen muchas redes y atributos, pero requieren muestras suficientes y comparación con reglas epidemiológicas simples. Para una red nacional única, un modelo mecanístico o métricas clásicas pueden ser más interpretables y estables.

13.11 Genómica metagenómica y vigilancia molecular

La secuenciación de genoma completo aumenta la resolución para identificar linajes, investigar conglomerados y describir genes de virulencia o resistencia. La genómica epidemiológica combina secuencias con tiempo, lugar, hospedador y exposición. Dos aislamientos cercanos genéticamente

apoyan una relación, pero no prueban transmisión directa: pueden compartir una fuente no muestreada.

El flujo analítico incluye calidad de lectura, ensamblaje o mapeo, control de contaminación, tipificación, identificación de variantes y análisis filogenético. Los parámetros deben documentarse porque cambios en cobertura, referencia o filtros alteran distancias. En vigilancia rutinaria conviene usar versiones fijadas, controles positivos y negativos y un sistema para reanalizar datos cuando cambien bases de referencia.

La metagenómica detecta material genético sin cultivo dirigido y puede revelar patógenos inesperados, microbiomas y resistomas. Sus dificultades incluyen abundancia baja, contaminación, secuencias del hospedador, bases de datos incompletas y dificultad para distinguir viabilidad. La clasificación taxonómica automatizada produce hipótesis que deben confirmarse, sobre todo ante agentes de alto impacto.

El aprendizaje automático puede clasificar secuencias, predecir fenotipos y priorizar variantes. La dimensión elevada exige selección, regularización y conjuntos independientes. La similitud filogenética debe formar parte del diseño de validación. Si genomas casi clonales se reparten al azar, el rendimiento medirá memorización de linaje. Las pruebas por clado, año, país o laboratorio son más exigentes y más informativas.

La revisión de Catalano et al. (2024) muestra que la genómica One Health puede investigar transmisión entre dominios, pero su valor depende de muestreo coordinado y metadatos comparables. La ausencia de una secuencia en un sector puede reflejar falta de muestreo. La vigilancia molecular debe declarar cobertura y no convertir un árbol filogenético incompleto en una narrativa causal cerrada.

13.12 Sesgos incertidumbre y comunicación del riesgo

Los sesgos más frecuentes se originan antes del modelado. El sesgo de selección aparece cuando las muestras proceden de animales con acceso a diagnóstico. El sesgo de notificación varía con incentivos y sanciones. El sesgo de verificación surge cuando solo algunos sospechosos reciben la prueba de referencia. El sesgo de espectro ocurre si se entrena con casos graves y controles sanos,

pero se aplica a cuadros ambiguos. La fuga de información incorpora datos posteriores o derivados del desenlace.

La incertidumbre incluye variación aleatoria, error de medición, falta de datos, incertidumbre de parámetros, elección de modelo y cambio futuro del sistema. Los intervalos de confianza o credibilidad cubren solo componentes específicos. La variación entre algoritmos, definiciones y escenarios debe presentarse mediante análisis de sensibilidad. En mapas, puede mostrarse la probabilidad de superar un umbral junto con el ancho del intervalo.

El desbalance exige métricas adecuadas. AUROC puede parecer alto con pocos brotes; AUPRC refleja mejor el desempeño sobre la clase positiva, aunque depende de prevalencia. Sensibilidad, especificidad, valores predictivos y matriz de confusión deben informarse en un umbral operativo. Para probabilidades se añaden Brier, pendiente e intercepto de calibración. Para brotes importa además el retraso de detección.

La validación externa responde si el modelo funciona en otra población, lugar o periodo. El cambio de dominio puede originarse en nuevas variantes, vacunación, clima, pruebas, flujos comerciales o definiciones. Un modelo debe contar con vigilancia de rendimiento, recalibración y criterios de suspensión. TRIPOD+AI ofrece una guía para informar modelos de predicción basados en regresión o aprendizaje automático (Collins et al., 2024); PROBAST+AI ayuda a evaluar calidad, sesgo y aplicabilidad (Moons et al., 2025).

La comunicación del riesgo debe indicar evento, población, periodo, magnitud, incertidumbre y acción. “Riesgo alto” carece de significado si no se define el umbral. Una formulación útil sería: “El modelo estima una probabilidad de 0,32 de al menos un foco confirmado en las próximas cuatro semanas; el intervalo de predicción es amplio y la señal se empleará para priorizar muestreo, no para declarar un brote”. Debe explicarse qué datos faltan y cuándo se actualizará la estimación.

Figura 13.3.

Representaciones analíticas para vigilancia veterinaria



Tabla 13.5

Lista de comprobación antes del uso operativo

Pregunta	Evidencia mínima
¿La población objetivo coincide con el entrenamiento?	Descripción de cobertura, especies y sistemas productivos
¿La validación reproduce el uso futuro?	Separación temporal, espacial y por unidad epidemiológica
¿Se evitó la fuga de información?	Preprocesamiento dentro de cada partición
¿Las probabilidades están calibradas?	Curva, intercepto y pendiente o Brier
¿El umbral corresponde a una acción?	Costos, capacidad y protocolo definidos
¿Se evaluó equidad territorial?	Rendimiento por región y nivel de acceso
¿Existe supervisión posterior?	Monitoreo, recalibración y retiro
¿La comunicación conserva incertidumbre?	Probabilidad, horizonte y límites explícitos

13.13 Caso didáctico predicción espacial de una enfermedad zoonótica

Planteamiento

Caso didáctico simulado. Una autoridad provincial desea priorizar vigilancia de leptospirosis en bovinos y su interfaz con casos humanos. La predicción se realizará por celdas de 5 por 5 km para las siguientes cuatro semanas. El objetivo no es declarar causalidad ni sustituir diagnóstico, sino seleccionar áreas para muestreo animal, evaluación de agua y coordinación con salud humana.

Pregunta: ¿qué celdas presentan mayor probabilidad de registrar al menos un evento animal confirmado durante las próximas cuatro semanas, utilizando información disponible al cierre de cada semana?

Hipótesis predictiva: la incorporación de precipitación rezagada, inundación, densidad bovina, proximidad a cuerpos de agua, notificaciones históricas y cobertura de muestreo mejorará la predicción respecto a un modelo basado solo en estacionalidad e incidencia previa.

Diseño y datos simulados

La base contiene 240 celdas observadas durante 156 semanas, equivalentes a 37 440 registros celda-semana. Se simulan 312 eventos positivos, por lo que el desenlace es raro. Las fuentes previstas son el laboratorio veterinario, registros de población bovina, imágenes de inundación, estaciones meteorológicas y reportes agregados de síndrome febril humano. Los datos humanos se incorporan con rezago y agregación suficiente para proteger privacidad.

La unidad de análisis es la celda-semana. El desenlace vale 1 si existe al menos un caso animal confirmado en las cuatro semanas siguientes. Los predictores se calculan con información previa: lluvia acumulada de una a cuatro semanas, anomalía de vegetación, superficie inundada, altitud, densidad bovina, distancia a río, eventos de las ocho semanas anteriores, número de muestras procesadas y distancia al laboratorio.

La variable “casos humanos futuros” queda excluida porque no estaría disponible. El número de muestras se incluye para representar esfuerzo de observación. Las coordenadas exactas de predios se reemplazan por celda y se mantiene una tabla segura de correspondencia fuera del conjunto analítico.

Estrategia de análisis

Se comparan tres modelos: regresión logística con splines para lluvia y estacionalidad; GAM con superficie espacial; y XGBoost con ponderación de clase. Un modelo nulo utiliza mes e incidencia rezagada. El preprocesamiento, imputación y selección se ajustan solo con datos de entrenamiento.

La validación se divide por tiempo y espacio. Las primeras 104 semanas forman el periodo de desarrollo. Las siguientes 26 se usan para ajuste. Las últimas 26 constituyen prueba temporal. Además, se aplica validación cruzada espacial en cinco bloques durante desarrollo. Las celdas vecinas permanecen juntas para reducir optimismo.

Las métricas principales son AUPRC, sensibilidad al umbral operativo, valor predictivo positivo, Brier, pendiente de calibración y número de celdas que deben inspeccionarse por evento detectado. AUROC se informa como complemento. El umbral se elige para que la lista semanal no exceda 25 celdas, capacidad real de las brigadas.

Resultados simulados

En la prueba temporal simulada, el modelo nulo alcanza AUPRC 0,18. La regresión logística obtiene 0,27; el GAM, 0,31; y XGBoost, 0,34. A un máximo de 25 celdas por semana, XGBoost presenta sensibilidad 0,71 y valor predictivo positivo 0,22. La pendiente de calibración es 0,76, lo que indica predicciones demasiado extremas. Después de recalibración isotónica ajustada en el conjunto de validación, el Brier disminuye de 0,041 a 0,037.

Estos valores son ficticios y sirven para interpretación. Un valor predictivo positivo de 0,22 significa que muchas inspecciones no encontrarán el evento; puede ser aceptable si el costo de perder un foco es alto y la visita aporta otras acciones preventivas. AUPRC debe compararse con la prevalencia del desenlace. La mejora frente al modelo nulo muestra valor incremental, pero no demuestra que lluvia o inundación causen cada caso.

Interpretación y decisión

El GAM ofrece menor discriminación que XGBoost, pero muestra relaciones espaciales y climáticas más interpretables. La autoridad podría seleccionar XGBoost para priorización y mantener el GAM como análisis de contraste. Si ambos señalan una zona, aumenta la confianza; si divergen, se revisan cobertura, extrapolación y datos faltantes.

El mapa operativo presentará tres capas: probabilidad recalibrada, incertidumbre por remuestreo y esfuerzo de muestreo histórico. Las zonas con riesgo alto y cobertura baja serán candidatas

prioritarias. La alerta no se comunicará como “brote probable”, sino como recomendación de vigilancia reforzada.

Limitaciones

La confirmación depende de acceso al laboratorio y puede concentrarse cerca de vías.

Los casos humanos agregados tienen retrasos y no establecen dirección de transmisión.

La resolución de 5 km puede ocultar focos locales y genera el problema de unidad espacial modificable.

Un evento extremo fuera del rango de entrenamiento requiere evaluación de extrapolación.

El modelo debe validarse prospectivamente antes de justificar restricciones o intervenciones costosas.

13.14 Caso didáctico clasificación de perfiles de resistencia antimicrobiana

Planteamiento

Caso didáctico simulado. Una red de laboratorios recibe aislamientos de Escherichia coli procedentes de bovinos, porcinos, aves, alimentos y aguas cercanas a unidades productivas. Se busca clasificar perfiles multirresistentes para priorizar confirmación, análisis genómico y auditoría de calidad. El modelo no reemplazará la prueba de susceptibilidad.

Pregunta: ¿puede un modelo predecir si un aislamiento presentará resistencia a tres o más clases, utilizando datos disponibles antes del panel completo?

Hipótesis predictiva: especie de origen, tipo de muestra, edad del animal, exposición antimicrobiana registrada, sistema productivo, región, año, resultados preliminares y marcadores genómicos seleccionados permitirán discriminar perfiles multirresistentes; la generalización disminuirá al evaluar años, regiones y linajes no observados.

Diseño y datos simulados

Se simulan 5 200 aislamientos recogidos entre 2021 y 2025. Cada registro contiene identificador de predio o establecimiento, sector, especie, muestra, fecha, laboratorio, método, panel de 12 antimicrobianos, genes de resistencia y linaje. El desenlace se define antes del análisis como resistencia fenotípica a tres o más clases, usando criterios armonizados.

Los duplicados del mismo animal y episodio se eliminan según una regla preestablecida. Para evitar dependencia, todos los aislamientos del mismo predio y clúster genómico se asignan a una sola partición. Los datos de 2021-2023 forman entrenamiento; 2024, validación; 2025, prueba temporal. Una región completa se reserva además para validación geográfica.

Los valores faltantes no se codifican automáticamente como ausencia. Se crean indicadores de prueba no realizada y se aplica imputación dentro del entrenamiento. Los genes identificados después del fenotipo no se utilizarían en una herramienta de alerta previa si no están disponibles a tiempo. Se construyen dos modelos: uno epidemiológico sin genómica y otro ampliado con marcadores rápidos.

Modelos y métricas

Se comparan regresión logística penalizada, random forest y LightGBM. La clase minoritaria se maneja con ponderación; no se aplica sobremuestreo antes de dividir por predio. La selección de hiperparámetros usa validación agrupada por establecimiento.

Las métricas principales son sensibilidad para multirresistencia, especificidad, AUPRC, valor predictivo positivo, Brier y calibración. Se presentan por sector, especie, laboratorio y región. Para cada combinación antimicrobiano-bacteria se añade una matriz de errores. La utilidad se evalúa por número de pruebas confirmatorias priorizadas y falsos negativos evitados.

Resultados simulados

En la prueba temporal, la regresión penalizada alcanza AUPRC 0,63, random forest 0,66 y LightGBM 0,69. El modelo ampliado con marcadores rápidos alcanza 0,77. Con un umbral elegido para sensibilidad 0,90, su especificidad es 0,61 y su valor predictivo positivo 0,58. En la región externa, AUPRC cae a 0,62 y la calibración sobreestima el riesgo.

La caída externa sugiere cambio de dominio. El análisis por linaje muestra que parte del rendimiento interno se debía a clones frecuentes. La recalibración regional mejora probabilidades, pero no corrige por completo la pérdida de discriminación. Se decide utilizar el modelo para priorizar secuenciación y revisión, no para informar susceptibilidad clínica.

Interpretación y límites

La importancia de "exposición antimicrobiana previa" no demuestra que esa exposición haya causado el perfil. Puede representar gravedad, prácticas de registro o selección de muestras. Para estimar el efecto de reducir un antimicrobiano se necesita un diseño causal con temporalidad, confusores y comparación apropiada.

El modelo se someterá a control trimestral de prevalencia, calibración y desempeño por sector. Si la sensibilidad cae por debajo de 0,85 o aparece un linaje nuevo con falsos negativos, se suspenderá la clasificación automática y se revisará el conjunto de entrenamiento.

Producto operativo

Cada aislamiento recibirá una probabilidad, una categoría de prioridad y una explicación limitada a variables disponibles. El informe indicará que se trata de una predicción. Los perfiles confirmados alimentarán nuevamente la vigilancia solo después de controles de calidad y cierre del episodio, evitando que etiquetas provisionales contaminen el aprendizaje.

13.15 Síntesis autoevaluación y actividad práctica

Síntesis

La inteligencia artificial amplía la epidemiología veterinaria cuando convierte datos heterogéneos en señales revisables, probabilidades calibradas y prioridades de acción. Su contribución depende de vigilancia de calidad, definiciones estables y una unidad de análisis vinculada con la decisión. En brotes raros, la oportunidad y la carga de falsas alertas importan tanto como la discriminación.

El enfoque One Health obliga a integrar preguntas, no solo bases. Los datos animales, humanos y ambientales requieren semántica, tiempo, espacio, procedencia y gobernanza compatibles. La

resistencia antimicrobiana y la vigilancia molecular muestran con claridad el valor de esa integración, así como el peligro de inferir transmisión desde muestras incompletas.

Los modelos espaciales, temporales y de redes deben validarse con particiones que respeten dependencia geográfica, cronológica y epidemiológica. La partición aleatoria suele ser optimista. Una herramienta destinada a nuevas regiones necesita prueba regional; una destinada al futuro necesita prueba temporal; una destinada a nuevos linajes necesita separación filogenética.

La salida del algoritmo no es la decisión sanitaria. Debe incorporarse a un protocolo con verificación, umbrales, capacidad operativa, comunicación de incertidumbre y vigilancia posterior. Ningún rendimiento interno compensa un sistema de notificación sesgado o una acción sin fundamento.

Preguntas de autoevaluación

1. ¿Por qué un mapa de conteos de casos no representa necesariamente un mapa de riesgo?
2. Diferencie una señal, un caso probable y un caso confirmado dentro de un sistema de vigilancia.
3. ¿Qué fuga de información puede ocurrir al calcular variables climáticas o incidencias móviles?
4. ¿Por qué una partición aleatoria es inadecuada cuando existen celdas vecinas o aislamientos clonales?
5. ¿Qué diferencia existe entre discriminación y calibración?
6. ¿Por qué AUPRC puede ser más informativa que exactitud en brotes raros?
7. ¿Qué variables permiten distinguir aumento de enfermedad de aumento de esfuerzo diagnóstico?
8. ¿Qué condiciones deben cumplirse antes de integrar datos humanos y animales?
9. ¿Por qué la cercanía genómica no demuestra transmisión directa?
10. ¿Qué aporta una red de contactos frente a un análisis tabular convencional?
11. ¿Cómo evaluaría el cambio de dominio entre producción intensiva y familiar?
12. ¿Qué daño puede producir una comunicación imprecisa de “alto riesgo”?
13. ¿Cuándo sería preferible un modelo de regresión o GAM frente a una red neuronal?

14. ¿Qué métricas utilizaría para un sistema de alerta con capacidad limitada de inspección?
15. ¿Qué evidencia adicional se requiere para pasar de asociación predictiva a recomendación causal?

Respuestas orientativas

1. Porque el conteo depende de población, muestreo y acceso diagnóstico.
2. La señal requiere verificación; el caso probable cumple criterios parciales; el confirmado cumple el estándar establecido.
3. Usar información posterior a la fecha de predicción o calcular transformaciones con toda la serie.
4. Porque comparte estructura casi idéntica entre entrenamiento y prueba y sobreestima generalización.
5. La discriminación ordena unidades; la calibración compara probabilidades con frecuencias observadas.
6. Porque se concentra en la clase positiva y su precisión frente a una prevalencia baja.
7. Número de pruebas, cobertura, distancia al laboratorio, notificadores activos y retraso.
8. Objetivo común, definiciones compatibles, base legal, protección, procedencia y plan de acción.
9. Puede existir una fuente intermedia no muestreada y la resolución depende del patógeno.
10. Representa rutas y nodos puente que no aparecen en variables independientes.
11. Con validación externa estratificada y análisis de cobertura, calibración y errores.
12. Estigma, pérdidas económicas, acciones desproporcionadas y pérdida de confianza.
13. Cuando la muestra es limitada, se requiere interpretabilidad o la estructura epidemiológica puede especificarse.
14. Sensibilidad, valor predictivo positivo, retraso, AUPRC, calibración y número de inspecciones.
15. Diseño causal, temporalidad, control de confusión, supuestos explícitos y análisis de intervención.

Actividad práctica integradora

Título: Diseño de un sistema de alerta temprana One Health.

Organice equipos de tres a cinco estudiantes. Cada equipo seleccionará una enfermedad zoonótica pertinente para su territorio, como rabia, leptospirosis, brucelosis, influenza aviar o fiebre del Nilo Occidental. El trabajo debe basarse en datos simulados o públicos y declarar su origen.

Producto 1. Formule una pregunta predictiva con población, unidad, desenlace, horizonte y acción. Dibuje el flujo desde fuente hasta respuesta. Identifique quién genera, custodia, analiza y recibe cada dato.

Producto 2. Elabore un diccionario de al menos 20 variables. Incluya definición, tipo, unidad, fuente, frecuencia, rezago, porcentaje esperado de ausencia, sensibilidad y disponibilidad en tiempo real.

Producto 3. Proponga un diseño de validación. Debe contener una prueba temporal, una prueba espacial o institucional, un modelo de referencia y una estrategia contra fuga de información. Justifique las métricas y el umbral operativo.

Producto 4. Presente un mapa o tablero simulado con riesgo e incertidumbre. Redacte un boletín de 150 palabras para veterinarios de campo y otro de 100 palabras para la comunidad. Ambos deben diferenciar predicción de confirmación.

Producto 5. Complete una matriz de riesgos éticos: privacidad, estigmatización, exclusión digital, uso punitivo, falsas alertas, omisión de focos y dependencia tecnológica. Para cada riesgo proponga una medida de control y un responsable.

Rúbrica sugerida

Criterio	Porcentaje	Desempeño esperado
Formulación epidemiológica	20	Unidad, desenlace, horizonte y decisión coherentes
Calidad e integración de datos	20	Fuentes, definiciones, rezagos y gobernanza documentados

Diseño analítico	20	Modelo de referencia, algoritmo y supuestos justificados
Validación e interpretación	20	Partición realista, métricas, calibración y límites
Comunicación y ética	20	Incertidumbre clara, lenguaje proporcional y controles

CAPÍTULO 14

VALIDACIÓN INTERPRETACIÓN Y EVALUACIÓN CRÍTICA DE MODELOS

Introducción

La calidad de un modelo predictivo no se demuestra porque alcance una cifra elevada en el conjunto con el que fue entrenado. Un sistema puede ordenar correctamente a muchos pacientes y, al mismo tiempo, asignar probabilidades exageradas, fallar en otra clínica, favorecer a determinados subgrupos o inducir decisiones que no mejoran la atención. En investigación veterinaria, donde los tamaños muestrales suelen ser modestos, las poblaciones son heterogéneas y las mediciones cambian entre equipos, laboratorios, especies y sistemas productivos, la distancia entre desempeño aparente y utilidad real puede ser considerable.

Validar significa someter las predicciones a una prueba independiente y cuantificar hasta qué punto conservan su significado fuera de la muestra de desarrollo. Interpretar significa comprender qué información utiliza el algoritmo, cómo responde a cambios plausibles y si sus explicaciones son compatibles con la fisiopatología y el proceso de medición. Evaluar críticamente exige ir un paso más allá: preguntar si el modelo añade valor sobre la práctica habitual, si sus errores son aceptables, si las decisiones derivadas benefician a los animales y si el rendimiento se mantiene con el paso del tiempo.

Este capítulo integra estos planos. Las métricas se presentan como instrumentos de razonamiento, no como un listado de fórmulas. La premisa central es sencilla: ningún valor aislado —ni exactitud, ni AUC, ni F1— certifica por sí mismo un modelo clínicamente válido. La evaluación debe corresponder al propósito, la población, la unidad de análisis, la prevalencia, los costos de error y el momento concreto de uso. La guía TRIPOD+AI refuerza esta exigencia de transparencia para modelos de predicción desarrollados con regresión o aprendizaje automático (Collins et al., 2024).

Objetivos de aprendizaje

1. Distinguir validación aparente, interna, temporal, geográfica, entre centros y entre especies.
2. Calcular e interpretar métricas de clasificación y regresión con incertidumbre y contexto veterinario.

3. Evaluar discriminación, calibración, utilidad clínica, robustez, explicabilidad y validez biológica.
4. Seleccionar umbrales según consecuencias clínicas y no únicamente por conveniencia matemática.
5. Comparar un modelo con profesionales y con la práctica habitual mediante diseños justos y reproducibles.
6. Diseñar un plan de evaluación de impacto y monitoreo posterior al despliegue.
7. Ejecutar una validación externa completa de un modelo de riesgo clínico mediante un caso didáctico simulado.

14.1 Validación interna y externa

La validación interna estima el optimismo que aparece cuando se desarrolla y evalúa un modelo con información procedente de la misma fuente. Su pregunta no es si el modelo funcionará en cualquier hospital, sino cuánto rendimiento perdería al aplicarse a nuevos individuos de la población de origen. La validación externa, en cambio, examina el desempeño en datos independientes que difieren en tiempo, lugar, centro, profesionales, equipos o composición biológica.

Una división aleatoria simple en entrenamiento y prueba es fácil de explicar, pero desperdicia información cuando la muestra es pequeña y puede producir resultados inestables. En conjuntos moderados suele preferirse bootstrap o validación cruzada repetida. El bootstrap permite estimar el optimismo del proceso completo: selección de variables, imputación, transformación, ajuste de hiperparámetros y entrenamiento. Si alguna de estas operaciones se realiza antes del remuestreo, la información del conjunto de evaluación se filtra hacia el modelo y el desempeño queda inflado.

La validación cruzada debe respetar la estructura de los datos. Si un perro aporta varias consultas, todas sus observaciones deben permanecer en el mismo pliegue. En datos de hatos, el agrupamiento debe hacerse por finca; en imágenes, por animal y no por imagen; en sensores, por individuo o unidad productiva. Separar fotografías del mismo animal entre entrenamiento y prueba crea una independencia ficticia. El algoritmo puede reconocer rasgos del paciente, del equipo o del fondo en lugar de la lesión.

La validación externa es una investigación autónoma. Requiere protocolo, criterios de elegibilidad, tamaño muestral, definiciones operativas, manejo de datos faltantes y análisis preespecificado. No debe utilizarse repetidamente para ajustar hiperparámetros: cuando el conjunto externo orienta cambios del modelo, deja de ser una prueba independiente y pasa a formar parte del desarrollo. La versión actualizada necesitará otra validación.

Tabla 14.1

Niveles de validación y pregunta científica

Nivel	Separación	Pregunta principal	Riesgo que detecta
Aparente	Mismos datos	¿Qué tan bien reproduce el modelo la muestra de desarrollo?	Optimismo no cuantificado
Interna	Remuestreo dentro de la fuente	¿Qué rendimiento se espera en nuevos sujetos similares?	Sobreajuste y variabilidad muestral
Temporal	Periodo posterior	¿Se mantiene ante cambios de frecuencia y práctica?	Deriva temporal
Geográfica o multicéntrica	Clínica región o finca distinta	¿Se transporta a otro contexto?	Cambios de mezcla de casos y medición
Prospectiva	Datos recogidos después de congelar el modelo	¿Funciona en el flujo previsto?	Sesgo retrospectivo y factibilidad
Impacto	Uso real frente a comparador	¿Mejora decisiones resultados o eficiencia?	Beneficio clínico no demostrado

14.2 Validación temporal geográfica y entre especies

La validación temporal utiliza pacientes atendidos después del periodo de desarrollo. Es más exigente que una división aleatoria porque incorpora cambios en prevalencia, tratamientos, criterios diagnósticos y registro. Un modelo de riesgo de mastitis entrenado antes de implantar sensores de conductividad puede deteriorarse después de modificar el sistema de ordeño o el protocolo de tratamiento. La separación debe respetar una fecha definida y evitar que la historia futura de una vaca aparezca como predictor de un evento pasado.

La validación geográfica evalúa transportabilidad. Una clínica universitaria recibe casos referidos y graves; una consulta primaria observa otra mezcla de enfermedad. En producción animal, altitud,

clima, genética, alimentación y manejo alteran tanto la incidencia como la relación entre predictores y desenlaces. Por ello, una AUC similar entre regiones no garantiza riesgos absolutos correctos: la calibración puede desplazarse cuando cambia la frecuencia del evento.

La transferencia entre especies merece una cautela especial. Perros y gatos comparten pruebas de laboratorio, pero no necesariamente distribuciones, puntos de corte, metabolismo ni mecanismos de enfermedad. Un modelo no queda validado para gatos porque sus variables tengan el mismo nombre. Debe justificarse la equivalencia biológica, armonizarse la medición y estimarse el desempeño por especie. La adaptación mediante transferencia de aprendizaje tampoco sustituye esta evaluación.

La heterogeneidad puede cuantificarse por centro, región, raza, sexo, edad, sistema productivo y equipo. Conviene reportar estimaciones con intervalos de confianza y evitar conclusiones fuertes en subgrupos con pocos eventos. Cuando existen múltiples centros, un análisis de efectos aleatorios o una validación interna-externa —entrenar con todos menos un centro y validar en el centro excluido— ayuda a estudiar la variabilidad de transportabilidad.

14.3 Sensibilidad especificidad y valores predictivos

Para un umbral dado, la matriz de confusión organiza verdaderos positivos, falsos positivos, verdaderos negativos y falsos negativos. La sensibilidad es la proporción de animales con el evento que el modelo identifica; la especificidad es la proporción sin el evento que clasifica como negativos. Ambas dependen del umbral y de la definición del estándar de referencia.

El valor predictivo positivo responde qué proporción de animales marcados como positivos realmente presenta o desarrollará el evento. El valor predictivo negativo responde lo equivalente entre los clasificados como negativos. A diferencia de sensibilidad y especificidad, cambian de forma directa con la prevalencia. Un clasificador de enfermedad rara puede conservar sensibilidad y especificidad y, sin embargo, producir numerosos falsos positivos cuando se usa como tamizaje poblacional.

Tabla 14.2

Interpretación clínica de métricas binarias

Métrica	Cálculo	Pregunta veterinaria	Advertencia
Sensibilidad	VP / (VP + FN)	De los animales enfermos ¿cuántos detecta?	Alta sensibilidad no garantiza pocos falsos positivos
Especificidad	VN / (VN + FP)	De los animales sanos ¿cuántos descarta?	Depende del umbral y del espectro clínico
VPP	VP / (VP + FP)	Si alerta ¿qué probabilidad hay de evento?	Muy sensible a prevalencia y selección
VPN	VN / (VN + FN)	Si no alerta ¿qué probabilidad hay de ausencia?	Puede ser alto solo porque el evento es raro
Razón de verosimilitud positiva	Sensibilidad / (1 - especificidad)	¿Cuánto aumenta la evidencia con un positivo?	Debe combinarse con probabilidad preprueba
Razón de verosimilitud negativa	(1 - sensibilidad) / especificidad	¿Cuánto reduce la evidencia con un negativo?	No expresa directamente riesgo posterior

Ejemplo: en 1.000 vacas con prevalencia de mastitis clínica de 5%, una prueba con sensibilidad de 90% y especificidad de 90% detectaría aproximadamente 45 casos, omitiría 5 y generaría 95 falsos positivos. El VPP sería cercano a 32%. La aparente simetría del 90% no evita que dos de cada tres alertas sean falsas. La utilidad dependerá del costo de revisar una vaca, del daño de retrasar el tratamiento y de la capacidad operativa del hato.

14.4 Curvas ROC y área bajo la curva

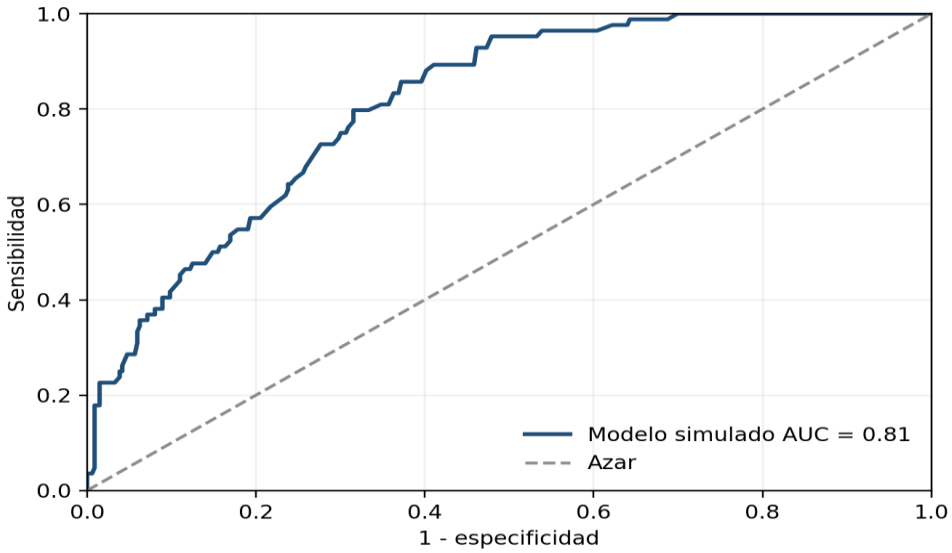
La curva ROC representa sensibilidad frente a 1 menos especificidad a través de todos los umbrales. El área bajo la curva puede interpretarse como la probabilidad de que un animal elegido al azar entre quienes presentan el evento reciba un puntaje mayor que otro elegido entre quienes no lo presentan. Evalúa ordenamiento, no exactitud de probabilidades.

Una AUC de 0,80 no significa 80% de diagnósticos correctos ni garantiza utilidad clínica. Dos modelos con la misma AUC pueden comportarse de modo distinto cerca del umbral relevante. También es posible que una variable mejore ligeramente la AUC sin modificar decisiones. Deben presentarse intervalos de confianza y, al comparar modelos sobre los mismos pacientes, utilizar métodos pareados o remuestreo.

La AUC puede resultar optimista desde el punto de vista práctico en eventos infrecuentes, porque incorpora la especificidad sobre una gran cantidad de negativos. Las curvas precisión-recall suelen revelar mejor el costo de los falsos positivos en ese escenario. Para pruebas de tamizaje debe examinarse el tramo de alta sensibilidad; para confirmar enfermedad puede interesar alta especificidad.

Figura 14.1

Curva ROC del caso didáctico simulado



Nota. Datos simulados para fines didácticos; la curva no representa el desempeño de un sistema clínico real.

14.5 Precision recall y puntuación F1

En la terminología de aprendizaje automático, precision equivale al valor predictivo positivo y recall a la sensibilidad. La puntuación F1 es la media armónica de ambas: penaliza que una sea alta y la otra baja. Estas métricas son útiles cuando la clase positiva es infrecuente y el interés se concentra en recuperarla, pero F1 ignora los verdaderos negativos y presupone implícitamente un equilibrio entre precision y recall que puede no corresponder al problema clínico.

La curva precision-recall y su área deben interpretarse contra una línea basal relacionada con la prevalencia. Si la prevalencia es 5%, una precision de 20% representa una mejora respecto al azar, aunque todavía implica cuatro falsas alertas por cada verdadera. Comparar áreas entre poblaciones con prevalencias distintas puede ser engañoso.

Existen variantes F-beta que asignan más peso al recall cuando omitir casos es costoso, o a precision cuando las intervenciones innecesarias son dañinas. En un sistema de alerta de torsión gástrica podría priorizarse sensibilidad; en una recomendación de cirugía invasiva, la tolerancia a falsos positivos sería mucho menor. La selección debe documentarse antes de examinar los resultados.

14.6 Error absoluto error cuadrático y coeficiente de determinación

Los desenlaces continuos requieren métricas coherentes con la escala y el costo del error. El error absoluto medio (MAE) promedia la magnitud de las diferencias entre valores observados y predichos. Se expresa en la unidad original, por ejemplo litros de leche por día, kilogramos de peso o minutos de recuperación anestésica. Su interpretación es directa y cada error contribuye de manera lineal.

El error cuadrático medio (MSE) eleva al cuadrado las diferencias y penaliza con fuerza los errores grandes. La raíz del MSE (RMSE) regresa a la unidad original. Resulta útil cuando desviaciones extremas son especialmente dañinas, pero también puede quedar dominada por pocos valores atípicos. Por ello conviene reportar MAE y RMSE juntos, acompañados de gráficos de residuos.

El coeficiente de determinación R^2 expresa la reducción relativa del error cuadrático frente a predecir la media. En validación externa puede ser negativo: indica que el modelo es peor que esa predicción ingenua. Un R^2 alto no demuestra ausencia de sesgo; el modelo podría sobreestimar sistemáticamente en animales de bajo peso y subestimar en los de alto peso. Deben examinarse calibración de intercepto y pendiente para resultados continuos, distribución de residuos y error por rangos clínicos.

Tabla 14.3

Métricas para predicción continua

Métrica	Ventaja	Limitación	Ejemplo de lectura
MAE	Unidad original y robustez relativa	No penaliza de modo especial errores extremos	MAE 1,8 L significa desviación absoluta media de 1,8 L/día
MSE	Penaliza errores grandes	Unidad al cuadrado y sensibilidad a atípicos	Útil para optimización pero menos intuitivo
RMSE	Unidad original y énfasis en errores grandes	Puede ocultar dirección del sesgo	RMSE 2,6 kg en predicción de peso
R^2 externo	Compara con predecir la media	No mide calibración ni relevancia biológica	R^2 negativo indica peor desempeño que la media
Error porcentual	Facilita comparar escalas	Inestable cerca de cero	No usar sin revisar ceros y valores pequeños

14.7 Calibración y utilidad clínica

La calibración describe la concordancia entre probabilidades predichas y frecuencias observadas. Si se asigna 20% de riesgo a un grupo de perros, aproximadamente 20 de cada 100 deberían presentar el desenlace dentro del horizonte definido. Esta propiedad es indispensable cuando la probabilidad desencadena una decisión. Un modelo con buena discriminación puede estar mal calibrado y conducir a sobretratamiento o infratratamiento (Van Calster et al., 2019).

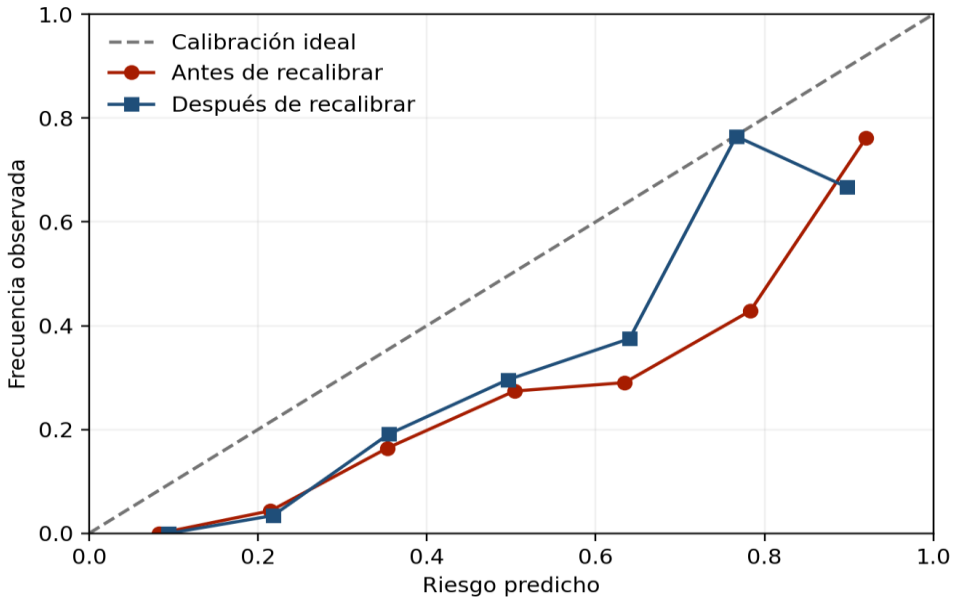
La calibración en el promedio compara riesgo medio predicho y frecuencia observada. El intercepto ideal es 0: un valor negativo suele indicar sobreestimación global y uno positivo, subestimación. La pendiente ideal es 1. Una pendiente menor que 1 sugiere predicciones demasiado extremas, hallazgo compatible con sobreajuste; una pendiente mayor que 1 indica poca dispersión. Una curva suavizada permite localizar regiones de riesgo con desajuste. El test de Hosmer-Lemeshow no debe ser la única evaluación: depende del tamaño muestral y de agrupaciones arbitrarias.

El puntaje de Brier promedia el error cuadrático entre probabilidad y desenlace binario. Combina discriminación y calibración, pero su magnitud depende de la prevalencia. Debe compararse con un modelo de referencia y, cuando sea útil, descomponerse o escalarse.

La utilidad clínica no se deduce de la calibración. El análisis de curva de decisión calcula beneficio neto para un rango de umbrales y compara el modelo con estrategias de intervenir a todos o a ninguno (Vickers & Elkin, 2006). El umbral representa una relación entre daño de falso positivo y daño de falso negativo. Su elección exige participación de clínicos, responsables del animal y gestores; no es un parámetro puramente estadístico.

Figura 14.2

Calibración antes y después de una actualización simulada



Nota. Datos simulados. La recalibración puede corregir riesgos absolutos sin alterar el orden relativo de los pacientes.

14.8 Desbalance de clases y selección de umbrales

El desbalance aparece cuando una clase es mucho menos frecuente. No constituye por sí mismo un defecto: puede reflejar la epidemiología real. El problema surge cuando la exactitud domina la evaluación. Con prevalencia de 2%, clasificar a todos los animales como negativos produce 98% de exactitud y ninguna capacidad de detección.

La estrategia debe comenzar por definir la métrica y el diseño de muestreo. La ponderación de clases, el submuestreo, el sobremuestreo y los métodos sintéticos pueden ayudar durante el entrenamiento, pero deben aplicarse solo dentro de cada pliegue. Alterar la proporción de clases modifica la interpretación probabilística; las predicciones pueden requerir recalibración a la prevalencia objetivo. El conjunto de prueba debe conservar la distribución de la población de uso, salvo que el objetivo del estudio sea otro y se corrija analíticamente.

El umbral no debe fijarse automáticamente en 0,50 ni seleccionarse solo por el índice de Youden. Debe derivarse del propósito: tamizaje, confirmación, priorización o asignación de recursos. Para cada umbral conviene presentar sensibilidad, especificidad, VPP, VPN, cantidad de alertas por 100 animales, falsos negativos y beneficio neto. Un rango de umbrales clínicamente razonables es más informativo que un punto óptimo inestable.

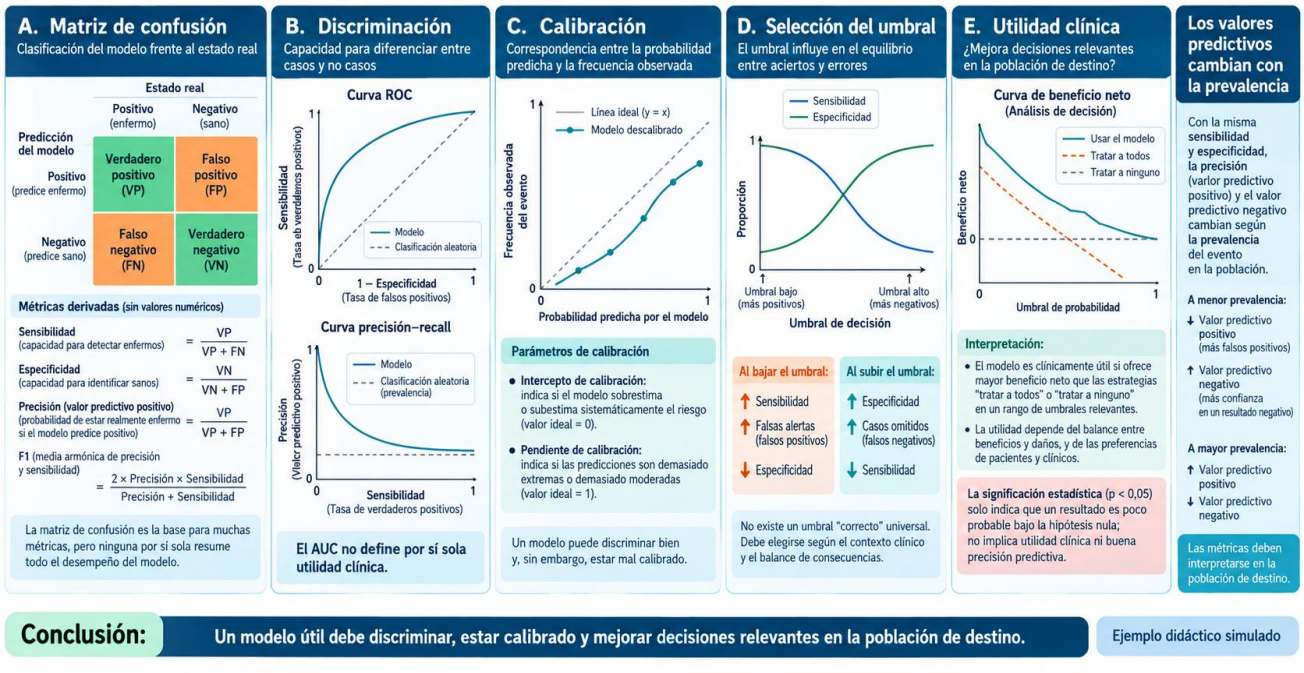
Tabla 14.4

Selección de umbral según la decisión

Uso	Error más costoso	Orientación	Información adicional
Tamizaje de sepsis	Falso negativo	Umbral bajo y alta sensibilidad	Capacidad de confirmación y carga de alertas
Indicación de procedimiento invasivo	Falso positivo	Umbral alto y alta especificidad	Complicaciones y alternativas
Priorización de revisión radiográfica	Ambos según flujo	Umbral adaptado a capacidad	Tiempo de espera y prevalencia por turno
Alerta de mastitis en hato	Fatiga por falsas alertas y caso omitido	Dos niveles de alerta	Costo de leche descartada y revisión

Figura 14.2.

Calibración antes y después de una actualización simulada



14.9 Importancia de variables

La importancia de variables responde qué información contribuye al desempeño del modelo, pero su significado depende del método. En árboles, la reducción de impureza puede favorecer variables continuas o con muchas categorías. La importancia por permutación mide cuánto empeora una métrica al desordenar una variable, aunque distribuye o distorsiona importancia cuando los predictores están correlacionados. En regresión, la magnitud de coeficientes solo es comparable si las escalas y codificaciones son consideradas.

Importancia predictiva no equivale a causalidad, mecanismo ni recomendación terapéutica. Si la administración de un fármaco aparece como predictor de mortalidad, puede actuar como marcador de gravedad por confusión por indicación. Del mismo modo, el código de la clínica puede capturar diferencias de referencia, equipos o registro. Eliminarlo no elimina necesariamente el sesgo; otras variables pueden funcionar como sustitutos.

Debe distinguirse importancia global —promedio en la población— de influencia local en un individuo. Es recomendable reportar estabilidad mediante remuestreo, contrastar varios métodos y evaluar si el resultado se conserva en centros externos. Una lista ordenada sin incertidumbre puede sugerir una precisión que los datos no sostienen.

14.10 SHAP y otros métodos de explicabilidad

Los valores SHAP distribuyen la diferencia entre una predicción y un valor basal entre las variables, apoyándose en ideas de teoría de juegos (Lundberg & Lee, 2017). Permiten resúmenes globales, gráficos de dependencia y explicaciones locales. Sin embargo, dependen de la población de referencia, del tratamiento de la dependencia entre predictores y de la versión del modelo. Dos implementaciones pueden producir atribuciones distintas.

Una explicación local señala qué variables empujaron una predicción hacia arriba o abajo; no prueba que modificar esas variables cambie el desenlace. Tampoco garantiza que el modelo haya aprendido un mecanismo clínico válido. Ghassemi et al. (2021) advierten que la explicabilidad post hoc puede ofrecer una apariencia de transparencia sin resolver problemas de causalidad, seguridad o calidad de datos.

Otros métodos incluyen LIME, permutación, gráficos de dependencia parcial, perfiles de efectos acumulados, contrafactuales y mapas de activación en imágenes. Los mapas de calor deben comprobarse con pruebas de oclusión y evaluación experta: una región coloreada no certifica razonamiento anatómico. Los contrafactuales deben respetar restricciones biológicas; no es útil sugerir que un perro reduzca su edad o cambie de raza para disminuir riesgo.

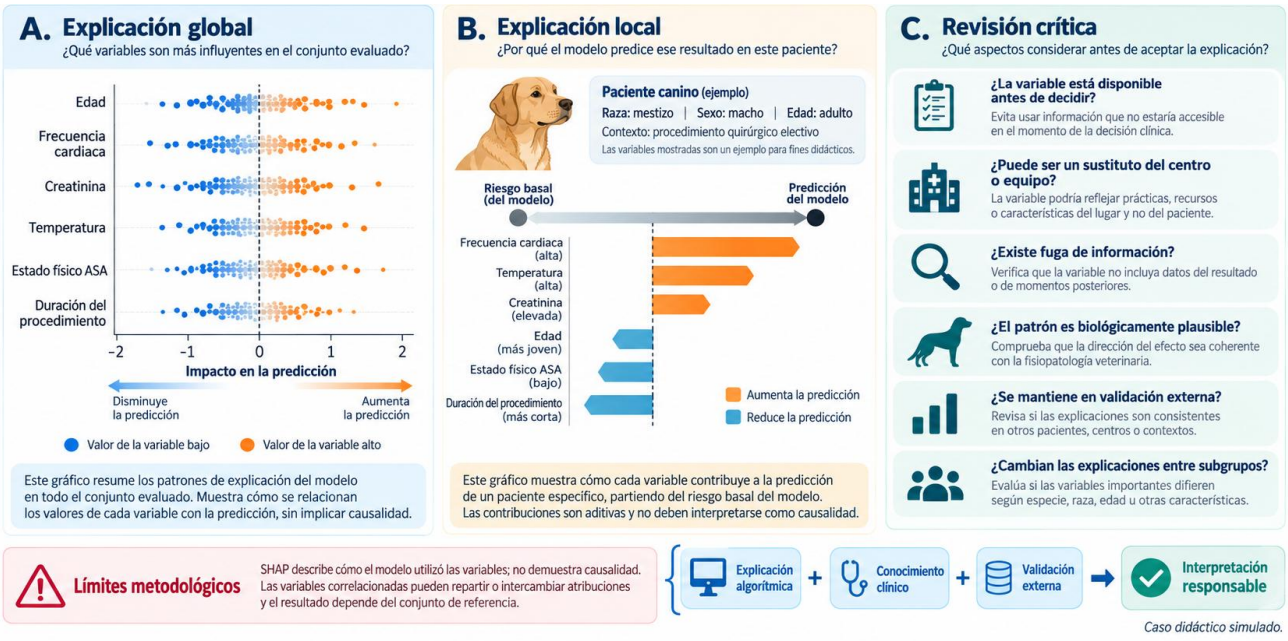
Tabla 14.5

Métodos de interpretación y principales cautelas

Método	Alcance	Pregunta	Cautela
Coeficientes	Global	¿Cómo cambia la predicción al variar un predictor?	Escala interacciones y no linealidad
Permutación	Global	¿Cuánto cae el rendimiento si se rompe esta información?	Predictores correlacionados
SHAP	Local y global	¿Cómo se distribuye la predicción respecto al basal?	Dependencia y población de referencia
Dependencia parcial	Global	¿Cuál es la respuesta promedio a una variable?	Combinaciones irreales si hay correlación
Efectos acumulados	Global	¿Cómo varía la predicción en regiones observadas?	Interpretación menos intuitiva
LIME	Local	¿Qué aproximación simple explica este caso?	Inestabilidad del vecindario
Grad CAM u oclusión	Local en imágenes	¿Qué regiones influyen?	Atención visual no equivale a lesión causal

Figura 14.2.

Interpretación global e individual mediante SHAP



14.11 Robustez y análisis de sensibilidad

Robustez es la capacidad de conservar un comportamiento aceptable frente a perturbaciones plausibles. Deben evaluarse cambios en unidades, redondeo, valores faltantes, ruido de medición, dispositivos, iluminación, prevalencia y criterios de inclusión. En radiología, pequeñas variaciones de contraste, colimación o marcadores pueden alterar un modelo; en laboratorio, un cambio de analizador puede desplazar distribuciones.

El análisis de sensibilidad modifica decisiones analíticas razonables y observa si cambian las conclusiones. Ejemplos: imputación múltiple frente a casos completos, definición estricta frente a amplia del desenlace, exclusión de una clínica, ventanas temporales alternativas, diferentes semillas, tratamiento de valores extremos y recalibración. El objetivo no es elegir la versión que produce el mejor resultado, sino identificar dependencias frágiles.

También se requieren pruebas de estrés. Un modelo puede evaluarse en subpoblaciones con mayor enfermedad, datos incompletos o equipos menos frecuentes. Las perturbaciones adversariales deliberadas deben ser clínicamente plausibles; alterar imperceptiblemente píxeles es relevante para seguridad informática, pero no reemplaza simular rotación, compresión, artefactos de movimiento o proyecciones deficientes.

14.12 Comparación con el desempeño profesional

Comparar IA y profesionales exige que ambos resuelvan la misma tarea con información equivalente y bajo condiciones compatibles. Un modelo con acceso a metadatos, evolución o resultados de laboratorio no debe compararse con veterinarios que solo ven una imagen. También deben definirse experiencia, especialidad, tiempo disponible, instrucciones, prevalencia y estándar de referencia.

El diseño cruzado permite que cada caso sea evaluado por modelo y profesionales. Las comparaciones deben ser pareadas y considerar que varios lectores examinan los mismos casos. Además de sensibilidad y especificidad, interesan tiempo, concordancia, calibración de confianza, variabilidad entre evaluadores y errores clínicamente graves. El estándar de referencia debe ser independiente; si el consenso de lectores sirve como verdad, comparar el algoritmo con esos mismos lectores introduce circularidad.

La pregunta más útil suele ser si la combinación profesional-modelo supera al profesional sin asistencia. La asistencia puede mejorar sensibilidad y reducir tiempo, pero también provocar sesgo de automatización: aceptación acrítica de una sugerencia errónea. Un estudio de impacto debe observar cómo cambia la decisión, no solo presentar una competencia retrospectiva.

Tabla 14.6

Diseños de comparación profesional modelo

Diseño	Ventaja	Amenaza	Análisis
Mismos casos sin asistencia	Comparación directa	Información o tiempo desiguales	Pruebas pareadas y efectos por lector/caso
Lectura antes y después de IA	Mide cambio individual	Memoria y efecto de orden	Aleatorizar orden y periodo de lavado
Cruzado asistido no asistido	Estima efecto de asistencia	Contaminación entre fases	Modelo mixto con lector y caso
Implementación por clínica	Evalúa flujo real	Confusión temporal y organizacional	Ensayo por conglomerados o serie temporal

14.13 Validez clínica y biológica

La validez clínica exige correspondencia entre población, momento de predicción, desenlace, horizonte y acción. Predecir mortalidad después de disponer de información obtenida durante la hospitalización no ayuda a la decisión de ingreso. Un biomarcador medido después del tratamiento no puede emplearse en una herramienta pretratamiento. El desempeño debe juzgarse en el punto real del flujo clínico.

La validez biológica pregunta si las relaciones son compatibles con conocimiento fisiopatológico y medición confiable. No obliga a que un modelo complejo replique una teoría previa, pero sí demanda investigar asociaciones imposibles, cambios abruptos y dependencias de variables espurias. La dirección de un efecto puede cambiar por colinealidad o interacción; por ello, la revisión experta debe realizarse junto con diagnósticos estadísticos.

La relevancia clínica tampoco coincide con significación estadística. Una mejora de AUC de 0,79 a 0,80 puede ser precisa y, sin embargo, no cambiar decisiones. Por el contrario, una mejora localizada en el rango de riesgo donde se decide hospitalizar puede ser valiosa aunque el cambio global sea pequeño. Deben considerarse bienestar animal, iatrogenia, costos, equidad de acceso y carga para el personal.

14.14 Evaluación de impacto en condiciones reales

Después de la validación externa viene la evaluación de impacto: demostrar que usar el modelo mejora el proceso o los resultados frente a una alternativa. Los estudios retrospectivos de desempeño no responden esta pregunta. DECIDE-AI propone una evaluación temprana, prospectiva y por etapas de sistemas de apoyo a decisiones basados en IA, prestando atención a factores humanos y flujo de trabajo (Vasey et al., 2022).

Los desenlaces pueden ser clínicos, productivos, operativos y económicos: mortalidad, complicaciones, dolor, uso de antimicrobianos, tiempo hasta diagnóstico, rendimiento, falsas alertas, carga de trabajo y costos. La selección debe evitar indicadores sustitutos que mejoran sin beneficiar al animal. Un sistema que acelera la lectura radiográfica pero aumenta errores graves no puede considerarse exitoso.

Los diseños incluyen ensayo aleatorizado individual, ensayo por conglomerados, diseño escalonado, estudio antes-después con control, serie temporal interrumpida y despliegue silencioso. Este último ejecuta el modelo sin mostrar resultados para comprobar estabilidad y logística antes de influir en decisiones. La evaluación debe documentar adherencia: una herramienta precisa puede no tener impacto si las alertas llegan tarde, no se comprenden o no existe capacidad de respuesta.

El análisis debe considerar efectos no deseados: desplazamiento de tareas, fatiga por alertas, dependencia, inequidad entre clínicas y cambios de conducta que alteran los datos futuros. El sistema socio-técnico —personas, protocolos, interfaz y recursos— es la intervención real, no el algoritmo aislado.

14.15 Monitoreo y deterioro del modelo

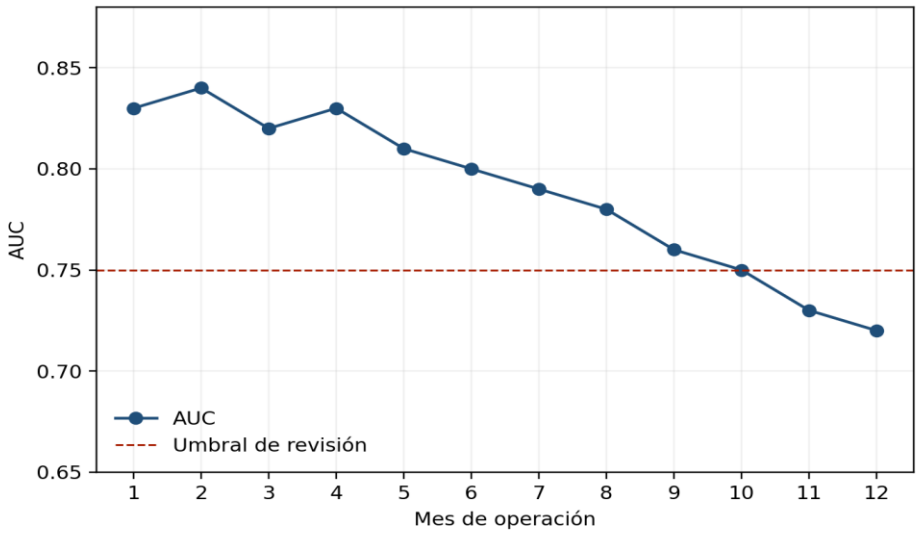
Un modelo validado no queda certificado de forma permanente. La deriva de datos ocurre cuando cambia la distribución de predictores; la deriva de prevalencia modifica la frecuencia del desenlace; la deriva conceptual altera la relación entre predictores y resultado. Nuevas vacunas, tratamientos, razas predominantes, protocolos, equipos o políticas de registro pueden producir estos cambios.

El monitoreo debe definir responsables, frecuencia, métricas, umbrales de alerta y acciones. Incluye integridad de datos, proporción de valores faltantes, rangos, distribución de variables, tasa de alertas, discriminación, calibración, errores por subgrupo y resultados clínicos. Las métricas de desempeño requieren desenlaces maduros; mientras llegan, pueden vigilarse indicadores de entrada y proceso.

No todo cambio justifica reentrenar. Primero debe investigarse si existe un error de interfaz, codificación o medición. Las opciones incluyen recalibrar el intercepto, ajustar intercepto y pendiente, reestimar coeficientes, incorporar nuevos predictores o desarrollar otro modelo. Cada actualización debe versionarse, probarse y aprobarse. El reentrenamiento automático sin supervisión puede introducir inestabilidad y ocultar degradación.

Figura 14.3

Ejemplo simulado de deterioro de discriminación durante la operación



Nota. Una señal estadística inicia revisión; no determina por sí sola la causa ni autoriza reentrenamiento automático.

Tabla 14.7

Plan mínimo de vigilancia

Dominio	Indicador	Frecuencia	Acción ante alerta
Entrada	Valores faltantes rangos equipo y codificación	Semanal o mensual	Auditar captura e interfaz
Población	Edad especie raza centro prevalencia	Mensual o trimestral	Evaluar cambio de mezcla
Rendimiento	AUC sensibilidad VPP MAE	Según maduración del desenlace	Validación focal y revisión de umbral
Calibración	Intercepto pendiente Brier y curva	Trimestral o semestral	Considerar recalibración
Equidad	Métricas por subgrupo relevante	Trimestral	Investigar datos y proceso
Impacto	Resultados daño carga y costos	Semestral o anual	Modificar o suspender uso

14.16 Caso didáctico validación externa de un modelo de riesgo clínico

Caso didáctico simulado. Todos los datos y resultados de esta sección fueron creados con fines educativos y no corresponden a pacientes reales ni a un producto disponible.

Problema y pregunta

Una red de hospitales veterinarios desarrolló un modelo de regresión logística para predecir, al momento del ingreso, el riesgo de complicación mayor en perros sometidos a cirugía abdominal. Se consideran complicación mayor la reintervención, ingreso no planificado a cuidados intensivos o muerte durante los siete días posteriores. El modelo original emplea edad, categoría ASA, frecuencia cardiaca, lactato, hematocrito, tipo de cirugía y carácter urgente.

La pregunta es si las probabilidades conservan discriminación, calibración y utilidad clínica en un hospital ecuatoriano que no participó en el desarrollo. La hipótesis operativa es que el modelo discriminará mejor que una regla basada solo en ASA, pero podría sobreestimar el riesgo debido a diferencias de referencia y prevalencia.

Diseño población y unidad de análisis

Se plantea una cohorte prospectiva consecutiva. La población objetivo son perros sometidos a cirugía abdominal durante doce meses. La unidad de análisis es el procedimiento índice por paciente; si un perro tiene reintervención durante el seguimiento, esta forma parte del desenlace y no una nueva observación. Se excluyen expedientes sin posibilidad de establecer el desenlace, pero no los casos con predictores faltantes, que se manejarán con el método definido antes del análisis.

Se incluyen 420 perros, 84 con complicación mayor y 336 sin ella. El modelo permanece congelado. Las variables se codifican con las mismas unidades y definiciones del desarrollo. El equipo de validación no modifica coeficientes ni selecciona el umbral a partir de los resultados externos.

Plan de análisis preespecificado

1. Describir elegibilidad, flujo de pacientes, variables, datos faltantes y diferencias respecto a la cohorte original.
2. Aplicar el modelo original y obtener probabilidades sin reentrenamiento.
3. Estimar AUC con intervalo de confianza mediante bootstrap estratificado.
4. Evaluar calibración en el promedio, intercepto, pendiente, curva suavizada y puntaje de Brier.
5. Calcular sensibilidad, especificidad, VPP y VPN en umbrales predefinidos de 10%, 20% y 30%.
6. Comparar beneficio neto con intervenir a todos, a ninguno y con la regla ASA.
7. Realizar análisis por urgencia y categoría de peso, presentando incertidumbre y evitando pruebas exploratorias concluyentes.
8. Si la discriminación se conserva y existe descalibración global, evaluar recalibración del intercepto como análisis secundario.

Resultados simulados

Tabla 14.8

Resultados simulados de validación externa

Dominio	Estimación simulada	Interpretación
Discriminación	AUC 0,81; IC 95% 0,76 a 0,86	Buen ordenamiento relativo con incertidumbre moderada
Calibración promedio	Riesgo medio 24%; observado 20%	Sobreestimación global de 4 puntos porcentuales
Intercepto	-0,31; IC 95% -0,52 a -0,10	Predicciones globalmente altas
Pendiente	0,82; IC 95% 0,65 a 0,99	Riesgos algo extremos y posible sobreajuste
Brier	0,139 frente a 0,160 del modelo nulo	Mejora global modesta
Umbral 20%	Sens. 0,76; espec. 0,72; VPP 0,40; VPN 0,92	Útil para descartar; muchas alertas no tendrán evento
Recalibración	Intercepto ajustado; AUC sin cambio	Mejora riesgos absolutos pero no ordenamiento

La AUC sugiere que el modelo conserva capacidad de ordenamiento. No obstante, el intercepto negativo y la curva de calibración muestran sobreestimación. En términos prácticos, algunos

perros serían clasificados por encima del umbral de vigilancia intensiva sin que su riesgo real lo justifique. La pendiente menor que 1 indica que las predicciones extremas requieren cautela.

En el umbral del 20%, el VPN de 92% parece alto, pero debe interpretarse junto con prevalencia de 20%. De 100 perros con una composición semejante, el modelo alertaría aproximadamente a 38; cerca de 15 presentarían complicación y 23 serían falsos positivos. Entre los 62 no alertados, unos 5 desarrollarían el evento. La aceptabilidad depende de si la vigilancia adicional es poco invasiva y de si el hospital puede atender 38 alertas por cada 100 cirugías.

El análisis de decisión simulado muestra beneficio neto frente a intervenir a todos o a ninguno entre umbrales de 12% y 28%. Fuera de ese rango no se demuestra ventaja. En cirugías urgentes la calibración empeora, pero el intervalo es amplio. Este hallazgo se considera señal para una validación dirigida, no evidencia definitiva de inequidad.

Decisión y actualización

El modelo no debería desplegarse con las probabilidades originales. Dado que conserva discriminación y la mayor desviación es global, se propone recalibrar el intercepto con datos locales y mantener los coeficientes. La versión recalibrada debe etiquetarse como una nueva versión, evaluarse prospectivamente en modo silencioso y someterse después a un estudio de impacto. No sería metodológicamente correcto presentar la misma cohorte usada para recalibrar como confirmación externa final.

Código orientativo en R

El siguiente esquema ilustra la evaluación. Los nombres de variables deben adaptarse al diccionario real. Los paquetes y versiones deben registrarse en el protocolo reproducible.

```
# Datos externos: desenlace (0/1) y probabilidad del modelo congelado
library(pROC)
library(rms)
library(rmda)

roc_ext <- roc(datos$complicacion, datos$prob_original, ci = TRUE)
auc(roc_ext); ci.auc(roc_ext)

# Calibración débil: intercepto y pendiente
lp <- qlogis(pmin(pmax(datos$prob_original, 1e-6), 1 - 1e-6))
cal_int <- glm(complicacion ~ offset(lp), family = binomial(), data = datos)
```

```
cal_slope <- glm(complicacion ~ lp, family = binomial(), data = datos)

# Métricas en un umbral clínicamente predefinido
u <- 0.20
pred <- datos$prob_original >= u
tab <- table(predicho = pred, observado = datos$complicacion)

# Recalibración del intercepto como análisis secundario
datos$prob_recal <- plogis(coef(cal_int)[1] + lp)
```

Limitaciones del caso

- El número de eventos permite una evaluación inicial, pero limita análisis finos por subgrupo.
- El desenlace compuesto reúne eventos con diferente gravedad y puede ocultar patrones.
- El seguimiento de siete días no capta complicaciones tardías.
- La recalibración local puede no transportarse a hospitales con otros flujos de referencia.
- La utilidad calculada supone una relación de costos implícita en cada umbral; debe discutirse con el equipo clínico.

14.17 Síntesis autoevaluación y actividad práctica

Síntesis

La evaluación rigurosa de un modelo veterinario combina varias preguntas. La validación interna cuantifica optimismo; la externa estudia transportabilidad. La discriminación ordena, la calibración da significado a los riesgos y el análisis de decisión conecta predicciones con consecuencias. Las métricas de clasificación dependen del umbral y, en el caso de los valores predictivos, de la prevalencia. Las métricas de regresión deben expresarse en unidades clínicamente comprensibles y complementarse con residuos y calibración.

La interpretación de variables no establece causalidad. SHAP, LIME o mapas de activación pueden ayudar a auditar el comportamiento, pero no convierten un modelo opaco en evidencia mecanística. La comparación con profesionales debe ser justa y la evaluación decisiva es la del sistema asistido en condiciones reales. Finalmente, todo despliegue requiere vigilancia, versionado, umbrales de revisión y posibilidad de suspender el sistema.

Lista de comprobación para evaluar un modelo veterinario

Pregunta	Sí	No o incierto
¿Población momento horizonte y desenlace coinciden con el uso previsto?	□	□
¿El conjunto externo es realmente independiente?	□	□
¿La separación respeta animal finca centro y tiempo?	□	□
¿Se reportan discriminación y calibración con incertidumbre?	□	□
¿Los umbrales se justifican por consecuencias clínicas?	□	□
¿Se presentan falsos positivos y negativos en números absolutos?	□	□
¿Se estudia desempeño por subgrupos biológicamente relevantes?	□	□
¿La explicabilidad se interpreta sin inferencias causales indebidas?	□	□
¿Existe comparación con práctica habitual y evaluación de impacto?	□	□
¿Hay plan de monitoreo actualización versionado y suspensión?	□	□

Preguntas de autoevaluación

1. Explique por qué una división aleatoria por imagen puede producir fuga de información cuando cada animal aporta varias radiografías.
2. Diferencie validación interna, temporal y geográfica mediante un ejemplo de mastitis bovina.
3. ¿Por qué el VPP cambia cuando se aplica el mismo modelo en una población con menor prevalencia?
4. ¿Qué interpreta realmente una AUC de 0,82 y qué no permite concluir?
5. ¿Cuándo resulta más informativa una curva precision-recall que una curva ROC?
6. Compare MAE y RMSE en una predicción de producción de leche.
7. ¿Cómo se interpretan un intercepto de calibración negativo y una pendiente menor que 1?
8. ¿Por qué el umbral 0,50 puede ser inadecuado para una alerta de sepsis?
9. Diferencie importancia predictiva de efecto causal.
10. Mencione tres limitaciones de SHAP para justificar una decisión clínica.
11. Diseñe una prueba de robustez para un clasificador de radiografías caninas.
12. ¿Qué condiciones debe cumplir una comparación justa entre IA y radiólogos veterinarios?
13. Diferencie validez estadística, clínica y biológica.
14. ¿Qué aporta un estudio de impacto que no aporta una validación externa retrospectiva?
15. Proponga cuatro indicadores para detectar deterioro de un modelo en operación.

Orientaciones de respuesta

1. Las imágenes del mismo animal comparten anatomía, técnica y artefactos; deben mantenerse en un solo pliegue.
2. Interna: misma fuente con remuestreo; temporal: periodo posterior; geográfica: otra finca o región.
3. Porque aumenta la proporción de falsos positivos respecto a verdaderos positivos aun si sensibilidad y especificidad se conservan.
4. Expresa ordenamiento entre un caso y un no caso; no es exactitud, calibración ni utilidad.
5. Cuando la clase positiva es rara y el costo de falsas alertas es central.
6. MAE resume error típico lineal; RMSE penaliza más los errores grandes.
7. Sobreestimación global y predicciones excesivamente extremas, respectivamente.
8. Porque el costo relativo de errores y la prevalencia pueden exigir un umbral mucho menor o mayor.
9. La primera describe contribución a predicción; la segunda requiere un estimando y supuestos causales.
10. Dependencia de referencia, correlación entre predictores y ausencia de interpretación causal.
11. Variar proyección, exposición, compresión, equipo y artefactos plausibles, con evaluación por subgrupo.
12. Mismos casos, información, tiempo, estándar independiente y análisis pareado con efectos de lector y caso.
13. Rendimiento cuantitativo; beneficio en decisiones; coherencia con fisiopatología y medición.
14. Mide cambios en decisiones, resultados, flujo, daño y adopción al usar el sistema.
15. Distribución de entradas, faltantes, tasa de alertas, calibración, discriminación y errores por subgrupo.

Actividad práctica

Título: Protocolo de validación externa y vigilancia de un modelo veterinario

Trabaje con un conjunto de datos propio anonimizado o con datos simulados. Elija un modelo de clasificación o regresión relacionado con medicina clínica, producción, epidemiología o bienestar animal. No utilice datos identificables ni copie registros clínicos en servicios externos sin autorización.

Tabla 14.9

Guía de la actividad práctica

Fase	Producto esperado	Criterio de calidad
1 Definición	Población desenlace horizonte unidad y momento	Uso clínico inequívoco
2 Datos	Diagrama de flujo y diccionario	Independencia y trazabilidad
3 Plan	Métricas umbrales subgrupos y faltantes	Preespecificación antes de resultados
4 Validación	Discriminación calibración y utilidad	IC y gráficos apropiados
5 Interpretación	Errores importancia robustez y validez biológica	Sin inferencias causales indebidas
6 Implementación	Comparador impacto y flujo de trabajo	Consecuencias y factores humanos
7 Monitoreo	Tablero periodicidad responsables y acciones	Versionado y criterio de suspensión

El informe final debe incluir un resumen de 250 palabras, protocolo, descripción de datos, resultados con dos tablas y dos figuras, interpretación clínica, limitaciones, plan de impacto y plan de monitoreo. Como ejercicio de transparencia, adjunte el código y una ficha de versión del modelo. Evalúe el informe con cuatro dimensiones de 0 a 3 puntos: rigor metodológico, interpretación clínica, reproducibilidad y comunicación de incertidumbre.

CAPÍTULO 15

REPRODUCIBILIDAD, COMUNICACIÓN CIENTÍFICA Y FUTURO DE LA IA VETERINARIA

Introducción

Un modelo puede alcanzar una excelente área bajo la curva y, aun así, carecer de valor científico si nadie logra reconstruir cómo se obtuvieron los datos, qué transformaciones se aplicaron, qué versión del código produjo las métricas o en qué población veterinaria podría utilizarse. La reproducibilidad no es una tarea administrativa posterior al análisis: es una propiedad del proceso de investigación. Abarca decisiones sobre organización, documentación, software, control de versiones, disponibilidad responsable de datos, reporte transparente y vigilancia después de la implementación.

Este capítulo cierra la obra integrando tres planos. El primero es operativo: cómo convertir un proyecto ejecutado en R en un flujo rastreable y repetible. El segundo es comunicacional: cómo describir modelos, incertidumbre, sesgos y utilidad sin reducir el artículo a una lista de métricas. El tercero es prospectivo: qué puede esperarse de los modelos multimodales, los gemelos digitales, la robótica y la medicina veterinaria de precisión, y qué condiciones científicas, económicas y profesionales deberán cumplirse para que esas tecnologías beneficien a animales, responsables, productores y sistemas de salud.

Objetivos de aprendizaje

1. Distinguir repetibilidad computacional, reproducibilidad analítica, replicabilidad científica y generalización externa.
2. Organizar proyectos veterinarios de datos mediante estructuras separadas para datos originales, derivados, código, resultados y documentación.
3. Construir flujos reproducibles con R, RStudio, scripts, cuadernos, reportes dinámicos y entornos de dependencias.
4. Aplicar control de versiones y principios FAIR a datos, código, modelos y metadatos sin confundir FAIR con apertura irrestricta.
5. Comunicar el desarrollo, validación y desempeño de algoritmos con transparencia clínica, biológica y estadística.
6. Seleccionar guías de reporte según el tipo de estudio y efectuar una revisión crítica por pares.

7. Planificar la transferencia de un modelo a la práctica considerando impacto, costos, infraestructura, equidad, seguridad y mantenimiento.
8. Analizar posibilidades y límites de modelos multimodales, gemelos digitales, robótica y sistemas autónomos.
9. Formular una agenda regional y desarrollar un proyecto veterinario asistido por IA desde la pregunta hasta la comunicación de resultados.

15.1. Investigación reproducible y ciencia abierta

La reproducibilidad comienza con una distinción conceptual. La repetibilidad computacional significa que el mismo equipo, con los mismos datos, instrucciones y entorno, obtiene el mismo resultado. La reproducibilidad analítica exige que otro equipo pueda ejecutar el procedimiento descrito y reconstruir tablas, figuras y métricas. La replicabilidad va más lejos: un estudio independiente, con nuevos animales o unidades productivas, encuentra un patrón compatible. La generalización pregunta si el modelo conserva utilidad en otra clínica, región, raza, especie, equipo diagnóstico o periodo. Estos niveles no son intercambiables (National Academies of Sciences, Engineering, and Medicine, 2019). Un script que devuelve exactamente el AUC publicado demuestra repetibilidad, no validez externa.

La crisis de reproducibilidad no se resuelve depositando un archivo al final del proyecto. Las decisiones que más afectan la trazabilidad suelen tomarse al inicio: definición del desenlace, identificación de la unidad experimental, reglas de inclusión, manejo de datos faltantes, prevención de fuga de información y separación temporal o geográfica de los conjuntos. En una investigación sobre mastitis bovina, por ejemplo, dividir registros por filas puede colocar lactancias o mediciones de la misma vaca tanto en entrenamiento como en prueba. El código será perfectamente ejecutable, pero la estimación de desempeño estará sesgada.

La ciencia abierta promueve acceso, transparencia, colaboración y reutilización. La Recomendación de la UNESCO sobre Ciencia Abierta ofrece un marco internacional que incluye conocimiento científico abierto, infraestructuras, participación social y diálogo con otros sistemas de conocimiento (UNESCO, 2021). Apertura, sin embargo, no equivale a exposición indiscriminada. Historias clínicas, localización de predios, datos de propietarios, genómica

sensible o información comercial pueden requerir anonimización, acceso controlado, acuerdos de uso o publicación de datos sintéticos. Un proyecto responsable hace tan abierto como sea posible y tan restringido como sea necesario.

En ciencias veterinarias, la apertura también debe reconocer relaciones de poder. Los datos de una granja pertenecen a una red de actores: productor, profesionales, laboratorio, proveedor de sensores e institución investigadora. Antes de prometer un repositorio público, se deben definir titularidad, consentimiento, beneficio compartido, restricciones sanitarias y condiciones de reutilización. La transparencia científica no autoriza a perjudicar la bioseguridad ni a revelar la identidad económica de una unidad productiva.

Tabla 15.1.

Niveles de verificación de una investigación computacional veterinaria.

Nivel	Pregunta de verificación	Ejemplo veterinario	Evidencia mínima
Repetibilidad	¿El mismo entorno reproduce el resultado?	El script recalcula sensibilidad y calibración del clasificador radiográfico.	Datos congelados, semilla, versiones y comando de ejecución.
Reproducibilidad	¿Un tercero reconstruye el análisis?	Otro grupo genera las tablas del estudio de dolor posoperatorio.	Protocolo, diccionario, código, dependencias y resultados esperados.
Replicabilidad	¿Nuevos datos sostienen el hallazgo?	Una cohorte independiente confirma la asociación entre señal del sensor y mastitis.	Protocolo independiente, población nueva y análisis predefinido.
Generalización	¿Funciona fuera del contexto original?	Una CNN entrenada en un hospital mantiene desempeño en equipos radiográficos distintos.	Validación externa, análisis por subgrupos, calibración e impacto.

15.2. Organización de proyectos de análisis

Una estructura de carpetas no es decoración informática; expresa la lógica del estudio. El principio central consiste en preservar los datos originales como objetos de solo lectura y derivar de ellos cada producto intermedio mediante código. Las correcciones manuales dentro de una hoja de

cálculo rompen la cadena de procedencia porque el cambio puede no quedar registrado. Si se detecta que “macho” aparece como M, Male y 1, la recodificación debe escribirse en un script y acompañarse de una regla documentada.

La organización útil separa funciones. Una carpeta de datos originales contiene exportaciones sin modificar. Otra conserva datos derivados que pueden regenerarse. El código se divide por etapas numeradas; los productos finales se destinan a resultados; las decisiones viven en un README, un diccionario y un registro de cambios. Los nombres deben describir contenido y versión lógica, no estados emocionales: análisis_final_ahora_si_v7.R es menos informativo que 04_validacion_externa.R.

```
proyecto_mastitis/
├── README.md
├── protocolo/
│   ├── protocolo_registrado.pdf
│   └── plan_analisis.md
├── datos/
│   ├── originales/        # sin edición manual
│   ├── externos/          # validación geográfica o temporal
│   └── derivados/         # regenerables
├── R/
│   ├── 01_importar.R
│   ├── 02_limpiar.R
│   ├── 03_describir.R
│   ├── 04_entrenar.R
│   ├── 05_validar.R
│   └── funciones.R
├── modelos/               # objetos serializados y fichas del modelo
├── resultados/            # tablas y figuras publicables
├── reporte/               # manuscrito y suplemento
├── renv.lock
└── LICENSE
```

Figura 15.1.

Estructura reproducible de un proyecto veterinario en R



La unidad de organización debe reflejar la unidad científica. En estudios longitudinales conviene conservar identificadores separados para animal, episodio, visita, lote y predio. El identificador público no debe contener cédula del propietario, número telefónico, nombre del animal o coordenadas exactas. Una tabla maestra de correspondencias, si es imprescindible, permanece cifrada y fuera del repositorio analítico.

Cada ejecución importante puede producir un manifiesto: fecha, versión de datos, versión del repositorio, sistema operativo, versión de R, paquetes, semilla aleatoria y huellas digitales de archivos. Este registro permite responder por qué dos resultados difieren. También facilita la auditoría cuando un laboratorio actualiza rangos de referencia o un proveedor cambia el firmware de un sensor.

15.3. Uso de R y RStudio en flujos reproducibles

R ofrece un lenguaje estadístico, un ecosistema de paquetes y mecanismos para automatizar análisis. RStudio agrega un entorno integrado, pero la reproducibilidad depende del código y del entorno, no de la interfaz. Una secuencia sólida importa datos, valida tipos y rangos, transforma variables, describe la población, ajusta modelos, evalúa desempeño y genera salidas sin pasos

ocultos. Cada etapa debe poder ejecutarse desde una sesión limpia, siguiendo prácticas computacionales explícitas y proporcionales al proyecto (Sandve et al., 2013; Wilson et al., 2017).

Las fuentes frecuentes de discrepancia son versiones de paquetes, aleatoriedad, configuración regional, zonas horarias y rutas absolutas. Las semillas controlan procesos pseudoaleatorios, aunque no garantizan identidad entre todas las bibliotecas o arquitecturas. Un archivo de bloqueo de dependencias, como renv.lock, registra versiones de paquetes. Las rutas relativas evitan que el proyecto funcione solo en el computador del autor. La información de sesión permite documentar el entorno que produjo el resultado.

```
# Inicio reproducible de un análisis veterinario
set.seed(1501)
options(stringsAsFactors = FALSE)

# Rutas relativas al proyecto
ruta <- file.path("datos", "derivados", "cohorte_mastitis.csv")
datos <- read.csv(ruta, na.strings = c("", "NA"))

# Comprobaciones que detienen el análisis ante errores
stopifnot(!anyDuplicated(datos$id_registro))
stopifnot(all(datos$dias_leche >= 0, na.rm = TRUE))
stopifnot(all(datos$desenlace_mastitis %in% c(0, 1)))

# Al cierre: registrar el entorno
sessionInfo()
```

Un flujo reproducible no debe convertir el código en una caja negra. Las funciones necesitan nombres clínicamente legibles, argumentos explícitos y pruebas para operaciones críticas. Si una función calcula incidencia por vaca-año, una prueba automatizada con un conjunto pequeño debe comprobar numerador, tiempo en riesgo y manejo de animales perdidos. El comentario útil explica por qué se eligió una regla; repetir en palabras lo que hace una línea agrega poco.

R también permite declarar objetivos y dependencias mediante sistemas de canalización. En un proyecto con imágenes, datos clínicos y laboratorio, el reporte final solo debería regenerar los componentes afectados por un cambio. Esta idea reduce tiempo y errores, pero no elimina la necesidad de revisar resultados. La automatización amplifica tanto las buenas decisiones como los defectos sistemáticos.

15.4. Scripts, cuadernos y reportes dinámicos

Un script es apropiado para operaciones estables y reutilizables. Un cuaderno combina narración, código y resultados, por lo que resulta valioso para exploración, docencia y auditoría. Un reporte dinámico genera un documento a partir de datos y código, y reduce la transcripción manual de cifras. La elección no es ideológica: depende de la etapa y del público. El análisis central puede residir en funciones y scripts probados, mientras un documento Quarto o R Markdown los ejecuta y presenta.

Los cuadernos plantean un riesgo conocido: ejecutar celdas fuera de orden y conservar objetos antiguos en memoria. Antes de entregar un reporte, debe renderizarse desde una sesión limpia. Los mensajes, advertencias y errores se gestionan de forma deliberada; ocultarlos sin examinarlos puede encubrir convergencia deficiente, separación completa en regresión logística o niveles nuevos en validación externa.

Los reportes dinámicos son especialmente útiles para estudios repetitivos: vigilancia mensual de leptospirosis, control de calidad de laboratorio o actualización de indicadores de bienestar. No deben confundirse con análisis confirmatorios flexibles. Si cada actualización modifica desenlaces, umbrales o población, el producto ya no es la misma investigación. Las decisiones preespecificadas deben estar separadas de los análisis exploratorios.

Tabla 15.2.

Funciones y controles de productos computacionales.

Objeto	Uso recomendado	Riesgo principal	Control
Script .R	Limpieza, funciones, modelos y validación.	Dependencias implícitas entre archivos.	Orden explícito, pruebas y canalización.
Cuaderno	Exploración comentada y enseñanza.	Estado oculto y ejecución desordenada.	Renderizar desde sesión limpia.
Reporte dinámico	Manuscritos, suplementos y vigilancia periódica.	Resultados desactualizados si no se recompila.	Compilación automatizada y registro de fecha.

Objeto	Uso recomendado	Riesgo principal	Control
Aplicación interactiva	Exploración clínica o productiva.	Confundir interfaz atractiva con evidencia.	Mostrar versión, población, límites e incertidumbre.

15.5. Control de versiones y documentación

El control de versiones registra cambios en archivos de texto y permite vincular una cifra publicada con el código que la produjo. Git no sustituye una copia de seguridad ni un repositorio institucional; resuelve otra pregunta: qué cambió, quién lo cambió y por qué (Bryan, 2018). Los commits deben representar unidades comprensibles, como "corrige definición temporal de mastitis clínica", y no acumulaciones de cientos de archivos bajo el mensaje "actualización".

Las ramas son útiles para probar una imputación alternativa o una nueva figura sin alterar la versión estable. Las etiquetas pueden marcar el análisis enviado a revisión y el aceptado. Los datos grandes, las credenciales, los identificadores personales y los archivos temporales no deben incorporarse al historial. El archivo .gitignore reduce inclusiones accidentales, pero no repara una filtración ya registrada.

La documentación mínima comprende README, licencia, diccionario de datos, instrucciones de ejecución, procedencia de datos, restricciones de uso, definición de desenlaces, estructura del repositorio y contacto responsable. Un modelo que se pretende reutilizar requiere además una ficha o model card: propósito, población objetivo, entradas, salidas, versión, desempeño global y por subgrupos, limitaciones, umbral, condiciones de no uso, fecha de evaluación y responsable de mantenimiento.

La documentación también debe explicar decisiones negativas: variables excluidas por medición inconsistente, centros omitidos, etiquetas inciertas, cambios respecto del protocolo y análisis que no convergieron. Estos elementos permiten una revisión científica más honesta que la simple presentación del modelo ganador.

15.6. Datos y código FAIR

FAIR significa localizable, accesible, interoperable y reutilizable. Los principios se formularon para mejorar la administración de objetos digitales y su procesamiento por humanos y máquinas (Wilkinson et al., 2016). FAIR no es sinónimo de abierto: un conjunto clínico puede ser localizable mediante metadatos y accesible bajo un procedimiento autorizado. Para el software, FAIR4RS adapta el marco a objetos que tienen versiones, dependencias, capacidad ejecutable y condiciones de licencia (Barker et al., 2022).

Tabla 15.3.

Aplicación práctica de FAIR a datos, software y modelos veterinarios.

Principio	Datos veterinarios	Código/modelo	Pregunta de auditoría
Localizable	Identificador persistente, título, palabras clave, especies y metadatos.	Repositorio indexado, versión etiquetada y cita de software.	¿Puede encontrarse la versión exacta?
Accesible	Protocolo de acceso abierto o controlado, con autenticación si procede.	Licencia y procedimiento de descarga o ejecución.	¿Se entiende cómo obtenerlo y bajo qué condiciones?
Interoperable	Vocabularios, unidades, fechas y formatos documentados.	Entradas y salidas estándar; dependencias declaradas.	¿Puede integrarse sin adivinar significados?
Reutilizable	Procedencia, calidad, consentimiento, licencia y diccionario.	Documentación, pruebas, ejemplos, versión y límites.	¿Puede evaluarse si sirve para otro contexto?

La interoperabilidad veterinaria enfrenta vocabularios heterogéneos. Una misma condición puede registrarse como “otitis”, “otitis externa”, código local o texto libre. Las unidades de laboratorio cambian, las razas se agregan de manera distinta y la especie puede estar implícita. Antes de entrenar un modelo, estas diferencias deben resolverse mediante diccionarios, catálogos de códigos y transformaciones versionadas. La interoperabilidad semántica es tan importante como el formato CSV o JSON.

Para depositar un producto conviene asignar identificadores persistentes y citar versiones, no solo la página principal del repositorio. La licencia debe diferenciar datos, código y documentación. Si

los datos no pueden compartirse, todavía es posible publicar esquema, diccionario, código, datos simulados, pruebas y un procedimiento de solicitud. La imposibilidad ética de abrir historias clínicas no justifica ocultar el método.

15.7. Comunicación de métodos y resultados algorítmicos

Un lector debe poder reconstruir quiénes integraron la población, cuándo y dónde se recogieron los datos, cuál fue la unidad de análisis, cómo se definieron predictores y desenlace, qué particiones se realizaron y qué decisiones fueron anteriores a observar el resultado. Decir “se utilizó random forest” es insuficiente. Se necesitan versión de software, variables candidatas, preprocesamiento, manejo de ausencias, estrategia de validación, búsqueda de hiperparámetros, métrica de selección y especificación final.

Los métodos deben identificar la posible dependencia entre observaciones. En ganadería, varias mediciones de la misma vaca, animales del mismo lote o granjas de una región violan la independencia si se tratan como filas aisladas. En animales de compañía, múltiples consultas de un paciente generan el mismo problema. La partición debe ocurrir al nivel que impida contaminación: animal, propietario, granja, hospital o periodo, según la pregunta.

Los resultados necesitan más que exactitud. Para clasificación se informan matriz de confusión, sensibilidad, especificidad, valores predictivos, discriminación, calibración e intervalos de confianza. Para regresión se presentan error absoluto, error cuadrático, distribución de residuales y relevancia clínica del error. Cuando existe un umbral, debe explicarse quién lo eligió, qué costo asigna a falsos negativos y falsos positivos y si fue fijado antes de la prueba.

La incertidumbre incluye variación muestral, error de etiquetado, cambio de dominio y decisiones analíticas. Los intervalos de confianza obtenidos con remuestreo deben respetar agrupamientos. También se informa desempeño por sexo, edad, raza, especie, centro, equipo y otros grupos pertinentes, evitando inferencias fuertes cuando los tamaños son pequeños. La equidad no consiste en exigir métricas idénticas sin contexto, sino en investigar diferencias, daño potencial y mecanismos de medición.

15.8. Tablas, figuras y visualización científica

Una tabla conserva valores exactos; una figura hace visible un patrón. La elección debe responder a una pregunta. La Tabla 1 de un artículo predictivo suele describir la población de desarrollo y validación por separado, con número de animales, eventos, ausencias y distribución de variables. No debe utilizar valores p para "demostrar" que dos particiones aleatorias son iguales; las diferencias descriptivas y su significado son más útiles.

Las curvas ROC muestran discriminación a través de umbrales, pero pueden ocultar rendimiento deficiente en eventos raros. Las curvas precisión-recall resultan informativas cuando la clase positiva es infrecuente. Una gráfica de calibración compara probabilidades estimadas y observadas; debe acompañarse de intercepto, pendiente y, cuando proceda, puntuación de Brier. Para impacto potencial, las curvas de decisión relacionan beneficio neto y umbrales, aunque requieren explicar supuestos clínicos.

Las visualizaciones deben mostrar denominadores, incertidumbre y escala. En series de sensores, agregar por día puede ocultar episodios breves de dolor o estrés. En mapas epidemiológicos, la intensidad aparente depende de la población animal en riesgo y de la resolución espacial. En imágenes explicables, un mapa de activación no prueba causalidad ni razonamiento clínico: indica regiones influyentes bajo un método específico y necesita validación por especialistas.

- Evitar gráficos tridimensionales, ejes truncados sin justificación, arcoíris y áreas proporcionales engañosas.
- Incluir unidades, periodo, población, número de observaciones y definición del evento en títulos o notas.
- Diferenciar datos observados, predicciones e intervalos; no unir con líneas categorías sin continuidad.
- Usar paletas legibles para personas con alteraciones de visión del color y verificar contraste en escala de grises.
- No repetir exactamente la misma información en texto, tabla y figura; el texto interpreta, no recita.

15.9. Redacción de artículos con modelos predictivos

El título y el resumen deben identificar si el estudio desarrolló, actualizó, validó o evaluó impacto. “Aplicación de IA al diagnóstico de enfermedad” no aclara la tarea ni el diseño. Un título informativo podría ser: “Desarrollo y validación temporal de un modelo para predecir mastitis clínica a siete días en vacas Holstein mediante registros de ordeño”. La palabra diagnóstico tampoco debe emplearse si el desenlace es un marcador sustituto.

La introducción parte del problema veterinario y de la decisión que el modelo pretende apoyar. Debe explicar qué ocurre sin el modelo, qué usuarios lo emplearían y por qué el horizonte temporal es relevante. El vacío científico no es “no se ha usado XGBoost”, sino la ausencia de una predicción válida, calibrada y útil en una población concreta.

Los métodos describen fuente de datos, elegibilidad, desenlace, predictores, tamaño muestral, ausencias, preprocesamiento, modelado, validación y análisis. El algoritmo final debe presentarse de forma que pueda aplicarse: ecuación y coeficientes para regresión, o código/objeto versionado para métodos complejos. Un artículo que muestra AUC pero no permite obtener una predicción individual deja incompleta su contribución.

En resultados se presenta flujo de inclusión, características de muestras, número de eventos, datos faltantes, desempeño e incertidumbre. La discusión contrasta hallazgos con evidencia previa, examina errores y explica aplicabilidad. La superioridad estadística sobre un comparador no implica beneficio clínico. Las limitaciones deben abordar sesgo de selección, calidad del referente, fuga, optimismo, cambio temporal, representatividad y costo de error.

15.10. Guías para reportar estudios de inteligencia artificial

Las guías de reporte son listas de información necesaria; no certifican calidad por sí mismas. TRIPOD+AI actualiza el reporte de modelos de predicción que usan regresión o aprendizaje automático y enfatiza transparencia, desempeño, equidad y ciencia abierta (Collins et al., 2024). PROBAST+AI, publicado después, evalúa calidad del desarrollo, riesgo de sesgo en la evaluación y aplicabilidad en cuatro dominios: participantes y fuentes de datos, predictores, desenlace y análisis (Moons et al., 2025).

Para exactitud diagnóstica de sistemas de IA, STARD-AI amplía los estándares de reporte e incorpora elementos específicos del índice algorítmico, datos y evaluación (Sounderajah et al., 2025). DECIDE-AI se dirige a evaluaciones clínicas tempranas en vivo de sistemas de apoyo a decisiones y presta atención al flujo de trabajo, factores humanos, seguridad y modificaciones del sistema (Vasey et al., 2022). CONSORT-AI y SPIRIT-AI abordan ensayos clínicos y sus protocolos (Liu et al., 2020; Rivera et al., 2020); CLAIM orienta estudios de IA en imagen médica (Mongan et al., 2020). Aunque muchas se originaron en medicina humana, sus principios son útiles en veterinaria si se explicitan especie, raza, unidad productiva, profesional usuario y condiciones de atención.

Tabla 15.4.

Selección orientativa de guías de reporte para investigaciones veterinarias con IA.

Tipo de estudio	**Guía principal**	**Aporte**	**Adaptación veterinaria imprescindible**
Desarrollo/validación de predicción	TRIPOD+AI	Reporte completo de datos, modelo, evaluación y acceso.	Especie, raza, unidad agrupada, horizonte y decisión clínica/productiva.
Evaluación de calidad o sesgo	PROBAST+AI	Calidad del desarrollo, sesgo de desempeño y aplicabilidad.	Representatividad de centros, referente veterinario y dependencia entre animales.
Exactitud diagnóstica con IA	STARD-AI	Transparencia del índice, referente y flujo diagnóstico.	Espectro de enfermedad, lectores veterinarios, equipos y especies.
Evaluación temprana en vivo	DECIDE-AI	Interacción humano-sistema, seguridad y flujo de trabajo.	Rol del veterinario, productor, tutor y contexto de uso.
Ensayo clínico con intervención IA	SPIRIT-AI / CONSORT-AI	Protocolo y reporte de ensayos que incorporan IA.	Bienestar, consentimiento del responsable y unidad experimental.
IA en imágenes	CLAIM	Adquisición, partición, referencia y evaluación de imágenes.	Modalidad, equipo, proyección, especie, anotadores y nivel de partición.

La guía se elige durante la planificación, no la noche previa al envío. Puede revelar que falta validación temporal, descripción de exclusiones o acceso al modelo. Cumplir una lista tampoco

corrige un diseño débil. Un estudio puede reportar con gran transparencia que usó una muestra pequeña y un referente imperfecto; será claro, pero seguirá siendo metodológicamente limitado.

15.11. Revisión crítica por pares

La revisión de un manuscrito algorítmico requiere competencia veterinaria, estadística y computacional. La primera pregunta no es qué algoritmo se utilizó, sino si la pregunta, el diseño y los datos permiten responder el problema. Después se examina si el desempeño fue estimado con separación real entre entrenamiento y evaluación. La comparación justa exige igual información, particiones y presupuesto de ajuste para todos los modelos, incluido un referente sencillo.

El revisor debe rastrear el denominador. ¿Cuántos animales, eventos, imágenes, predios y centros existen? Diez mil radiografías de doscientos perros no equivalen a diez mil pacientes independientes. También debe identificar exclusiones posteriores, datos faltantes, selección del "mejor" resultado entre múltiples intentos y uso de la prueba para ajustar el modelo. La ausencia de calibración es una señal grave si se pretende comunicar riesgo individual.

Reproducir no significa aceptar. El código puede ejecutar una evaluación sesgada. PROBAST+AI ayuda a separar calidad del desarrollo y riesgo de sesgo del desempeño (Moons et al., 2025). La revisión debe pedir archivos o explicaciones cuando sean necesarios, respetando confidencialidad. Si los datos no pueden compartirse, el autor puede facilitar datos simulados, funciones, especificaciones, listas de paquetes y resultados intermedios verificables.

- ¿La población representa el uso previsto y se reportan exclusiones?
- ¿Predictores y desenlace estaban disponibles en el momento real de decisión?
- ¿La referencia fue independiente, cegada y clínicamente defendible?
- ¿La partición evitó dependencia por animal, granja, centro, equipo y tiempo?
- ¿Todo preprocesamiento se estimó solo con entrenamiento?
- ¿Se comparó con práctica habitual y con modelos simples?
- ¿Se informan calibración, incertidumbre, errores y subgrupos?
- ¿Las conclusiones corresponden al diseño y no prometen impacto no evaluado?
- ¿Código, modelo, datos o restricciones de acceso están documentados?

- ¿Conflictos, financiamiento y participación del proveedor son transparentes?

15.12. Transferencia del modelo a la práctica veterinaria

La transferencia comienza al definir la decisión. Un modelo de riesgo anestésico puede mostrarse antes de la inducción, pero su valor depende de una acción: modificar monitorización, solicitar pruebas, posponer el procedimiento o referir al paciente. Si el equipo no acuerda qué hacer ante cada nivel de riesgo, el sistema añade información sin cambiar la atención. Esta distancia entre rendimiento técnico e impacto clínico es uno de los problemas centrales de la implementación de IA en salud (Kelly et al., 2019).

La ruta de implementación comprende validación local silenciosa, prueba de usabilidad, evaluación prospectiva, estudio de impacto y vigilancia. En modo silencioso, el algoritmo corre sin influir en decisiones; se mide calibración y disponibilidad de variables. Después se estudia cómo el usuario interpreta la salida, cuánto tarda y qué errores introduce. Solo entonces es razonable evaluar resultados clínicos, bienestar, eficiencia, costos y posibles daños.

El despliegue debe especificar responsables de actualización, umbrales de alarma, tiempo de respuesta, registro de recomendaciones, posibilidad de anulación y plan de contingencia. El profesional conserva autoridad, pero “humano en el circuito” no garantiza seguridad si confía de manera automática o ignora sistemáticamente las alertas. Deben medirse adherencia, fatiga de alarmas, desacuerdos y consecuencias.

El cambio de distribución es inevitable. Se modifican protocolos, equipos, prevalencia, razas atendidas, personal y prácticas de registro. La vigilancia compara entradas, prevalencia, discriminación, calibración y resultados por grupos. Una caída puede requerir recalibración, reentrenamiento o retiro. Actualizar sin control transforma un modelo validado en otro distinto; cada versión necesita documentación y evaluación proporcional al riesgo, de acuerdo con una perspectiva de ciclo de vida (Food and Drug Administration, 2021).

15.13. Costos, infraestructura y competencias profesionales

El costo total no se limita al entrenamiento. Incluye captura y etiquetado, integración de sistemas, almacenamiento, cómputo, seguridad, validación, soporte, capacitación, mantenimiento, auditoría y reemplazo. En muchos proyectos veterinarios, la limpieza y armonización consumen más recursos que el algoritmo. Una solución local con variables rutinarias puede ser preferible a una arquitectura compleja que dependa de conectividad irregular o sensores difíciles de calibrar.

Tabla 15.5.

Componentes del costo total de un sistema veterinario de IA.

Componente	Costo visible	Costo frecuentemente omitido	Pregunta de decisión
Datos	Sensores, imágenes o pruebas.	Etiquetado experto, gobernanza y corrección.	¿El dato mejora la decisión lo suficiente?
Cómputo	Servidor, nube o estación de trabajo.	Transferencia, energía, latencia y respaldo.	¿Puede ejecutarse en el contexto objetivo?
Integración	Interfaz y conexión al sistema.	Cambios de flujo, soporte y caída del servicio.	¿Quién actúa cuando falla?
Validación	Estudio local y análisis.	Tiempo profesional y seguimiento de daños.	¿Qué evidencia habilita el uso?
Mantenimiento	Actualizaciones y licencias.	Deriva, recalibración y retiro de versiones.	¿Quién responde durante el ciclo de vida?

Las competencias necesarias forman un equipo, no una profesión imposible. El veterinario aporta fisiopatología, flujo clínico y juicio sobre daño; el bioestadístico, diseño e inferencia; el científico de datos, ingeniería y evaluación; el especialista en sistemas, integración y seguridad; y los usuarios finales, condiciones reales. En producción animal deben participar productores y técnicos. La gobernanza requiere además ética, derecho, bienestar animal y administración; los principios de autonomía, seguridad, transparencia, responsabilidad e inclusión propuestos para IA en salud

ofrecen un referente adaptable, no una sustitución del análisis veterinario (World Health Organization, 2021).

Para instituciones con recursos limitados, la estrategia no debería empezar comprando infraestructura. Conviene priorizar problemas de alta carga, datos ya disponibles, modelos simples, estándares abiertos y alianzas regionales. El cálculo de retorno debe incluir desenlaces animales y sociales, no solo ahorro. Un sistema barato que incrementa tratamientos innecesarios desplaza costos y puede deteriorar bienestar o resistencia antimicrobiana.

15.14. Modelos multimodales y medicina veterinaria de precisión

Un modelo multimodal integra fuentes diferentes: texto clínico, imagen, laboratorio, señales fisiológicas, genómica, ambiente y comportamiento. La promesa es representar mejor el paciente o el sistema productivo. En un perro con sospecha de cardiopatía, la combinación de historia, examen, biomarcadores, electrocardiograma y ecocardiografía podría apoyar estratificación. En bovinos, producción, rumia, actividad, temperatura y clima podrían anticipar trastornos metabólicos.

Integrar no significa acumular. Las modalidades se miden en tiempos y escalas distintos, pueden faltar de forma informativa y tener calidad desigual. Un modelo puede aprender que disponer de una tomografía identifica a pacientes de mayor gravedad, no una característica biológica de la imagen. La evaluación debe comparar el sistema multimodal con cada modalidad y con el estándar clínico, cuantificar ganancia incremental y estudiar si esa ganancia justifica costo y demora.

La medicina veterinaria de precisión adapta prevención, diagnóstico o tratamiento a características del animal y su contexto. Su unidad puede ser un paciente, un rebaño o una población. La precisión no debe degenerar en determinismo: una probabilidad individual conserva incertidumbre. Además, la diversidad de especies y razas fragmenta datos; los modelos fundacionales o de transferencia pueden ayudar, pero también transportar sesgos entre dominios biológicamente distintos.

Un diseño prospectivo debe definir qué modalidad estará disponible en el momento de uso, cómo se sincronizan mediciones y qué ocurre cuando falta una fuente. Es recomendable evaluar rutas de

degradación segura: si el sensor falla, el sistema debe abstenerse o operar con un modelo reducido validado, no imputar silenciosamente una señal crítica.

15.15. Gemelos digitales, robótica y sistemas autónomos

Un gemelo digital es una representación computacional actualizable de un objeto o proceso físico. En veterinaria podría representar una vaca, un lote, una unidad acuícola o un flujo hospitalario. Recibe datos y simula estados o respuestas: balance energético, riesgo térmico, crecimiento, ocupación de camas o efecto de intervenciones. No es un tablero con indicadores ni una réplica gráfica; requiere un vínculo dinámico, un modelo explícito y validación de predicciones.

La combinación de mecanismos biológicos y aprendizaje automático es atractiva. Un modelo fisiológico puede imponer coherencia, mientras el componente aprendido corrige patrones residuales. Sin embargo, un gemelo detallado exige mediciones frecuentes y parámetros identificables. Si múltiples combinaciones producen la misma salida, la personalización puede ser aparente. La validación debe incluir escenarios no usados en calibración y análisis de sensibilidad.

La robótica ya participa en ordeño, alimentación, limpieza, monitoreo y cirugía asistida. Los sistemas autónomos perciben, deciden y actúan, por lo que un error algorítmico puede traducirse en daño físico inmediato. La seguridad exige límites mecánicos, detección de presencia animal y humana, parada de emergencia, registros de eventos, mantenimiento preventivo y supervisión. La evaluación de bienestar debe observar si el animal puede elegir, adaptarse y evitar interacciones aversivas.

La autonomía es gradual. Un sistema puede recomendar, ejecutar con confirmación o actuar sin intervención. Cuanto mayor sea el potencial de daño, mayor debe ser la evidencia, redundancia y capacidad de recuperación. En cirugía, administración de fármacos o decisiones de sacrificio, la delegación plena plantea umbrales éticos particularmente altos. El futuro técnicamente posible no define por sí mismo el futuro aceptable.

15.16. Agenda de investigación para América Latina

América Latina combina biodiversidad, producción pecuaria estratégica, enfermedades endémicas, grandes centros urbanos y zonas rurales con conectividad desigual. Importar modelos entrenados en otros continentes puede fallar por diferencias genéticas, epidemiológicas, climáticas, productivas y de registro. La región necesita datos propios y cooperación que evite concentrar beneficios en pocas instituciones.

1. Construir cohortes multicéntricas de calidad para enfermedades prioritarias de animales de compañía y producción, con diccionarios comunes.
2. Desarrollar repositorios regionales con acceso gobernado, identificadores persistentes y reconocimiento de quienes generan datos.
3. Validar externamente modelos por país, altitud, clima, especie, raza, sistema productivo y nivel de atención.
4. Crear bancos de imágenes veterinarias anotadas mediante protocolos de consenso y medir desacuerdo entre especialistas.
5. Integrar datos One Health para zoonosis, resistencia antimicrobiana, movilidad, ambiente y fauna silvestre.
6. Diseñar soluciones frugales que funcionen sin conexión permanente y en equipos de bajo costo.
7. Evaluar impacto sobre bienestar, uso de antimicrobianos, productividad, equidad y carga laboral, no solo AUC.
8. Investigar sesgos lingüísticos y de representación en modelos generativos usados por profesionales hispanohablantes y lusófonos.
9. Fortalecer formación en bioestadística, programación, gobernanza, ética, ciberseguridad y comunicación.
10. Establecer redes regulatorias y de evaluación de tecnologías veterinarias adaptadas al riesgo y al ciclo de vida.

Una agenda regional debe proteger la soberanía de datos sin aislar la ciencia. Los acuerdos pueden permitir análisis federado o metaanálisis con datos armonizados, manteniendo información sensible en cada institución. También deben evitar extractivismo: recolectar datos locales para

desarrollar productos inaccesibles a las comunidades que los generaron. La participación de universidades públicas, servicios veterinarios, colegios profesionales, productores y comunidades es una condición de legitimidad.

15.17. Formación del investigador veterinario ante la IA

El investigador veterinario no necesita programar cada arquitectura desde cero, pero sí formular preguntas, reconocer diseños inválidos, leer código elemental, interpretar métricas y detectar afirmaciones excesivas. La alfabetización en IA empieza por método científico y bioestadística. Sin comprensión de muestreo, confusión, dependencia, incertidumbre y causalidad, una interfaz accesible puede acelerar errores.

Tabla 15.6.

Competencias para la formación del investigador veterinario ante la IA.

Dominio	Competencia observable	Producto de aprendizaje
Problema y diseño	Convierte una necesidad clínica en pregunta, población, desenlace y decisión.	Protocolo con unidad experimental y plan de validación.
Datos	Audita procedencia, calidad, ausencias, etiquetas y sesgos.	Diccionario y reporte de calidad.
Análisis	Selecciona modelos simples o complejos y evita fuga.	Pipeline reproducible con comparador base.
Evaluación	Interpreta discriminación, calibración, utilidad, incertidumbre y subgrupos.	Informe crítico con límites clínicos.
Reproducibilidad	Versiona código, dependencias y productos.	Repositorio ejecutable y documentado.
Comunicación	Reporta con guía adecuada y visualiza sin distorsión.	Manuscrito, suplemento y ficha de modelo.
Ética e implementación	Evalúa daño, privacidad, equidad, costo y vigilancia.	Plan de despliegue y retiro.

La enseñanza debería usar proyectos con datos imperfectos, porque la práctica rara vez ofrece conjuntos limpios. Los estudiantes deben experimentar una partición incorrecta y observar el optimismo, comparar un modelo complejo con regresión logística, recalibrar una predicción y

redactar una limitación honesta. La evaluación puede valorar el razonamiento y la trazabilidad más que la cifra máxima.

Los modelos generativos pueden apoyar explicación de código, borradores y depuración, pero la persona investigadora verifica cada afirmación, referencia y resultado. Nunca se introducen datos clínicos identificables en servicios no autorizados. La formación debe incluir registro del uso de estas herramientas y responsabilidad autoral: una salida plausible no constituye evidencia.

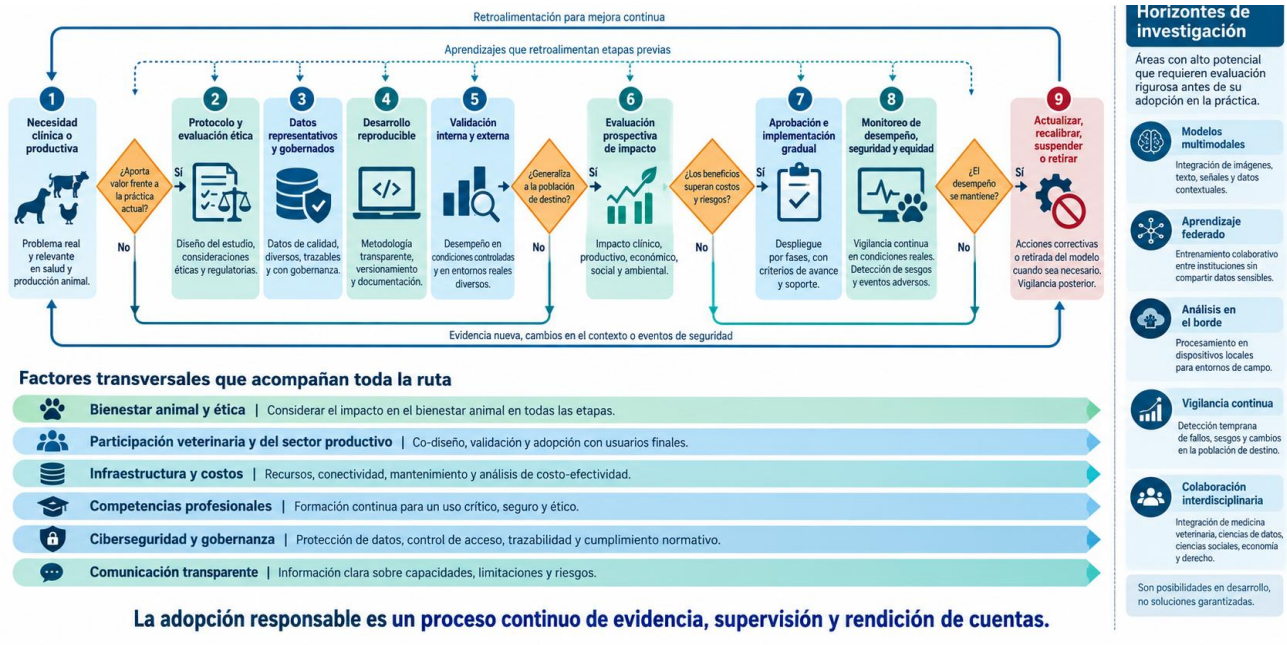
15.18. Proyecto integrador: desarrollo de una investigación veterinaria asistida por IA

Caso didáctico simulado: predicción temprana de mastitis clínica bovina

Este proyecto simulado integra diseño, datos, modelado, validación, reporte e implementación. Su finalidad es pedagógica; ninguna cifra debe interpretarse como evidencia clínica real. La situación propone una red de 18 fincas lecheras de dos provincias que desea identificar vacas con alto riesgo de mastitis clínica durante los siguientes siete días para intensificar observación y realizar evaluación veterinaria dirigida, no para administrar antibióticos automáticamente.

Figura 15.3.

Ruta del proyecto integrador sobre mastitis bovina



15.18.1. Problema, pregunta e hipótesis

Problema: la detección se produce después de signos visibles, mientras registros de ordeño y actividad contienen cambios tempranos. Pregunta: ¿un modelo que integra conductividad, producción, rumia, actividad, días en leche, paridad y antecedentes predice mastitis clínica a siete días con calibración suficiente en fincas no usadas para entrenar? Hipótesis: el modelo multimodal supera a una regla basada solo en conductividad y conserva calibración aceptable en validación geográfica.

Población objetivo: vacas lecheras lactantes bajo sistemas con sensores compatibles. Unidad de predicción: vaca-día. Unidad de partición externa: finca. Desenlace: primer episodio clínico confirmado durante los siete días posteriores, según protocolo veterinario predefinido. El episodio y el periodo de exclusión posterior evitan contar días correlacionados como eventos nuevos.

15.18.2. Diseño y gobernanza

Se plantea una cohorte retrospectiva para desarrollo y validación interna con datos de 14 fincas, seguida de validación externa en cuatro fincas reservadas. Una fase prospectiva silenciosa evaluará disponibilidad de variables, calibración y flujo antes de cualquier intervención. El protocolo se registra antes de acceder al conjunto externo. Los acuerdos establecen titularidad, finalidad, acceso, conservación, publicación y devolución de resultados a productores.

Los identificadores de finca y animal se seudonimizan. Coordenadas exactas, nombres y datos comerciales no ingresan al repositorio analítico. El comité correspondiente evalúa bienestar, privacidad y uso secundario. Un comité con veterinarios, productores, estadístico y responsable de datos revisa definiciones y cambios.

15.18.3. Variables y control de calidad

Tabla 15.7.

Diccionario resumido del caso didáctico simulado.

Elemento	Definición operativa	Control

Desenlace	Mastitis clínica confirmada dentro de 1–7 días.	Auditoría ciega de muestra y concordancia.
Producción	Litros por ordeño y cambio respecto de mediana individual.	Rango, unidad, frecuencia y fallos del medidor.
Conductividad	Valor por cuarto y desviación respecto de línea basal.	Calibración del equipo y marca.
Rumia/actividad	Minutos diarios y cambios de 3 y 7 días.	Días sin transmisión y calidad del collar.
Contexto	Días en leche, paridad, estación, finca.	Categorías, fechas y coherencia longitudinal.
Antecedentes	Episodios previos y recuento celular disponible.	Ventana temporal; prevenir datos futuros.

La auditoría documenta duplicados, interrupciones, unidades, desfases horarios y ausencias. Una variable medida después del inicio clínico queda excluida. La imputación se estima dentro de cada pliegue de entrenamiento. Se agregan indicadores de ausencia cuando el patrón puede tener significado operativo, pero se evalúa si el modelo aprende fallos del sensor asociados a una finca.

15.18.4. División, modelos y validación

Las cuatro fincas externas permanecen selladas hasta fijar el pipeline. En las 14 restantes se usa validación cruzada agrupada por finca. Se comparan: regla de conductividad, regresión logística penalizada y gradient boosting. El criterio principal no es el AUC aislado; se evalúan AUC ROC, área precisión-recall, sensibilidad a umbral operativo, valor predictivo positivo, Brier, intercepto y pendiente de calibración. Los intervalos se obtienen con remuestreo por finca.

```
# Esquema conceptual; requiere adaptar nombres y paquetes al conjunto real
set.seed(15018)

# 1. Reservar fincas externas ANTES del ajuste
externas <- c("F15", "F16", "F17", "F18")
entrenamiento <- subset(datos, !(id_finca %in% externas))
prueba_externa <- subset(datos, id_finca %in% externas)

# 2. Crear pliegues agrupados por finca, no por fila
# folds <- group_vfold_cv(entrenamiento, group = id_finca, v = 7)

# 3. Dentro de cada pliegue: imputar, transformar, ajustar y predecir
# Nunca estimar medianas, escalas o selección de variables con prueba externa.

# 4. Congelar receta, hiperparámetros y umbral antes de abrir prueba externa.
```

El umbral se fija para alcanzar una sensibilidad preespecificada en validación interna, sujeta a un número máximo de alertas por cada 100 vacas-día. Esta restricción convierte la métrica en una decisión de capacidad. En la prueba externa se informa el rendimiento sin retocar el umbral. Si la calibración falla pero la discriminación se mantiene, una recalibración se considera un modelo actualizado y se evalúa por separado.

15.18.5. Resultados simulados e interpretación

Supóngase que el conjunto externo contiene 22 400 vaca-días y 112 eventos. El modelo de boosting obtiene AUC 0,82 (IC 95 % simulado: 0,78–0,86), área precisión-recall 0,18 y sensibilidad 0,76 con valor predictivo positivo 0,09. La pendiente de calibración es 0,74 y el intercepto 0,31, compatible con predicciones demasiado extremas y riesgo promedio subestimado. La regresión penalizada alcanza AUC 0,80, área precisión-recall 0,16 y mejor calibración.

La interpretación no es “82 % de exactitud”. Con prevalencia baja, solo alrededor de nueve de cada cien alertas corresponderían a un evento en la ventana; la mayoría serían falsos positivos. Si la acción consiste en observación breve, el costo puede ser aceptable. Si la alerta condujera a tratamiento, el sistema sería clínicamente inadecuado. La pequeña ganancia de discriminación del boosting frente a regresión debe compararse con calibración, explicabilidad, mantenimiento y costo.

El análisis por finca muestra menor sensibilidad donde faltan datos de rumia. Esto sugiere dependencia de infraestructura, no necesariamente diferencia biológica. Se informa sin ajustar retrospectivamente. La fase prospectiva silenciosa deberá medir pérdida de señal, retraso, carga de alertas y concordancia con evaluación veterinaria.

15.18.6. Reproducibilidad y productos

- Protocolo y plan de análisis fechados; desviaciones registradas.
- Diccionario completo, reglas de limpieza y diagrama de flujo de inclusión.
- Repositorio versionado con scripts, pruebas, renv.lock y licencia.
- Datos simulados con el mismo esquema para ejecutar el pipeline.
- Acceso controlado a datos reales mediante comité y acuerdo.

- Modelo congelado, ficha de modelo y huella de versión.
- Reporte dinámico con tablas, figuras, métricas e intervalos.
- Checklist TRIPOD+AI y evaluación crítica con PROBAST+AI.
- Plan prospectivo, indicadores de deriva, criterios de pausa y retiro.

15.18.7. Plan de comunicación e implementación

El manuscrito se dirige a audiencia científica; un informe ejecutivo explica a productores cuántas alertas esperar y qué acción se propone; una ficha técnica informa a veterinarios entradas, límites y condiciones de abstención. Ningún material presenta el sistema como diagnóstico autónomo. La interfaz muestra probabilidad, ventana, datos ausentes, versión y acción recomendada.

La evaluación de impacto compara atención habitual con flujo asistido, idealmente mediante diseño por conglomerados o escalonado si es ética y logísticamente apropiado. Desenlaces: tiempo hasta evaluación, mastitis grave, uso de antimicrobianos, descarte de leche, bienestar, carga laboral, costos y eventos adversos. El modelo solo avanzaría si mejora resultados relevantes sin generar daño o inequidad entre fincas.

15.18.8. Preguntas para discusión del proyecto

11. ¿Qué fuga aparecería si se usan mediciones tomadas el mismo día del diagnóstico?
12. ¿Por qué la finca, y no la vaca-día, gobierna la partición externa?
13. ¿Qué cambia si el objetivo es reducir antibióticos en vez de anticipar evaluación?
14. ¿Cuándo una recalibración constituye actualización que debe revalidarse?
15. ¿Qué desenlace de bienestar podría contradecir una mejora productiva?
16. ¿Qué información debería recibir una finca sin conectividad continua?

15.19. Síntesis y reflexión final

La investigación asistida por IA alcanza valor científico cuando conserva una cadena explícita desde la pregunta hasta la decisión. La reproducibilidad hace visible esa cadena; la comunicación permite examinarla; la validación y el seguimiento determinan si resiste otros contextos. Abrir código o mostrar una métrica no basta. Se requieren definiciones clínicas, unidades correctas, datos

con procedencia, particiones sin fuga, comparadores razonables, calibración, incertidumbre y límites de uso.

Las tecnologías futuras ampliarán la capacidad de integrar señales, simular escenarios y automatizar acciones. También aumentarán el número de puntos donde un error puede propagarse. En medicina veterinaria, la diversidad de especies, contextos productivos y recursos obliga a desconfiar de soluciones universales. La precisión tendrá sentido si mejora decisiones concretas sin desplazar daño hacia animales, responsables, trabajadores o comunidades.

El criterio final no es cuán sofisticado parece un sistema, sino si produce conocimiento verificable y una práctica más segura, justa y biológicamente fundada. Un modelo simple, bien diseñado, reproducible y útil supera a una arquitectura opaca que nadie puede auditar o mantener. La formación del investigador veterinario debe preservar esa jerarquía: primero la pregunta y el diseño; después los datos y el análisis; finalmente la tecnología.

Preguntas de autoevaluación

1. Diferencie repetibilidad, reproducibilidad, replicabilidad y generalización mediante un ejemplo veterinario.
2. Explique por qué un conjunto de datos FAIR no necesariamente es abierto.
3. Proponga una estructura de proyecto para un estudio de supervivencia en oncología canina.
4. ¿Qué información del entorno de R debe conservarse para reconstruir un análisis?
5. Compare ventajas y riesgos de scripts, cuadernos y reportes dinámicos.
6. ¿Qué debe contener una ficha de modelo veterinario?
7. Justifique por qué la partición por imágenes puede producir fuga en radiología veterinaria.
8. Seleccione una guía de reporte para un estudio de exactitud de una CNN y explique su elección.
9. ¿Por qué un AUC alto no garantiza utilidad clínica?
10. Diseñe una secuencia de validación local, evaluación prospectiva y vigilancia.
11. Enumere cinco costos ocultos de implementar un sistema de IA en una clínica o finca.
12. Analice dos riesgos científicos de los modelos multimodales.
13. ¿Qué diferencia un gemelo digital de un tablero de monitoreo?

14. Proponga tres prioridades de investigación en IA veterinaria para Ecuador o América Latina.
15. Explique cómo documentaría el uso de un modelo generativo durante la redacción de un artículo.

Actividad práctica

Auditoría reproducible de un estudio veterinario. Seleccione un artículo reciente que desarrolle o valide un modelo predictivo veterinario. Trabaje con la guía pertinente y elabore un informe de 1 500 a 2 000 palabras con los siguientes productos:

- Identificación de la pregunta, población, unidad de análisis, desenlace, horizonte y decisión prevista.
- Diagrama del flujo de datos desde la fuente hasta la métrica final.
- Evaluación de partición, fuga, datos faltantes, comparadores, calibración, incertidumbre y subgrupos.
- Lista de información o artefactos necesarios para reproducir el análisis.
- Valoración del riesgo de sesgo y aplicabilidad con PROBAST+AI cuando corresponda.
- Rediseño de una tabla y una figura para mejorar comunicación.
- Plan de validación externa en un contexto latinoamericano y estimación cualitativa de costos.
- Conclusión razonada: usar, validar, rediseñar o descartar el modelo, con condiciones explícitas.

Bibliografía general

Adisasmito, W. B., Almuhairi, S., Behravesh, C. B., Bilivogui, P., Bukachi, S. A., Casas, N., Cediel Becerra, N., Charron, D. F., Chaudhary, A., Ciacci Zanella, J. R., Cunningham, A. A., Dar, O., Debnath, N., Dungu, B., Farag, E., Gao, G. F., Hayman, D. T. S., Khaitsa, M., Koopmans, M. P. G., … Zhou, L. (2022). One Health: A new definition for a sustainable and healthy future. *PLOS Pathogens*, *18*(6), e1010537. https://doi.org/10.1371/journal.ppat.1010537

Aguilar-Lazcano, C. A., Espinosa-Curiel, I. E., Ríos-Martínez, J. A., Madera-Ramírez, F. A., & Pérez-Espinosa, H. (2023). Machine learning-based sensor data fusion for animal monitoring: Scoping review. *Sensors*, *23*(12), 5732. https://doi.org/10.3390/s23125732

Akinsulie, O. C., Idris, I., Aliyu, V. A., Shahzad, S., Banwo, O. G., Ogunleye, S. C., Olorunshola, M., Okedoyin, D. O., Ugwu, C., Oladapo, I. P., Gbadegoye, J. O., Akande, Q. A., Babawale, P., Rostami, S., & Soetan, K. O. (2024). The potential application of artificial intelligence in veterinary clinical practice and biomedical research. *Frontiers in Veterinary Science*, *11*, 1347550. https://doi.org/10.3389/fvets.2024.1347550

Alonso Sousa, S., Bukhari, S. S. U. H., Steagall, P. V., Bęczkowski, P. M., Giuliano, A., & Flay, K. J. (2025). Performance of large language models on veterinary undergraduate multiple-choice examinations: A comparative evaluation. *Frontiers in Veterinary Science*, *12*, 1616566. https://doi.org/10.3389/fvets.2025.1616566

Alsahaf, A., Gheorghe, R., Hidalgo, A. M., Petkov, N., & Azzopardi, G. (2023). Pre-insemination prediction of dystocia in dairy cattle. *Preventive Veterinary Medicine*, *210*, 105812. https://doi.org/10.1016/j.prevetmed.2022.105812

Antonucci, F., & Costa, C. (2020). Precision aquaculture: A short review on engineering innovations. *Aquaculture International*, *28*, 41–57. https://doi.org/10.1007/s10499-019-00443-w

Appleby, R. B., & Basran, P. S. (2022). Artificial intelligence in veterinary medicine. *Journal of the American Veterinary Medical Association, 260*(8), 819–824. https://doi.org/10.2460/javma.22.03.0093

Appleby, R. B., DiFazio, M., Cassel, N., Hennessey, R., & Basran, P. S. (2025). American College of Veterinary Radiology and European College of Veterinary Diagnostic Imaging position statement on artificial intelligence. *Journal of the American Veterinary Medical Association, 263*(6), 773–776. https://doi.org/10.2460/javma.25.01.0027

Asamblea Nacional del Ecuador. (2021). *Ley Orgánica de Protección de Datos Personales.* Registro Oficial, Quinto Suplemento No. 459, 26 de mayo de 2021. https://www.asambleanacional.gob.ec/es/multimedios-legislativos/63464-ley-organica-de-proteccion-de-datos

Banzato, T., Wodzinski, M., Burti, S., Osti, V. L., Rossoni, V., Atzori, M., & Zotti, A. (2021). Automatic classification of canine thoracic radiographs using deep learning. *Scientific Reports, 11*, 3964. https://doi.org/10.1038/s41598-021-83515-3

Banzato, T., Wodzinski, M., Burti, S., Vettore, E., Muller, H., & Zotti, A. (2023). An AI-based algorithm for the automatic evaluation of image quality in canine thoracic radiographs. *Scientific Reports, 13*, 17024. https://doi.org/10.1038/s41598-023-44089-4

Bao, J., & Xie, Q. (2022). Artificial intelligence in animal farming: A systematic literature review. *Journal of Cleaner Production, 331*, 129956. https://doi.org/10.1016/j.jclepro.2021.129956

Barker, M., Chue Hong, N. P., Katz, D. S., Lamprecht, A.-L., Martinez-Ortiz, C., Psomopoulos, F., Harrow, J., Castro, L. J., Gruenpeter, M., Martinez, P. A., & Honeyman, T. (2022). Introducing the FAIR Principles for research software. *Scientific Data, 9*, 622. https://doi.org/10.1038/s41597-022-01710-x

Basran, P. S., & Appleby, R. B. (2022). The unmet potential of artificial intelligence in veterinary medicine. *American Journal of Veterinary Research, 83*(5), 385–392. https://doi.org/10.2460/ajvr.22.03.0038

Basran, P. S., & Appleby, R. B. (2024). What's in the box? A toolbox for safe deployment of artificial intelligence in veterinary medicine. *Journal of the American Veterinary Medical Association, 262*(8), 1090–1098. https://doi.org/10.2460/javma.24.01.0027

Bellamy, J. E. C. (2023). Artificial intelligence in veterinary medicine requires regulation. *Canadian Veterinary Journal, 64*(10), 968–970. https://pmc.ncbi.nlm.nih.gov/articles/PMC10506349/

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). *On the dangers of stochastic parrots: Can language models be too big? Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, 610–623.* https://doi.org/10.1145/3442188.3445922

Berckmans, D. (2017). General introduction to precision livestock farming. *Animal Frontiers, 7*(1), 6–11. https://doi.org/10.2527/af.2017.0102

Bertout, J. A., Baneux, P. J. R., & Robertson-Plouch, C. K. (2021). Recommendations for ethical review of veterinary clinical trials. *Frontiers in Veterinary Science, 8*, 715926. https://doi.org/10.3389/fvets.2021.715926

Bertram, C. A., Donovan, T. A., Tecilla, M., Bartenschlager, F., Fragoso, M., Wilm, F., Marzahl, C., Breininger, K., Maier, A., Klopfleisch, R., & Aubreville, M. (2021). *Dataset on bi- and multi-nucleated tumor cells in canine cutaneous mast cell tumors.* In Bildverarbeitung für die Medizin 2021. Springer Vieweg. https://doi.org/10.1007/978-3-658-33198-6_33

Bobbo, T., Biffani, S., Taccioli, C., Penasa, M., & Cassandro, M. (2021). Comparison of machine learning methods to predict udder health status based on somatic cell counts in dairy cows. *Scientific Reports, 11*, 13642. https://doi.org/10.1038/s41598-021-93056-4

Boguslav, M. R., Kiehl, A., Kott, D., Strecker, G. J., Webb, T. L., Saklou, N., Ward, T., & Kirby, M. (2026). Fine-tuning foundational models to code diagnoses from veterinary health records. *PLOS Digital Health, 5*(2), e0001147. https://doi.org/10.1371/journal.pdig.0001147

Bouhali, O., Bensmail, H., Sheharyar, A., David, F., & Johnson, J. P. (2022). A review of radiomics and artificial intelligence and their application in veterinary diagnostic imaging. *Veterinary Sciences, 9*(11), 620. https://doi.org/10.3390/vetsci9110620

Bovo, M., Agrusti, M., Benni, S., Torreggiani, D., & Tassinari, P. (2021). Random forest modelling of milk yield of dairy cows under heat stress conditions. *Animals, 11*(5), 1305. https://doi.org/10.3390/ani11051305

Breiman, L. (2001). Random forests. *Machine Learning, 45*, 5–32. https://doi.org/10.1023/A:1010933404324

Bryan, J. (2018). *Excuse me, do you have a moment to talk about version control? The American Statistician, 72(1), 20–27.* https://doi.org/10.1080/00031305.2017.1399928

Burton, J., Farrell, S., Mäntylä Noble, P.-J., & Al Moubayed, N. (2024). Explainable text-tabular models for predicting mortality risk in companion animals. *Scientific Reports, 14*, 14217. https://doi.org/10.1038/s41598-024-64551-1

Buschi, D., Curti, N., Cola, V., Carlini, G., Sala, C., Dall'Olio, D., Castellani, G., Pizzi, E., Del Magno, S., Foglia, A., Giunti, M., Pisoni, L., & Giampieri, E. (2023). Automated wound image segmentation: Transfer learning from human to pet via active semi-supervised learning. *Animals, 13*(6), 956. https://doi.org/10.3390/ani13060956

Cabezas, J., Yubero, R., Visitación, B., Navarro-García, J., Algar, M. J., Cano, E. L., & Ortega, F. (2022). Analysis of accelerometer and GPS data for cattle behaviour identification and anomalous events detection. *Entropy, 24*(3), 336. https://doi.org/10.3390/e24030336

Carlson, C. J., Albery, G. F., Merow, C., Trisos, C. H., Zipfel, C. M., Eskew, E. A., Olival, K. J., Ross, N., & Bansal, S. (2022). Climate change increases cross-species viral transmission risk. *Nature, 607*, 555-562. https://doi.org/10.1038/s41586-022-04788-w

Catalano, S., Battelli, F., Traore, Z. I., Raghwani, J., Faust, C. L., & Standley, C. J. (2024). Pathogen genomics and One Health: A scoping review of current practices in zoonotic disease research. *IJID One Health, 4*, 100031. https://doi.org/10.1016/j.ijidoh.2024.100031

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research, 16*, 321–357. https://doi.org/10.1613/jair.953

Chen, L., Thorup, V. M., Kudahl, A. B., & Østergaard, S. (2024). Effects of heat stress on feed intake, milk yield, milk composition, and feed efficiency in dairy cows: A meta-analysis. *Journal of Dairy Science, 107*(5), 3207–3218. https://doi.org/10.3168/jds.2023-24059

Chen, T., & Guestrin, C. (2016). *XGBoost: A scalable tree boosting system.* In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785–794). Association for Computing Machinery. https://doi.org/10.1145/2939672.2939785

Chu, C. P. (2024). ChatGPT in veterinary medicine: A practical guidance of generative artificial intelligence in clinics, education, and research. *Frontiers in Veterinary Science, 11*, 1395934. https://doi.org/10.3389/fvets.2024.1395934

Coghlan, S., & Quinn, T. (2023). Ethics of using artificial intelligence in veterinary medicine. *AI & Society, 39*, 2337–2348. https://doi.org/10.1007/s00146-023-01686-1

Cohen, E. B., & Gordon, I. K. (2022). First, do no harm: Ethical and legal issues of artificial intelligence and machine learning in veterinary radiology and radiation oncology. *Veterinary Radiology & Ultrasound*, *63*(Suppl. 1), 840–850. https://doi.org/10.1111/vru.13171

Collins, G. S., Moons, K. G. M., Dhiman, P., Riley, R. D., Beam, A. L., Van Calster, B., Ghassemi, M., Liu, X., Reitsma, J. B., van Smeden, M., Boulesteix, A.-L., Camaradou, J. C., Celi, L. A., Denaxas, S., Denniston, A. K., Glocker, B., Golub, R. M., Harvey, H., Heinze, G., … Logullo, P. (2024). TRIPOD+AI statement: Updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ, 385*, e078378. https://doi.org/10.1136/bmj-2023-078378

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*, 273–297. https://doi.org/10.1007/BF00994018

Cruz Rivera, S., Liu, X., Chan, A.-W., Denniston, A. K., Calvert, M. J., & SPIRIT-AI and CONSORT-AI Working Group. (2020). Guidelines for clinical trial protocols for interventions involving artificial intelligence: The SPIRIT-AI extension. *Nature Medicine*, *26*(9), 1351–1363. https://doi.org/10.1038/s41591-020-1037-7

Curti, P. F., Selli, A., Pinto, D. L., Merlos-Ruiz, A., Balieiro, J. C. C., & Ventura, R. V. (2023). Applications of livestock monitoring devices and machine learning algorithms in animal production and reproduction: An overview. *Animal Reproduction*, *20*(2), e20230077. https://doi.org/10.1590/1984-3143-AR2023-0077

Davies, H., Nenadic, G., Alfattni, G., Arguello Casteleiro, M., Al Moubayed, N., Farrell, S., Radford, A. D., & Noble, P.-J. M. (2024). Text mining for disease surveillance in veterinary clinical data: Part two, training computers to identify features in clinical text. *Frontiers in Veterinary Science, 11*, 1352726. https://doi.org/10.3389/fvets.2024.1352726

Davies, H., Noble, P.-J. M., Fins, I. S., Pinchbeck, G., Singleton, D., Pirmohamed, M., & Killick, D. (2025). Developing electronic health records as a source of real-world data for veterinary pharmacoepidemiology. *Frontiers in Veterinary Science, 12*, 1550468. https://doi.org/10.3389/fvets.2025.1550468

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). *BERT: Pre-training of deep bidirectional transformers for language understanding.* Proceedings of NAACL-HLT 2019, 4171–4186. https://doi.org/10.18653/v1/N19-1423

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2021). *An image is worth 16×16 words: Transformers for image recognition at scale.* Proceedings of the International Conference on Learning Representations.

European Commission. (2026). *AI Act.* Shaping Europe's digital future. https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai

European Food Safety Authority & European Centre for Disease Prevention and Control. (2024). The European Union summary report on antimicrobial resistance in zoonotic and indicator bacteria from humans, animals and food in 2021-2022. *EFSA Journal, 22*(2), e8583. https://doi.org/10.2903/j.efsa.2024.8583

European Medicines Agency. (2022). *European veterinary big data strategy 2022–2027.* https://www.ema.europa.eu/en/documents/other/european-veterinary-big-data-strategy-2022-2027_en.pdf

European Parliament & Council of the European Union. (2024). *Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence.* Official Journal of the European Union. https://eur-lex.europa.eu/eli/reg/2024/1689/oj

Ezanno, P., Picault, S., Beaunée, G., Bailly, X., Muñoz, F., Duboz, R., Monod, H., & Guégan, J.-F. (2021). Research perspectives on animal health in the era of artificial intelligence. *Veterinary Research, 52*, 40. https://doi.org/10.1186/s13567-021-00902-4

Farrell, S., Appleton, C., Noble, P.-J. M., & Al Moubayed, N. (2023). PetBERT: Automated ICD-11 syndromic disease coding for outbreak detection in first opinion veterinary electronic health records. *Scientific Reports, 13*, 18015. https://doi.org/10.1038/s41598-023-45155-7

Festing, M. F. W. (2020). The "completely randomised" and the "randomised block" are the only experimental designs suitable for widespread use in pre-clinical research. *Scientific Reports, 10*, 17577. https://doi.org/10.1038/s41598-020-74538-3

Finlayson, S. G., Subbaswamy, A., Singh, K., Bowers, J., Kupke, A., Zittrain, J., Kohane, I. S., & Saria, S. (2021). The clinician and dataset shift in artificial intelligence. *New England Journal of Medicine, 385*(3), 283–286. https://doi.org/10.1056/NEJMc2104626

Food and Agriculture Organization of the United Nations. (2023). *Safe food for everyone: FAO's work on food safety.* FAO. https://www.fao.org/3/cc4347en/cc4347en.pdf

Food and Agriculture Organization of the United Nations, United Nations Environment Programme, World Health Organization, & World Organisation for Animal Health. (2022). *One Health joint plan of action 2022-2026: Working together for the health of humans, animals, plants and the environment.* World Health Organization. https://www.who.int/publications/i/item/9789240059139

Food and Agriculture Organization of the United Nations, United Nations Environment Programme, World Health Organization, & World Organisation for Animal Health. (2023). *A quadripartite global One Health intelligence framework.* FAO. https://openknowledge.fao.org/handle/20.500.14283/cc7621en

Food and Agriculture Organization of the United Nations, United Nations Environment Programme, World Health Organization, & World Organisation for Animal Health. (2025). *Quadripartite guidance on One Health integrated surveillance of antimicrobial resistance and use.* World Health Organization. https://www.who.int/publications/i/item/9789240109865

Food and Drug Administration. (2001). *CVM GFI #85 (VICH GL9): Good clinical practice.* https://www.fda.gov/regulatory-information/search-fda-guidance-documents/cvm-gfi-85-vich-gl9-good-clinical-practice

Food and Drug Administration. (2021). *Artificial intelligence/machine learning (AI/ML)-based software as a medical device (SaMD) action plan.* U.S. Department of Health and Human Services.

Fragoso-Garcia, M., Wilm, F., Bertram, C. A., Merz, S., Schmidt, A., Donovan, T., Fuchs-Baumgartinger, A., Bartel, A., Marzahl, C., Diehl, L., Puget, C., Maier, A., Aubreville, M., Breininger, K., & Klopfleisch, R. (2023). Automated diagnosis of 7 canine skin tumors using machine learning on H&E-stained whole slide images. *Veterinary Pathology, 60*(6), 865–875. https://doi.org/10.1177/03009858231189205

Frederick, C. E. (2024). Obtaining informed consent in veterinary clinical trials mini review. *Frontiers in Veterinary Science, 11*, 1426014. https://doi.org/10.3389/fvets.2024.1426014

Gao, C. A., Howard, F. M., Markov, N. S., Dyer, E. C., Ramesh, S., Luo, Y., & Pearson, A. T. (2023). Comparing scientific abstracts generated by ChatGPT to real abstracts with detectors and blinded human reviewers. *npj Digital Medicine, 6*, 75. https://doi.org/10.1038/s41746-023-00819-6

Ghassemi, M., Oakden-Rayner, L., & Beam, A. L. (2021). The false hope of current approaches to explainable artificial intelligence in health care. *The Lancet Digital Health, 3*(11), e745–e750. https://doi.org/10.1016/S2589-7500(21)00208-9

Gómez, Y., Stygar, A. H., Boumans, I. J. M. M., Bokkers, E. A. M., Pedersen, L. J., Niemi, J. K., Pastell, M., Manteca, X., & Llonch, P. (2021). A systematic review on validated precision livestock farming technologies for pig production and its potential to assess animal welfare. *Frontiers in Veterinary Science, 8*, 660565. https://doi.org/10.3389/fvets.2021.660565

Grange, Z. L., Goldstein, T., Johnson, C. K., Anthony, S., Gilardi, K., Daszak, P., Olival, K. J., O'Rourke, T., Murray, S., Olson, S. H., Togami, E., Vidal, G., Expert Panel, & Mazet, J. A. K. (2021). Ranking the risk of animal-to-human spillover for newly discovered viruses. *Proceedings of the National Academy of Sciences, 118*(15), e2002324118. https://doi.org/10.1073/pnas.2002324118

Grubb, T., Sager, J., Gaynor, J. S., Montgomery, E., Parker, J. A., Shafford, H., & Tearney, C. (2020). 2020 AAHA anesthesia and monitoring guidelines for dogs and cats. *Journal of the American Animal Hospital Association, 56*(2), 59–82. https://doi.org/10.5326/JAAHA-MS-7055

Gruen, M. E., Lascelles, B. D. X., Colleran, E., Gottlieb, A., Johnson, J., Lotsikas, P., Marcellin-Little, D., & Wright, B. (2022). 2022 AAHA pain management guidelines for dogs and cats. *Journal of the American Animal Hospital Association, 58*(2), 55–76. https://doi.org/10.5326/JAAHA-MS-7292

Hao, W., Zhang, K., Zhang, L., Han, M., Hao, W., Li, F., & Yang, G. (2023). TSML: A new pig behavior recognition method based on two-stream mutual learning network. *Sensors, 23*(11), 5092. https://doi.org/10.3390/s23115092

Health Level Seven International. (2023). *FHIR Release 5*. https://hl7.org/fhir/

Heinlein, M. (2026). Ethical considerations of artificial intelligence in veterinary medicine decision-making. *Frontiers in Veterinary Science, 13*, 1780868. https://doi.org/10.3389/fvets.2026.1780868

Hennessey, E., DiFazio, M., Hennessey, R., & Cassel, N. (2022). Artificial intelligence in veterinary diagnostic imaging: A literature review. *Veterinary Radiology & Ultrasound, 63*(S1), 851–870. https://doi.org/10.1111/vru.13163

Hernán, M. A., & Robins, J. M. (2020). *Causal inference: What if.* Chapman & Hall/CRC.

Hyndman, R. J., & Athanasopoulos, G. (2021). *Forecasting: Principles and practice (3rd ed.).* OTexts. https://otexts.com/fpp3/

Intergovernmental Panel on Climate Change. (2023). *Climate change 2023: Synthesis report.* https://doi.org/10.59327/IPCC/AR6-9789291691647

International Committee of Medical Journal Editors. (2026). *Recommendations for the conduct, reporting, editing, and publication of scholarly work in medical journals: Use of artificial intelligence in publishing.* https://www.icmje.org/recommendations/browse/artificial-intelligence/

International Committee of Medical Journal Editors. (2026). *Use of artificial intelligence in publishing: Use of AI by authors.*

https://www.icmje.org/recommendations/browse/artificial-intelligence/ai-use-by-authors.html

James, G., Witten, D., Hastie, T., Tibshirani, R., & Taylor, J. (2023). *An introduction to statistical learning: With applications in Python.* Springer. https://doi.org/10.1007/978-3-031-38747-0

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y., Chen, D., Dai, W., Chan, H. S., Madotto, A., & Fung, P. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys*, *55*(12), Article 248. https://doi.org/10.1145/3571730

Jiang, B., Tang, W., Cui, L., & Deng, X. (2023). Precision livestock farming research: A global scientometric review. *Animals, 13*(13), 2096. https://doi.org/10.3390/ani13132096

Kapoor, S., & Narayanan, A. (2023). Leakage and the reproducibility crisis in machine-learning-based science. *Patterns, 4*(9), 100804. https://doi.org/10.1016/j.patter.2023.100804

Kaur, U., Malacco, V. M. R., Bai, H., Price, T. P., Datta, A., Xin, L., Senecal, A. G., Voyles, R. M., & Brito, L. F. (2023). Invited review: Integration of technologies and systems for precision animal agriculture—A case study on precision dairy farming. *Journal of Animal Science, 101*, skad206. https://doi.org/10.1093/jas/skad206

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). *LightGBM: A highly efficient gradient boosting decision tree.* In Advances in Neural Information Processing Systems 30 (pp. 3146–3154).

Kelly, C. J., Karthikesalingam, A., Suleyman, M., Corrado, G., & King, D. (2019). Key challenges for delivering clinical impact with artificial intelligence. *BMC Medicine, 17*, 195. https://doi.org/10.1186/s12916-019-1426-2

Kim, J. Y., Han, M. G., Chun, J. H., Huh, E. A., & Lee, S. J. (2022). Developing a diagnosis model for dry eye disease in dogs using object detection. *Scientific Reports, 12*, 21351. https://doi.org/10.1038/s41598-022-25867-y

Kim, Y., Kim, J., Kim, S., Youn, H., Choi, J., & Seo, K. (2023). Machine learning-based risk prediction model for canine myxomatous mitral valve disease using electronic health record data. *Frontiers in Veterinary Science, 10*, 1189157. https://doi.org/10.3389/fvets.2023.1189157

Kong, C. Y., Vasquez, P., Farhoodimoghadam, M., Brandt, C., Brown, T. C., Reagan, K. L., Zwingenberger, A., & Keller, S. M. (2025). Anna: An open-source platform for real-time integration of machine learning classifiers with veterinary electronic health records. *BMC Veterinary Research, 21*, 557. https://doi.org/10.1186/s12917-025-05000-7

Leung, W. T. M., Rudge, J. W., & Fournié, G. (2023). Simulating contact networks for livestock disease epidemiology: A systematic review. *Journal of the Royal Society Interface, 20*(202), 20220890. https://doi.org/10.1098/rsif.2022.0890

Liu, X., Cruz Rivera, S., Moher, D., Calvert, M. J., Denniston, A. K., & SPIRIT-AI and CONSORT-AI Working Group. (2020). Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: The CONSORT-AI extension. *Nature Medicine, 26*(9), 1364–1374. https://doi.org/10.1038/s41591-020-1034-x

Lundberg, S. M., & Lee, S.-I. (2017). *A unified approach to interpreting model predictions.* In I. Guyon, U. von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), Advances in neural information processing systems 30 (pp. 4765–4774). Curran Associates.

Luo, W., Dong, Q., & Feng, Y. (2023). Risk prediction model of clinical mastitis in lactating dairy cows based on machine learning algorithms. *Preventive Veterinary Medicine, 221*, 106059. https://doi.org/10.1016/j.prevetmed.2023.106059

Martínez-López, B., Díaz-Cao, J. M., Clavijo, M. J., González-Crespo, C., Gómez-Vázquez, J. P., Rupasinghe, R., & Liu, X. (2025). Toward precision veterinary epidemiology: Applications, challenges, and opportunities of digitalization and the Big Data revolution in livestock health. *Journal of the American Veterinary Medical Association, 263*(5), 650-657. https://doi.org/10.2460/javma.25.01.0026

McDonald, A., Novo Matos, J., Silva, J., Partington, C., Lo, E. J. Y., Luis Fuentes, V., Barron, L., Watson, P., & Agarwal, A. (2024). A machine-learning algorithm to grade heart murmurs and stage preclinical myxomatous mitral valve disease in dogs. *Journal of Veterinary Internal Medicine, 38*(6), 2994–3004. https://doi.org/10.1111/jvim.17224

McInnes, L., Healy, J., & Melville, J. (2018). UMAP: Uniform manifold approximation and projection. *Journal of Open Source Software, 3*(29), 861. https://doi.org/10.21105/joss.00861

Mongan, J., Moy, L., & Kahn, C. E., Jr. (2020). Checklist for Artificial Intelligence in Medical Imaging (CLAIM): A guide for authors and reviewers. *Radiology: Artificial Intelligence, 2*(2), e200029. https://doi.org/10.1148/ryai.2020200029

Monteiro, B. P., Lascelles, B. D. X., Murrell, J., Robertson, S., Steagall, P. V. M., & Wright, B. (2023). 2022 WSAVA guidelines for the recognition, assessment and treatment of pain. *Journal of Small Animal Practice, 64*(4), 177–254. https://doi.org/10.1111/jsap.13566

Moons, K. G. M., Damen, J. A. A. G., Kaul, T., Hooft, L., Andaur Navarro, C. L., Dhiman, P., Beam, A. L., Van Calster, B., Ghassemi, M., Liu, X., Reitsma, J. B., van Smeden, M., Boulesteix, A. L., Camaradou, J. C., Celi, L. A., Denaxas, S., Denniston, A. K., Glocker, B., Golub, R. M., … Collins, G. S. (2025). PROBAST+AI: An updated quality, risk of bias, and applicability assessment tool for prediction models using regression or artificial intelligence methods. *BMJ, 388*, e082505. https://doi.org/10.1136/bmj-2024-082505

Moore, S. S., Costa, A., Penasa, M., Callegaro, S., & De Marchi, M. (2023). How heat stress conditions affect milk yield, composition, and price in Italian Holstein herds. *Journal of Dairy Science, 106*(6), 4042–4058. https://doi.org/10.3168/jds.2022-22640

Murray, C. J. L., Ikuta, K. S., Sharara, F., Swetschinski, L., Robles Aguilar, G., Gray, A., Han, C., Bisignano, C., Rao, P., Wool, E., Johnson, S. C., Browne, A. J., Chipeta, M. G., Fell, F., Hackett, S., Haines-Woodhouse, G., Kashef Hamadani, B. H., Kumaran, E. A. P., McManigal, B., … Naghavi, M. (2022). Global burden of bacterial antimicrobial resistance

in 2019: A systematic analysis. *The Lancet, 399*(10325), 629-655. https://doi.org/10.1016/S0140-6736(21)02724-0

National Academies of Sciences, Engineering, and Medicine. (2019). *Reproducibility and replicability in science.* The National Academies Press. https://doi.org/10.17226/25303

National Electrical Manufacturers Association. (n.d.). DICOM standard. Retrieved September 16, 2026, from https://www.dicomstandard.org/current

Neupane, R., Aryal, A., Haeussermann, A., Hartung, E., Pinedo, P., & Paudyal, S. (2024). Evaluating machine learning algorithms to predict lameness in dairy cattle. *PLOS ONE, 19*(7), e0301167. https://doi.org/10.1371/journal.pone.0301167

Niemeyer, F., Galbusera, F., Beukers, M., Jonas, R., Tao, Y., Fusellier, M., Tryfonidou, M. A., Neidlinger-Wilke, C., Kienle, A., & Wilke, H.-J. (2024). Automatic grading of intervertebral disc degeneration in lumbar dog spines. *JOR Spine, 7*(2), e1326. https://doi.org/10.1002/jsp2.1326

Norori, N., Hu, Q., Aellen, F. M., Faraci, F. D., & Tzovara, A. (2021). Addressing bias in big data and AI for health care: A call for open science. *Patterns, 2*(10), 100347. https://doi.org/10.1016/j.patter.2021.100347

O'Connor, A. M., Sargeant, J. M., Dohoo, I. R., Erb, H. N., Cevallos, M., Egger, M., Ersbøll, A. K., Martin, S. W., Nielsen, L. R., Pearl, D. L., Pfeiffer, D. U., Sanchez, J., Torrence, M. E., Vigre, H., Waldner, C., & Ward, M. P. (2016). Explanation and elaboration document for the STROBE-Vet statement: Strengthening the reporting of observational studies in epidemiology—Veterinary extension. *Journal of Veterinary Internal Medicine, 30*(6), 1896–1928. https://doi.org/10.1111/jvim.14592

Ojo, R. O., Ajayi, A. O., Owolabi, H. A., Oyedele, L. O., & Akanbi, L. A. (2022). Internet of Things and machine learning techniques in poultry health and welfare management: A systematic literature review. *Computers and Electronics in Agriculture, 200*, 107266. https://doi.org/10.1016/j.compag.2022.107266

Oren, A., Türkcü, J. D., Meller, S., Lazebnik, T., Wiegel, P., Mach, R., Volk, H. A., & Zamansky, A. (2023). BrachySound: Machine learning based assessment of respiratory sounds in dogs. *Scientific Reports, 13*, 20300. https://doi.org/10.1038/s41598-023-47308-0

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., … Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ, 372*, n71. https://doi.org/10.1136/bmj.n71

Pang, G., Shen, C., Cao, L., & van den Hengel, A. (2021). Deep learning for anomaly detection: A review. *ACM Computing Surveys, 54*(2), Article 38. https://doi.org/10.1145/3439950

Papakonstantinou, G. I., Voulgarakis, N., Terzidou, G., Fotos, L., Giamouri, E., & Papatsiros, V. G. (2024). Precision livestock farming technology: Applications and challenges of animal welfare and climate change. *Agriculture, 14*(4), 620. https://doi.org/10.3390/agriculture14040620

Parsons, I. L., Karisch, B. B., Stone, A. E., Webb, S. L., Norman, D. A., Street, G. M., & Dillard, S. L. (2024). Machine learning methods and visual observations to categorize behavior of grazing cattle using accelerometer signals. *Sensors, 24*(10), 3171. https://doi.org/10.3390/s24103171

Penev, Y. P. (2024). Electronic health record data quality and performance assessments: Scoping review. *JMIR Medical Informatics, 12*, e58130. https://doi.org/10.2196/58130

Percie du Sert, N., Ahluwalia, A., Alam, S., Avey, M. T., Baker, M., Browne, W. J., Clark, A., Cuthill, I. C., Dirnagl, U., Emerson, M., Garner, P., Holgate, S. T., Howells, D. W., Hurst, V., Karp, N. A., Lazic, S. E., Lidster, K., MacCallum, C. J., Macleod, M., & Petersen, O. H. (2020b). Reporting animal research: Explanation and elaboration for the ARRIVE guidelines 2.0. *PLOS Biology, 18*(7), e3000411. https://doi.org/10.1371/journal.pbio.3000411

Percie du Sert, N., Hurst, V., Ahluwalia, A., Alam, S., Avey, M. T., Baker, M., Browne, W. J., Clark, A., Cuthill, I. C., Dirnagl, U., Emerson, M., Garner, P., Holgate, S. T., Howells, D. W., Karp, N. A., Lazic, S. E., Lidster, K., MacCallum, C. J., Macleod, M., … Würbel, H. (2020a). The ARRIVE guidelines 2.0: Updated guidelines for reporting animal research. *PLOS Biology*, *18*(7), e3000410. https://doi.org/10.1371/journal.pbio.3000410

Pillai, N., Ramkumar, M., & Nanduri, B. (2022). Artificial intelligence models for zoonotic pathogens: A survey. *Microorganisms*, *10*(10), 1911. https://doi.org/10.3390/microorganisms10101911

Piper, S. K., Zocholl, D., Toelch, U., Roehle, R., Stroux, A., Hoessler, J., Zinke, A., & Konietschke, F. (2023). Statistical review of animal trials—A guideline. *Biometrical Journal*, *65*(2), 2200061. https://doi.org/10.1002/bimj.202200061

Presidencia de la República del Ecuador. (2023). *Reglamento General a la Ley Orgánica de Protección de Datos Personales (Decreto Ejecutivo No.* 904). Registro Oficial, Tercer Suplemento No. 435, 13 de noviembre de 2023. https://www.telecomunicaciones.gob.ec/reglamento-ley-organica-de-proteccion-de-datos-personales/

Reagan, K. L., Deng, S., Sheng, J., Sebastian, J., Wang, Z., Huebner, S. N., Wenke, L. A., Michalak, S. R., Strohmer, T., & Sykes, J. E. (2022). Use of machine-learning algorithms to aid in the early detection of leptospirosis in dogs. *Journal of Veterinary Diagnostic Investigation*, *34*(4), 612–621. https://doi.org/10.1177/10406387221096781

Redondo, J. I., Rubio, M., Soler, G., Serra, I., Soler, C., & Gómez-Villamandos, R. J. (2024). Anaesthetic mortality in dogs: A worldwide analysis and risk assessment. *Veterinary Record*, *194*, e3604. https://doi.org/10.1002/vetr.3604

Regenstrief Institute. (n.d.). LOINC: The international standard for identifying health measurements, observations, and documents. Retrieved September 16, 2026, from https://loinc.org/

Rethlefsen, M. L., Kirtley, S., Waffenschmidt, S., Ayala, A. P., Moher, D., Page, M. J., Koffel, J. B., & PRISMA-S Group. (2021). PRISMA-S: An extension to the PRISMA statement for reporting literature searches in systematic reviews. *Systematic Reviews, 10*, 39. https://doi.org/10.1186/s13643-020-01542-z

Riley, R. D., Ensor, J., Snell, K. I. E., Harrell, F. E., Jr., Martin, G. P., Reitsma, J. B., Moons, K. G. M., Collins, G., & van Smeden, M. (2020). Calculating the sample size required for developing a clinical prediction model. *BMJ, 368*, m441. https://doi.org/10.1136/bmj.m441

Riley, R. D., Snell, K. I. E., Ensor, J., Burke, D. L., Harrell, F. E., Jr., Moons, K. G. M., & Collins, G. S. (2019). Minimum sample size for developing a multivariable prediction model: Part II—Binary and time-to-event outcomes. *Statistics in Medicine, 38*(7), 1276–1296. https://doi.org/10.1002/sim.7992

Roberts, M., Driggs, D., Thorpe, M., Gilbey, J., Yeung, M., Ursprung, S., Aviles-Rivero, A. I., Etmann, C., McCague, C., Beer, L., Weir-McCall, J. R., Teng, Z., Gkrania-Klotsas, E., Rudd, J. H. F., Sala, E., & Schönlieb, C.-B. (2021). Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans. *Nature Machine Intelligence, 3*, 199–217. https://doi.org/10.1038/s42256-021-00307-0

Russell, S. J., & Norvig, P. (2021). *Artificial intelligence: A modern approach (4th ed.).* Pearson.

Saito, T., & Rehmsmeier, M. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLOS ONE, 10*(3), e0118432. https://doi.org/10.1371/journal.pone.0118432

Salvi, M., Molinari, F., Iussich, S., Muscatello, L. V., Pazzini, L., Benali, S., Banco, B., Abramo, F., De Maria, R., & Aresu, L. (2021). Histopathological classification of canine cutaneous round cell tumors using deep learning: A multi-center study. *Frontiers in Veterinary Science, 8*, 640944. https://doi.org/10.3389/fvets.2021.640944

Sandve, G. K., Nekrutenko, A., Taylor, J., & Hovig, E. (2013). Ten simple rules for reproducible computational research. *PLOS Computational Biology, 9*(10), e1003285. https://doi.org/10.1371/journal.pcbi.1003285

Sargeant, J. M., O'Connor, A. M., Dohoo, I. R., Erb, H. N., Cevallos, M., Egger, M., Ersbøll, A. K., Martin, S. W., Nielsen, L. R., Pearl, D. L., Pfeiffer, D. U., Sanchez, J., Torrence, M. E., Vigre, H., Waldner, C., & Ward, M. P. (2016). Methods and processes of developing the Strengthening the Reporting of Observational Studies in Epidemiology—Veterinary (STROBE-Vet) statement. *Zoonoses and Public Health, 63*(8), 651–661. https://doi.org/10.1111/zph.12314

Schillings, J., Bennett, R., & Rose, D. C. (2021). Animal welfare and other ethical implications of precision livestock farming technology. *CABI Agriculture and Bioscience, 2*, 17. https://doi.org/10.1186/s43170-021-00037-8

Shim, H., Lee, J., Choi, S., Kim, J., Jeong, J., Cho, C., Kim, H., Kim, J.-I., Kim, J., & Eom, K. (2023). Deep learning-based diagnosis of stifle joint diseases in dogs. *Veterinary Radiology & Ultrasound, 64*(1), 113–122. https://doi.org/10.1111/vru.13181

Smith, A. J., Clutton, R. E., Lilley, E., Hansen, K. E. A., & Brattelid, T. (2018). PREPARE: Guidelines for planning animal research and testing. *Laboratory Animals, 52*(2), 135–141. https://doi.org/10.1177/0023677217724823

Sounderajah, V., Guni, A., Liu, X., Collins, G. S., Karthikesalingam, A., Markar, S. R., Golub, R. M., Denniston, A. K., Shetty, S., Moher, D., Bossuyt, P. M., Darzi, A., Ashrafian, H., & STARD-AI Steering Committee. (2025). The STARD-AI reporting guideline for diagnostic accuracy studies using artificial intelligence. *Nature Medicine, 31*(10), 3283–3289. https://doi.org/10.1038/s41591-025-03953-8

Steagall, P. V., Monteiro, B. P., Marangoni, S., Moussa, M., & Sautié, M. (2023). Fully automated deep learning models with smartphone applicability for prediction of pain using the Feline Grimace Scale. *Scientific Reports, 13*, 21584. https://doi.org/10.1038/s41598-023-49031-2

Steyerberg, E. W. (2019). *Clinical prediction models: A practical approach to development, validation, and updating (2nd ed.).* Springer. https://doi.org/10.1007/978-3-030-16399-0

Stimmer, L., & Kuiper, R. V. (2025). *Natural language processing in veterinary pathology: A commentary on opportunities, challenges, and future directions.* Veterinary Pathology. Advance online publication. https://doi.org/10.1177/03009858251343215

Stygar, A. H., Gómez, Y., Berteselli, G. V., Dalla Costa, E., Canali, E., Niemi, J. K., Llonch, P., & Pastell, M. (2021). A systematic review on commercially available and validated sensor technologies for welfare assessment of dairy cattle. *Frontiers in Veterinary Science, 8*, 634338. https://doi.org/10.3389/fvets.2021.634338

Szlosek, D., Coyne, M., Riggott, J., Knight, K., McCrann, D. J., & Kincaid, D. (2024). Development and validation of a machine learning model for clinical wellness visit classification in cats and dogs. *Frontiers in Veterinary Science, 11*, 1348162. https://doi.org/10.3389/fvets.2024.1348162

Tabassi, E. (2023). *Artificial intelligence risk management framework (AI RMF 1.0) (NIST AI 100-1).* National Institute of Standards and Technology. https://doi.org/10.6028/NIST.AI.100-1

Tejani, A. S., Klontzas, M. E., Gatti, A. A., Mongan, J. T., Moy, L., Park, S. H., Kahn, C. E., & CLAIM 2024 Update Panel. (2024). Checklist for Artificial Intelligence in Medical Imaging (CLAIM): 2024 update. *Radiology: Artificial Intelligence, 6*(4), e240300. https://doi.org/10.1148/ryai.240300

Thompson, J. S., Green, M. J., Hyde, R., Bradley, A. J., & O'Grady, L. (2023). The use of machine learning to predict somatic cell count status in dairy cows post-calving. *Frontiers in Veterinary Science, 10*, 1297750. https://doi.org/10.3389/fvets.2023.1297750

Tian, H., Zhou, X., Wang, H., Xu, C., Zhao, Z., Xu, W., & Deng, Z. (2024). The prediction of clinical mastitis in dairy cows based on milk yield, rumination time, and milk electrical

conductivity using machine learning algorithms. *Animals, 14*(3), 427. https://doi.org/10.3390/ani14030427

Tuyttens, F. A. M., Molento, C. F. M., & Benaissa, S. (2022). Twelve threats of precision livestock farming for animal welfare. *Frontiers in Veterinary Science, 9*, 889623. https://doi.org/10.3389/fvets.2022.889623

UNESCO. (2021). *Recommendation on the ethics of artificial intelligence.* https://unesdoc.unesco.org/ark:/48223/pf0000381137

UNESCO. (2021). *UNESCO recommendation on open science.* UNESCO.

van Buuren, S. (2018). *Flexible imputation of missing data (2nd ed.).* CRC Press. https://doi.org/10.1201/9780429492259

Van Calster, B., McLernon, D. J., van Smeden, M., Wynants, L., & Steyerberg, E. W. (2019). Calibration: The Achilles heel of predictive analytics. *BMC Medicine, 17,* 230. https://doi.org/10.1186/s12916-019-1466-7

van de Schoot, R., de Bruin, J., Schram, R., Zahedi, P., de Boer, J., Weijdema, F., Kramer, B., Huijts, M., Hoogerwerf, M., Ferdinands, G., Harkema, A., Willemsen, J., Ma, Y., Fang, Q., Hindriks, S., Tummers, L., & Oberski, D. (2021). An open source machine learning framework for efficient and transparent systematic reviews. *Nature Machine Intelligence, 3*(2), 125–133. https://doi.org/10.1038/s42256-020-00287-7

van Dis, E. A. M., Bollen, J., Zuidema, W., van Rooij, R., & Bockting, C. L. (2023). ChatGPT: Five priorities for research. *Nature, 614,* 224–226. https://doi.org/10.1038/d41586-023-00288-7

Varkoulis, K., Savvas, I., Anagnostou, T., Kazakos, G., & Pavlidou, K. (2023). A retrospective study on canine and feline mortality during anaesthesia at a university clinic in Greece. *Animals, 13*(15), 2486. https://doi.org/10.3390/ani13152486

Vasey, B., Nagendran, M., Campbell, B., Clifton, D. A., Collins, G. S., Denaxas, S., Denniston, A. K., Faes, L., Geerts, B., Ibrahim, M., Liu, X., Mateen, B. A., Mathur, P., McCradden, M. D., Morgan, L., Ordish, J., Rogers, C., Saria, S., Ting, D. S. W., … DECIDE-AI Expert Group. (2022). Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. *Nature Medicine*, *28*(5), 924–933. https://doi.org/10.1038/s41591-022-01772-9

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, *30*, 5998–6008.

VeNom Coding Group. (n.d.). VeNom codes. https://www.vetxml.co.uk/venom-codes/

Veterinary Terminology Services Laboratory. (2026). *Veterinary extension of SNOMED CT*. Virginia-Maryland College of Veterinary Medicine. https://vtsl.vetmed.vt.edu/

Vickers, A. J., & Elkin, E. B. (2006). Decision curve analysis: A novel method for evaluating prediction models. *Medical Decision Making*, *26*(6), 565–574. https://doi.org/10.1177/0272989X06295361

Vidal, G., Sharpnack, J., Pinedo, P., Tsai, I.-C., Lee, A. R., Martínez-López, B., & Aly, S. S. (2023). Impact of sensor data pre-processing strategies and selection of machine learning algorithm on the prediction of metritis events in dairy cattle. *Preventive Veterinary Medicine*, *214*, 105903. https://doi.org/10.1016/j.prevetmed.2023.105903

Villarino, N. F. (2025). Recommendations for a complete reporting of statistical methods in veterinary pharmacology. *Journal of Veterinary Pharmacology and Therapeutics*, *48*(4), 221–233. https://doi.org/10.1111/jvp.70001

Walker, M. A., Tan, L. M., Dang, L. H., Van Khang, P., Ha, H. T. T., & Blackburn, J. K. (2022). Spatiotemporal patterns of anthrax, Vietnam, 1990–2015. *Emerging Infectious Diseases*, *28*(11), 2206–2213. https://doi.org/10.3201/eid2811.212584

Wang, Z., Cao, Q., Liu, Q., Dufe, D., Wouters, P., Deng, L., & Pong, A. (2025). How many subjects are enough in a veterinary trial? Literature review and insights from industrial statisticians. *Research in Veterinary Science, 186*, 105569. https://doi.org/10.1016/j.rvsc.2025.105569

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA statement on p-values: Context, process, and purpose. *The American Statistician, 70*(2), 129–133. https://doi.org/10.1080/00031305.2016.1154108

Weng, H. Y., & Messam, L. L. M. (2025). Reporting and interpreting statistical results in veterinary medicine: Calling for change. *Journal of Veterinary Internal Medicine, 39*(1), e17258. https://doi.org/10.1111/jvim.17258

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., … Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data, 3*, 160018. https://doi.org/10.1038/sdata.2016.18

Wilson, G., Bryan, J., Cranston, K., Kitzes, J., Nederbragt, L., & Teal, T. K. (2017). Good enough practices in scientific computing. *PLOS Computational Biology, 13*(6), e1005510. https://doi.org/10.1371/journal.pcbi.1005510

World Health Organization. (2021). *Ethics and governance of artificial intelligence for health: WHO guidance.* https://www.who.int/publications/i/item/9789240029200

World Health Organization. (2022). *Global antimicrobial resistance and use surveillance system GLASS report 2022.* World Health Organization. https://www.who.int/publications/i/item/9789240062702

World Health Organization. (2024). *Ethics and governance of artificial intelligence for health: Guidance on large multi-modal models.* https://www.who.int/publications/i/item/9789240084759

World Organisation for Animal Health. (2024). *Annual report on antimicrobial agents intended for use in animals.* Eighth report. WOAH.

World Organisation for Animal Health. (2024). *Use of animals in research and education (Chapter 7.8).* In Terrestrial Animal Health Code. https://www.woah.org/fileadmin/Home/eng/Health_standards/tahc/2024/en_chapitre_aw_research_education.htm

Wulcan, J. M., Jacques, K. L., Lee, M. A., Kovacs, S. L., Dausend, N., Prince, L. E., Wulcan, J., Marsilio, S., & Keller, S. M. (2025). Classification performance and reproducibility of GPT-4 omni for information extraction from veterinary electronic health records. *Frontiers in Veterinary Science, 11*, 1490030. https://doi.org/10.3389/fvets.2024.1490030

Xiao, S., Dhand, N. K., Wang, Z., Hu, K., Thomson, P. C., House, J. K., & Khatkar, M. S. (2025). Review of applications of deep learning in veterinary diagnostics and animal health. *Frontiers in Veterinary Science, 12*, 1511522. https://doi.org/10.3389/fvets.2025.1511522

Yang, L., Liu, Y., Yu, H., Fang, X., Song, L., Li, D., & Chen, Y. (2021). Computer vision models in intelligent aquaculture with emphasis on fish detection and behavior analysis: A review. *Archives of Computational Methods in Engineering, 28*, 2785–2816. https://doi.org/10.1007/s11831-020-09486-2

Yang, X., Bist, R. B., Paneru, B., Liu, T., Applegate, T., Ritz, C., Kim, W., Regmi, P., & Chai, L. (2024). Computer vision-based cybernetics systems for promoting modern poultry farming: A critical review. *Computers and Electronics in Agriculture, 226*, 109339. https://doi.org/10.1016/j.compag.2024.109339

Zielinski, C., Winker, M., Aggarwal, R., Ferris, L., Heinemann, M., Lapeña, J. F., Pai, S. A., Ing, E., Citrome, L., Alam, M., Voight, M., Habibzadeh, F., & Bazdaric, K. (2023). Chatbots, generative AI, and scholarly manuscripts: WAME recommendations on chatbots and generative artificial intelligence in relation to scholarly publications. *Current Medical Research and Opinion, 40*(1), 11. https://doi.org/10.1080/03007995.2023.2286102

ANEXOS

Glosario de términos

Las definiciones se formulan para investigación veterinaria. Cuando un término posee un significado estadístico estricto, se diferencia de su uso cotidiano o computacional.

Abstención del modelo. Capacidad programada para no emitir una clasificación cuando la entrada es insuficiente, está fuera del dominio previsto o la incertidumbre supera un límite de seguridad.

Acelerómetro. Sensor que registra cambios de aceleración y permite derivar actividad, postura, locomoción o patrones de conducta; sus salidas requieren validación contra observación o medición de referencia.

Ajuste de hiperparámetros. Selección de configuraciones que controlan el aprendizaje, como profundidad de árboles, penalización o tasa de aprendizaje, utilizando únicamente datos de desarrollo.

Aleatorización. Asignación probabilística de unidades experimentales a tratamientos para reducir sesgo de selección y sostener comparabilidad causal.

Algoritmo. Conjunto definido de operaciones que transforma entradas en resultados. El término no implica autonomía, comprensión ni validez científica.

Algoritmo adaptativo. Sistema cuyos parámetros pueden cambiar después del despliegue. Exige control de versiones, límites de actualización, reevaluación y vigilancia reforzada.

Alucinación. Salida generativa plausible pero falsa, no sustentada o internamente inventada, incluidas referencias, datos, resultados o explicaciones.

Anonimización. Transformación destinada a impedir razonablemente la reidentificación. No equivale a eliminar solo el nombre o el número de historia clínica.

Análisis de decisión. Evaluación que combina rendimiento predictivo, umbrales y consecuencias de acciones para estimar utilidad clínica o sanitaria.

Análisis de sensibilidad. Conjunto de análisis que examina cuánto cambian las conclusiones ante supuestos, definiciones, datos faltantes, valores extremos o decisiones analíticas alternativas.

Aprendizaje automático. Métodos que estiman patrones a partir de datos para clasificar, predecir, agrupar o representar información.

Aprendizaje no supervisado. Análisis sin una variable objetivo etiquetada, usado para explorar estructura, grupos, dimensiones o anomalías.

Aprendizaje por refuerzo. Marco en el que un agente aprende acciones mediante recompensas. Su aplicación sobre animales exige restricciones éticas y de seguridad antes de cualquier interacción real.

Aprendizaje supervisado. Entrenamiento con pares de predictores y resultados conocidos para estimar una salida en casos nuevos.

Aumento de datos. Creación de variantes de observaciones durante el entrenamiento. Debe conservar plausibilidad anatómica, biológica y de adquisición.

AUPRC. Área bajo la curva precisión-recall. Resulta informativa cuando la clase positiva es rara y se desea valorar la relación entre sensibilidad y valor predictivo positivo.

Base de conveniencia. Conjunto disponible por facilidad de acceso, cuya representatividad no está garantizada y debe justificarse.

Benchmark. Conjunto, tarea o procedimiento usado para comparar métodos bajo reglas comunes; no reemplaza la validación en la población de uso.

Bioestadística. Disciplina que aplica razonamiento estadístico al diseño, medición, análisis e interpretación de fenómenos biológicos, clínicos y poblacionales.

Bloqueo. Agrupación de unidades semejantes antes de aleatorizar para controlar una fuente conocida de variabilidad, como centro, camada o peso basal.

Calibración. Concordancia entre probabilidades predichas y frecuencias observadas. Una predicción del 20 % está bien calibrada si el evento ocurre cerca del 20 % en casos comparables.

Cambio de dominio. Diferencia entre los datos de desarrollo y los de aplicación, causada por especie, raza, centro, dispositivo, protocolo, época o prevalencia.

Causalidad. Relación referida al resultado que habría ocurrido bajo intervenciones alternativas. Requiere diseño y supuestos causales, no solo asociación o importancia predictiva.

Censura. Situación en análisis de supervivencia en la que el tiempo del evento no se observa por completo durante el seguimiento.

Ciencia abierta. Prácticas que favorecen transparencia, acceso responsable, trazabilidad, colaboración y reutilización de protocolos, materiales, datos y código.

Clasificación. Predicción de una categoría, como positivo/negativo, grado histológico o nivel de riesgo.

Clustering. Agrupamiento exploratorio según similitud. Los grupos obtenidos no son automáticamente fenotipos ni entidades biológicas naturales.

CNN. Red neuronal convolucional diseñada para aprender representaciones espaciales, usada con frecuencia en radiografías, ecografías, histología y video.

Colisionador. Variable causada por dos variables del sistema causal; condicionarla puede crear una asociación espuria.

Confusión. Distorsión de una asociación por causas comunes de exposición y desenlace que no han sido controladas adecuadamente.

Conjunto de prueba. Datos congelados y no usados para seleccionar variables, arquitectura, hiperparámetros, umbral o preprocesamiento final.

Control de versiones. Registro sistemático de cambios en datos, código, documentos, dependencias y modelos para reconstruir el estado exacto de un análisis.

Curva ROC. Relación entre sensibilidad y tasa de falsos positivos a través de múltiples umbrales.

Dato faltante. Valor no observado. Su tratamiento depende del mecanismo de ausencia, el diseño, la variable y el objetivo analítico.

Deep learning. Familia de redes neuronales con múltiples capas de representación, útil en datos complejos pero demandante en volumen, cómputo y validación.

Deriva. Cambio de distribución, relación predictor-desenlace o desempeño después del despliegue.

Desbalance de clases. Desigual frecuencia entre categorías del desenlace. No se resuelve informando solo exactitud ni aplicando sobremuestreo sin validación correcta.

Desenlace. Variable definida para representar el evento, estado o magnitud que responde la pregunta científica.

Detección. Localización de objetos o eventos, como nódulos en una imagen o episodios de cojera en una señal.

Diccionario de datos. Documento que especifica nombre, definición, tipo, unidad, codificación, dominio, momento, fuente y reglas de calidad de cada variable.

Discriminación. Capacidad de ordenar o separar unidades con y sin el desenlace; no informa si las probabilidades tienen magnitud correcta.

Diseño factorial. Diseño que combina niveles de dos o más factores y permite estimar efectos principales e interacciones.

Efecto clínicamente importante. Magnitud mínima que modificaría una decisión o tendría relevancia para el paciente, tutor, productor o sistema sanitario.

Embedding. Representación numérica densa de texto, imagen u otra entrada, diseñada para conservar relaciones útiles para una tarea.

Enmascaramiento. Procedimiento que impide que cuidadores, evaluadores, participantes o analistas conozcan la intervención cuando ello puede sesgar conductas o mediciones.

Especificidad. Proporción de verdaderos negativos entre unidades sin la condición de referencia.

Estimando. Cantidad precisa que se desea estimar, definida por población, comparación, desenlace, tiempo y manejo de eventos intercurrentes.

Exactitud. Proporción global de clasificaciones correctas. Puede resultar engañosa con clases desbalanceadas.

Explicabilidad. Métodos que describen factores asociados con la salida del modelo. No demuestran mecanismos biológicos ni efectos de intervención.

F1. Media armónica de precisión y recall; resume ambas pero no incorpora calibración, especificidad ni costos clínicos.

FAIR. Principios para que datos y objetos digitales sean localizables, accesibles bajo condiciones claras, interoperables y reutilizables.

Falso negativo. Caso con condición presente clasificado como negativo. Su gravedad depende de las consecuencias de retrasar o perder una acción.

Falso positivo. Caso sin condición clasificado como positivo, con posibles costos, pruebas, tratamiento, alarma o daño innecesarios.

Fuga de información. Uso directo o indirecto de información que no estaría disponible en el momento real de predicción o que conecta indebidamente desarrollo y evaluación.

GAM. Modelo aditivo generalizado que representa relaciones no lineales mediante funciones suaves penalizadas manteniendo una estructura interpretable.

Gemelo digital. Representación dinámica de un sistema real que integra estado, procesos y simulación para explorar evolución o intervenciones; no es solo un tablero de indicadores.

Generalización. Capacidad de mantener desempeño en observaciones nuevas pertenecientes al dominio de uso previsto.

Hiperparámetro. Configuración elegida fuera del ajuste directo de parámetros, como profundidad máxima, número de vecinos o intensidad de regularización.

Imputación. Estimación de valores faltantes bajo supuestos explícitos. Debe aprenderse dentro de los datos de entrenamiento para evitar fuga.

Inferencia. Proceso de estimar cantidades poblacionales y su incertidumbre a partir de datos y un diseño.

Interacción. Situación en la que el efecto o asociación de un factor depende del nivel de otro.

Interoperabilidad. Capacidad de intercambiar datos y conservar su significado mediante formatos, identificadores, vocabularios y reglas compatibles.

Intervalo de confianza. Rango de valores compatibles con los datos y el modelo a un nivel especificado; expresa precisión y no una probabilidad posterior del parámetro.

IoT. Internet de las cosas: red de sensores y dispositivos conectados que registran o transmiten datos de animales, ambiente o procesos.

LIME. Método de explicación local que aproxima el comportamiento del modelo cerca de una observación; depende de perturbaciones y configuración.

MAE. Error absoluto medio. Resume el tamaño promedio del error en las unidades originales y penaliza linealmente.

Matriz de confusión. Tabla de verdaderos positivos, falsos positivos, verdaderos negativos y falsos negativos para un umbral determinado.

Medidas repetidas. Observaciones tomadas varias veces sobre la misma unidad, por lo que no son independientes.

Metadatos. Información que describe procedencia, estructura, significado, condiciones de acceso, transformaciones y responsables de un conjunto de datos.

Modelo de lenguaje. Modelo probabilístico entrenado para representar y generar secuencias lingüísticas. Sus respuestas no son evidencia científica por sí mismas.

Modelo lineal mixto. Modelo que combina efectos fijos con efectos aleatorios para representar agrupamiento o repetición.

Multimodalidad. Integración de fuentes distintas, como texto, imagen, laboratorio, genómica y sensores. Requiere estudiar ausencia y desalineación entre modalidades.

NMDS. Escalamiento multidimensional no métrico que representa relaciones de disimilitud en pocas dimensiones; su calidad se evalúa, entre otros elementos, mediante stress.

One Health. Enfoque integrado que reconoce interdependencias entre salud animal, humana y ambiental.

Overfitting. Sobreajuste: aprendizaje de peculiaridades o ruido del desarrollo que no se mantienen en datos nuevos.

Partición agrupada. Separación de datos manteniendo juntas todas las observaciones correlacionadas de un animal, hato, centro, lote o episodio.

Parámetro. Cantidad estimada durante el ajuste del modelo, como un coeficiente de regresión o los pesos de una red.

PCA. Análisis de componentes principales que resume variación lineal en combinaciones ortogonales; su interpretación depende de escala y cargas.

PERMANOVA. Prueba por permutaciones para diferencias multivariadas entre grupos basada en distancias; debe examinar dispersión y respetar el diseño de permutación.

Precisión o valor predictivo positivo. Proporción de verdaderos positivos entre todas las alertas positivas; depende de prevalencia y contexto de aplicación.

Predicción. Estimación de una salida en casos nuevos. Puede ser útil sin identificar causas.

Preprocesamiento. Transformaciones previas al ajuste, como limpieza, imputación, codificación y escalamiento. Debe formar parte del flujo validado.

Prevalencia. Proporción de una población con la condición en un momento o periodo; modifica los valores predictivos.

Pseudorreplicación. Tratamiento de submediciones dependientes como si fueran réplicas experimentales independientes.

Random forest. Conjunto de árboles entrenados con remuestreo y selección aleatoria de variables, cuyas predicciones se agregan.

Recall. Sinónimo frecuente de sensibilidad en clasificación binaria: proporción de positivos detectados.

Red neuronal. Modelo compuesto por capas de transformaciones parametrizadas que aprende representaciones y relaciones complejas.

Regularización. Penalización o restricción destinada a controlar complejidad y mejorar estabilidad fuera de muestra.

Replicabilidad. Obtención de resultados consistentes en un estudio nuevo que aborda la misma pregunta con datos independientes.

Reproducibilidad. Capacidad de obtener los mismos resultados usando los mismos datos, código y flujo documentado.

RMSE. Raíz del error cuadrático medio. Se expresa en unidades del desenlace y penaliza con mayor intensidad los errores grandes.

Sensibilidad. Proporción de verdaderos positivos entre unidades que poseen la condición de referencia.

Serie temporal. Observaciones ordenadas en el tiempo cuya dependencia, estacionalidad y tendencia deben modelarse y validarse cronológicamente.

SHAP. Familia de métodos de atribución basados en valores de Shapley para explicar predicciones; sus resultados dependen del modelo, referencia y correlaciones.

Sobreajuste. Véase overfitting.

Transformer. Arquitectura basada en mecanismos de atención, empleada en lenguaje, visión y datos multimodales.

Trazabilidad. Capacidad de reconstruir origen, transformaciones, decisiones, versiones y responsables de datos y modelos.

Umbral. Valor de corte que convierte una puntuación o probabilidad en una acción o categoría; debe elegirse según consecuencias y capacidad operativa.

Unidad experimental. Unidad asignada independientemente a una intervención o condición.

Unidad observacional. Entidad sobre la que se registra una medición; puede ser distinta y estar anidada dentro de la unidad experimental.

Validación cruzada. Remuestreo interno que rota subconjuntos de entrenamiento y validación; debe respetar grupos, tiempo y preprocesamiento.

Validación externa. Evaluación de un modelo congelado en datos independientes de otra población, centro, región, equipo o periodo relevante.

Valor p. Probabilidad, bajo el modelo nulo y sus supuestos, de observar un resultado al menos tan incompatible con el nulo como el obtenido.

Variable proxy. Variable que representa indirectamente otra característica, a veces sensible o causalmente distinta, como centro, costo o código de procedimiento.

Visión computacional. Métodos para extraer información de imágenes o video mediante clasificación, detección, segmentación, seguimiento o cuantificación.

XGBoost. Implementación de gradient boosting con árboles, regularización y optimización eficiente; requiere ajuste y validación rigurosos.

Área bajo la curva ROC (AUC). Probabilidad de que el modelo asigne mayor puntuación a un caso positivo que a uno negativo elegido al azar. Mide ordenamiento, no calibración ni utilidad.

Anexo 2. Guía para seleccionar métodos estadísticos y algoritmos

La selección comienza por la pregunta, el diseño y la unidad independiente. El tipo de archivo disponible o la popularidad de un algoritmo no determina el método. Primero se define qué se desea estimar o predecir; después se elige el procedimiento más sencillo que responda adecuadamente.

Regla central. Ningún algoritmo corrige una mala definición del desenlace, pseudorreplicación, confusión no controlada, fuga de información o ausencia de una población de validación pertinente.

2.1 Secuencia mínima de decisión

Definir la pregunta: descripción, comparación, asociación, efecto causal, diagnóstico, pronóstico, clasificación, agrupamiento o predicción temporal.

Precisar población, unidad experimental, unidad observacional, nivel de agrupamiento y momento cero.

Identificar tipo de desenlace: continuo, binario, ordinal, conteo, tiempo hasta evento, multivariado, imagen, texto o serie temporal.

Reconstruir el diseño: transversal, cohorte, caso-control, ensayo, bloques, factorial, repetido, jerárquico o vigilancia.

Establecer estimando o uso previsto, comparador, horizonte y consecuencia de los errores.

Examinar tamaño muestral, eventos, calidad, faltantes, distribución y dependencia.

Seleccionar un método base interpretable y añadir complejidad solo si existe una ganancia verificable.

Planificar validación, métricas, análisis de sensibilidad, reporte y evaluación de utilidad.

Pregunta o dato	Método estadístico inicial	Comprobaciones esenciales	Cuándo considerar IA/ML
Describir una variable continua	Media/DE o mediana/RIQ según distribución y objetivo	Distribución, unidad y valores extremos	No se requiere ML
Comparar dos grupos independientes	t de Student/Welch; Mann-Whitney para rangos o distribuciones	Independencia, escala, heterocedasticidad	Modelos complejos rara vez aportan
Comparar antes-después en los mismos animales	t pareada; Wilcoxon pareado	Pares completos y diferencia por animal	Modelo mixto si hay más tiempos o desequilibrio
Comparar tres o más tratamientos	ANOVA/ANCOVA; Kruskal-Wallis	Unidad experimental, contrastes, multiplicidad	Árboles no sustituyen estimación de efectos
Desenlace binario	Regresión logística; GEE o mixto si hay dependencia	Eventos, linealidad del logit, agrupamiento	RF/XGBoost/SVM si el objetivo es predicción y hay validación
Conteos	Poisson; binomial negativa; modelos cero-inflados con justificación	Exposición, sobredispersión, ceros	Boosting de conteo si mejora predicción externa
Tiempo hasta evento	Kaplan-Meier, log-rank, Cox, modelos paramétricos	Censura, tiempo cero, riesgos proporcionales	Survival forests/boosting con comparación y calibración temporal
Medidas repetidas o datos por hato/centro	Modelos mixtos, GEE	Estructura de correlación y niveles	Partición y validación agrupadas
Relación no lineal	Splines o GAM	Grado de suavizado, soporte de datos	Boosting si prima predicción y se valida
Composición microbiológica	Distancias, ordenación, PERMANOVA	Profundidad, composición, dispersión, permutación	Clustering o embeddings con estabilidad
Agrupar perfiles sin etiqueta	PCA + clustering jerárquico/k-means	Escala, métrica, estabilidad	UMAP/t-SNE solo como exploración visual
Pronóstico tabular	Regresión penalizada como base	Momento cero, fuga, eventos, calibración	RF, XGBoost, LightGBM, SVM

Imágenes	Lectura experta/medición validada como referencia	Partición por animal/centro y calidad	CNN o vision transformer si tamaño y validación lo permiten
Texto clínico	Reglas o modelos lineales como base	Negación, temporalidad, sujeto, privacidad	Transformers para tareas complejas con evaluación manual
Series de sensores	Modelos longitudinales/estado-espacio/ARIMA	Orden temporal, deriva, fallos de sensor	Boosting, redes recurrentes o transformers temporales si aportan
Efecto de una intervención	Ensayo aleatorizado; métodos causales observacionales	Estimando, confusión, positividad, consistencia	ML puede modelar nuisance functions, no reemplazar identificación causal

2.2 Perfiles de métodos frecuentes

t de Student y Welch

Comparar medias de dos grupos independientes. Welch es preferible cuando las varianzas o tamaños difieren. Informe diferencia de medias, IC y tamaño del efecto; no reduzca la interpretación al valor p.

ANOVA y ANCOVA

ANOVA estima diferencias entre varios grupos; ANCOVA ajusta covariables basales preespecificadas y puede mejorar precisión. Defina contrastes, compruebe residuos y respete la unidad de asignación.

Métodos no paramétricos

Mann-Whitney, Wilcoxon o Kruskal-Wallis trabajan con rangos y son útiles cuando la escala u objetivos lo justifican. No prueban automáticamente igualdad de medianas y no resuelven dependencia.

Modelos lineales generalizados

Conectan el desenlace con predictores mediante una distribución y enlace apropiados. Logística para binarios, Poisson/binomial negativa para conteos y variantes ordinales para categorías ordenadas.

Modelos mixtos y GEE

Representan correlación por animal, camada, corral, hato, clínica o tiempo. Los mixtos modelan heterogeneidad mediante efectos aleatorios; GEE estima efectos promedio poblacionales con correlación de trabajo.

Supervivencia

Kaplan-Meier describe, log-rank compara curvas y Cox estima razones de riesgos bajo proporcionalidad. Defina tiempo cero, evento, censura y horizonte; reporte supervivencia absoluta además de HR.

PCA, NMDS y clustering

Son exploratorios. La escala, distancia, número de componentes o grupos y estabilidad deben justificarse. Un grupo algorítmico necesita validación externa y significado biológico antes de denominarse fenotipo.

PERMANOVA

Contrasta diferencias multivariadas mediante permutaciones. La interpretación requiere revisar dispersión, estructura de dependencia y esquema de permutación compatible con bloques o medidas repetidas.

GAM

Apropiado para curvas no lineales como temperatura-producción o edad-riesgo. La suavidad se penaliza; deben inspeccionarse soporte, incertidumbre, residuos y extrapolación.

Regresión penalizada

Lasso, ridge y elastic net controlan complejidad y ofrecen líneas base sólidas en datos tabulares. El preprocesamiento y la penalización se seleccionan dentro de la validación.

Árboles, random forest y boosting

Capturan interacciones y no linealidades. Deben compararse con modelos simples, calibrarse y explicarse sin confundir importancia con causalidad.

SVM y kNN

SVM puede funcionar en espacios de alta dimensión; kNN predice por proximidad. Ambos dependen de escalamiento, distancia y ajuste interno. Su utilidad disminuye cuando la interpretación o el tiempo de consulta son críticos.

Redes neuronales profundas

Adecuadas para imágenes, audio, secuencias y grandes volúmenes. Requieren partición al nivel del animal, control de atajos, evaluación externa, análisis de fallos y recursos de mantenimiento.

Modelos de lenguaje

Útiles para extracción, clasificación y resumen bajo un esquema verificable. Deben evaluarse negación, temporalidad, alucinaciones, evidencia textual y confidencialidad.

Métodos causales

DAG, estandarización, ponderación, emparejamiento o g-métodos ayudan bajo supuestos explícitos. El objetivo es un contraste de intervenciones, no maximizar AUC.

2.3 Métricas según la decisión

Tarea	Métricas principales	Interpretación responsable
Clasificación equilibrada	Sensibilidad, especificidad, AUC, matriz de confusión	Añadir IC, umbral y análisis por subgrupos.

Evento raro	AUPRC, VPP, sensibilidad, falsas alertas por animal-tiempo	Exactitud puede ser engañosa.
Probabilidad individual	Calibración, Brier, intercepto y pendiente	AUC no basta.
Regresión	MAE, RMSE, sesgo, R^2, gráficos de residuos	Interpretar en unidades clínicas y por rango.
Supervivencia	C-index/AUC temporal, calibración temporal, Brier	Incluir horizontes y censura.
Segmentación	Dice, IoU, error de área/volumen, Hausdorff	Evaluar por animal, lesión y dificultad.
Alerta operativa	Tiempo de anticipación, cobertura, carga de alertas, acción y beneficio neto	Medir efecto sobre flujo y desenlaces.
Clustering	Estabilidad, silueta, replicación, asociación externa	No tratar la silueta como validez biológica.
NLP	F1 por clase, exactitud factual, completitud, evidencia, error de negación	Auditoría humana estratificada.

2.4 Selección de guías de reporte

Tipo de estudio	Guía sugerida	Uso
Experimento in vivo	ARRIVE 2.0	Diseño, animales, aleatorización, enmascaramiento, tamaño, resultados y bienestar.
Estudio observacional veterinario	STROBE-Vet	Población animal, niveles de organización, elegibilidad, sesgos, medición y análisis.
Ensayo en ganado o seguridad alimentaria	REFLECT	Reporte de ensayos aleatorizados con unidades y contextos productivos.

Protocolo de ensayo con IA	SPIRIT-AI + SPIRIT	Intervención algorítmica, entradas, interacción humano-IA, errores y actualización.
Reporte de ensayo con IA	CONSORT-AI + CONSORT	Flujo, intervención, desempeño, errores, integración y resultados clínicos.
Modelo diagnóstico o pronóstico	TRIPOD+AI	Desarrollo y evaluación de modelos de regresión o aprendizaje automático.
Riesgo de sesgo de modelos	PROBAST+AI	Participantes, predictores, desenlace, análisis, calidad y aplicabilidad.
Exactitud diagnóstica con IA	STARD-AI + STARD	Datos, referencia, evaluación del sistema, errores y generalización.
Evaluación clínica temprana de IA	DECIDE-AI	Uso en vivo, interacción humana, factores contextuales, seguridad y flujo.

2.5 Señales para preferir un método simple

La pregunta busca estimar un efecto o diferencia interpretable y el diseño es pequeño o confirmatorio.

El número de eventos es limitado frente a los predictores candidatos.

La relación puede representarse con pocos términos, transformaciones o splines.

Se requiere una probabilidad bien calibrada y explicable en tiempo real.

La mejora de un algoritmo complejo es pequeña, inestable o no se mantiene externamente.

Los costos de despliegue, mantenimiento o auditoría superan el beneficio esperado.

Anexo 3. Banco de casos didácticos veterinarios

Todos los casos de esta sección son casos didácticos simulados. Las cifras propuestas sirven para planificar ejercicios y no representan resultados clínicos reales.

Caso 1. Cicatrización posoperatoria canina: apósito convencional frente a miel

Componente	Desarrollo
Población y unidad	Perros sometidos a cirugía electiva; el perro es la unidad experimental.
Pregunta científica	¿El apósito con miel reduce días hasta epitelización completa y dolor sin aumentar infección?
Diseño propuesto	Ensayo aleatorizado, evaluador enmascarado, seguimiento longitudinal; bloqueo por clínica o tipo de herida.
Variables principales	Tratamiento; días a cicatrización; dolor repetido; infección; rescate; tamaño basal de herida.
Método o algoritmo	Supervivencia o tiempo hasta evento para cicatrización; modelo mixto para dolor; riesgo relativo para infección.
Validación y resultados	Efecto con IC, diferencia clínica, eventos adversos y curvas; evitar usar fotografías como réplicas independientes.
Limitación crítica	Partición por perro para cualquier modelo de imagen; validación externa por clínica.

Actividad sugerida: redactar el objetivo primario, dibujar la estructura de unidades, elaborar el diccionario mínimo, proponer el plan de análisis y anticipar tres decisiones que cambiarían ante resultados distintos.

Caso 2. Predicción de mastitis bovina con sensores

Componente	Desarrollo
Población y unidad	Vacas lecheras anidadas en hatos, con registros diarios.

Pregunta científica	¿Puede anticiparse mastitis clínica 48 horas antes usando producción, conductividad, rumia y temperatura?
Diseño propuesto	Cohorte temporal multicéntrica; origen del pronóstico diario; referencia clínica y microbiológica predefinida.
Variables principales	Ventanas previas; episodio de mastitis; hato; lactancia; días en leche; dispositivo.
Método o algoritmo	Regresión logística penalizada como base; XGBoost como comparación; validación agrupada y temporal.
Validación y resultados	AUPRC, sensibilidad, VPP, calibración, falsas alertas por 100 vaca-días y anticipación.
Limitación crítica	Evitar días de una vaca en ambos conjuntos y variables registradas después de la sospecha.

Actividad sugerida: redactar el objetivo primario, dibujar la estructura de unidades, elaborar el diccionario mínimo, proponer el plan de análisis y anticipar tres decisiones que cambiarían ante resultados distintos.

Caso 3. Dolor posoperatorio en perros

Componente	Desarrollo
Población y unidad	Perros sometidos a OVH con evaluaciones 2, 6, 12 y 24 h.
Pregunta científica	¿Meloxicam más tramadol reduce dolor y necesidad de rescate frente a meloxicam?
Diseño propuesto	Ensayo paralelo; aleatorización; evaluador ciego; protocolo de rescate obligatorio.
Variables principales	CMPS-SF, rescate, sedación, eventos adversos, dolor basal, centro.
Método o algoritmo	Modelo mixto ordinal o continuo según escala; regresión logística para rescate.
Validación y resultados	Trayectorias, contraste a tiempos predefinidos, IC y proporción con rescate.

Limitación crítica	No automatizar analgesia: una alerta visual puede apoyar observación, no reemplazar examen.

Actividad sugerida: redactar el objetivo primario, dibujar la estructura de unidades, elaborar el diccionario mínimo, proponer el plan de análisis y anticipar tres decisiones que cambiarían ante resultados distintos.

Caso 4. Clasificación radiográfica de displasia de cadera

Componente	Desarrollo
Población y unidad	Perros con varias proyecciones por episodio y lecturas de especialistas.
Pregunta científica	¿Una CNN clasifica grados con rendimiento transportable entre hospitales?
Diseño propuesto	Estudio de exactitud multicéntrico; referencia por consenso; centros externos reservados.
Variables principales	Imagen, grado, raza, edad, equipo, proyección, calidad, lector.
Método o algoritmo	Modelo lineal/CNN; aprendizaje por transferencia; calibración y abstención.
Validación y resultados	Sensibilidad por grado, matriz de confusión, AUC ordinal, calibración y cobertura.
Limitación crítica	Separar por perro y centro; auditar marcadores, colimación y atajos de equipo.

Actividad sugerida: redactar el objetivo primario, dibujar la estructura de unidades, elaborar el diccionario mínimo, proponer el plan de análisis y anticipar tres decisiones que cambiarían ante resultados distintos.

Caso 5. Supervivencia en mastocitoma canino

Componente	Desarrollo
Población y unidad	Perros con mastocitoma grado II seguidos desde cirugía.
Pregunta científica	¿Qué factores pronostican recurrencia y supervivencia, y cuál es la asociación del tratamiento adyuvante?
Diseño propuesto	Cohorte; tiempo cero común; criterios explícitos de censura; confusión por indicación.
Variables principales	Tiempo, evento, grado, margen, estadio, tratamiento, edad, centro.
Método o algoritmo	Kaplan-Meier y Cox; modelo penalizado si pocos eventos; análisis causal separado.
Validación y resultados	HR con IC, supervivencia absoluta a 12/24 meses, calibración temporal.
Limitación crítica	No interpretar el coeficiente de tratamiento como efecto causal sin abordar indicación.

Actividad sugerida: redactar el objetivo primario, dibujar la estructura de unidades, elaborar el diccionario mínimo, proponer el plan de análisis y anticipar tres decisiones que cambiarían ante resultados distintos.

Caso 6. Conteo de huevos por gramo en alpacas

Componente	Desarrollo
Población y unidad	Alpacas medidas antes y después de antiparasitario, agrupadas por rebaño.
Pregunta científica	¿Qué tratamiento reduce HPG y cuánto varía la respuesta entre rebaños?
Diseño propuesto	Ensayo por animal o grupo según administración; repetir medición; bloqueo por rebaño.
Variables principales	HPG, tratamiento, tiempo, peso, edad, rebaño.
Método o algoritmo	Binomial negativa mixta; contraste de razones de medias; reducción fecal complementaria.
Validación y resultados	Razón con IC, sobredispersión, ceros, resistencia probable y relevancia sanitaria.
Limitación crítica	No usar ANOVA simple sobre conteos fuertemente asimétricos sin justificar.

Actividad sugerida: redactar el objetivo primario, dibujar la estructura de unidades, elaborar el diccionario mínimo, proponer el plan de análisis y anticipar tres decisiones que cambiarían ante resultados distintos.

Caso 7. Microbiota intestinal y suplementación con orégano

Componente	Desarrollo
Población y unidad	Animales asignados a dieta, con muestras repetidas.
Pregunta científica	¿La suplementación modifica composición microbiana más allá de cambios de dispersión?
Diseño propuesto	Ensayo longitudinal; muestras anidadas; secuenciación y lote técnico documentados.
Variables principales	Matriz de abundancias, dieta, tiempo, animal, lote, covariables.
Método o algoritmo	Distancias, NMDS, PERMANOVA restringida y análisis de dispersión.
Validación y resultados	Tamaño de efecto multivariado, stress, estabilidad y taxones exploratorios con FDR.
Limitación crítica	La significación de PERMANOVA no identifica causalidad de taxones individuales.

Actividad sugerida: redactar el objetivo primario, dibujar la estructura de unidades, elaborar el diccionario mínimo, proponer el plan de análisis y anticipar tres decisiones que cambiarían ante resultados distintos.

Caso 8. Producción de leche frente a estrés térmico

Componente	Desarrollo
Población y unidad	Vacas con producción y temperatura-humedad diaria durante dos años.
Pregunta científica	¿Existe un umbral no lineal a partir del cual disminuye la producción?
Diseño propuesto	Cohorte longitudinal; estacionalidad, lactancia y autocorrelación.
Variables principales	Leche diaria, THI, días en leche, vaca, hato, estación, dieta.
Método o algoritmo	GAM mixto o modelo distribuido de rezagos.
Validación y resultados	Curva con IC, derivada o punto de inflexión con incertidumbre, error temporal.
Limitación crítica	Evitar atribuir causalidad si dieta y manejo cambian con la estación.

Actividad sugerida: redactar el objetivo primario, dibujar la estructura de unidades, elaborar el diccionario mínimo, proponer el plan de análisis y anticipar tres decisiones que cambiarían ante resultados distintos.

Caso 9. Detección automatizada de cojera bovina

Componente	Desarrollo
Población y unidad	Vacas filmadas en pasarela con puntuación locomotora ciega.
Pregunta científica	¿El video identifica cojera moderada o grave con suficiente anticipación?
Diseño propuesto	Exactitud diagnóstica y cohorte prospectiva; varias granjas y cámaras.
Variables principales	Trayectoria, postura, puntuación, lesión, superficie, cámara, granja.
Método o algoritmo	Características biomecánicas + logística como base; CNN/seguimiento como extensión.
Validación y resultados	Sensibilidad, VPP, AUPRC, falsas alertas, acuerdo y cobertura.
Limitación crítica	Evaluar fallos por suciedad, oclusión, raza y superficie.

Actividad sugerida: redactar el objetivo primario, dibujar la estructura de unidades, elaborar el diccionario mínimo, proponer el plan de análisis y anticipar tres decisiones que cambiarían ante resultados distintos.

Caso 10. Riesgo anestésico en pequeños animales

Componente	Desarrollo
Población y unidad	Perros y gatos sometidos a anestesia en hospitales con protocolos distintos.
Pregunta científica	¿Puede estimarse riesgo de hipotensión dentro de los siguientes 15 minutos?
Diseño propuesto	Cohorte temporal; ventanas móviles; momento cero repetido; evaluación prospectiva silenciosa.
Variables principales	Signos vitales previos, fármacos ya administrados, procedimiento, especie, estado ASA.
Método o algoritmo	Logística mixta o boosting temporal; partición por paciente y hospital.
Validación y resultados	Calibración, sensibilidad, alertas por hora, anticipación y beneficio neto.
Limitación crítica	Excluir variables posteriores al episodio y estudiar automatización complaciente.

Actividad sugerida: redactar el objetivo primario, dibujar la estructura de unidades, elaborar el diccionario mínimo, proponer el plan de análisis y anticipar tres decisiones que cambiarían ante resultados distintos.

Caso 11. Citología digital de tumores de células redondas

Componente	Desarrollo
Población y unidad	Láminas de perros y gatos, varios campos por lámina.
Pregunta científica	¿Un modelo clasifica tipos tumorales y detecta campos no evaluables?
Diseño propuesto	Exactitud multicéntrica; diagnóstico histopatológico/consenso como referencia.
Variables principales	Parches, lámina, animal, tinción, escáner, centro, diagnóstico.
Método o algoritmo	CNN con agregación por lámina; comparación con patólogos.
Validación y resultados	F1 macro, sensibilidad por tumor, calibración, cobertura y errores por centro.
Limitación crítica	Parches no son réplicas; el resultado clínico debe agregarse por caso.

Actividad sugerida: redactar el objetivo primario, dibujar la estructura de unidades, elaborar el diccionario mínimo, proponer el plan de análisis y anticipar tres decisiones que cambiarían ante resultados distintos.

Caso 12. Brote zoonótico y riesgo espacial

Componente	Desarrollo
Población y unidad	Predios pecuarios en una región, con notificación, clima y movilidad.
Pregunta científica	¿Qué predios priorizar para inspección la próxima semana?
Diseño propuesto	Vigilancia espacio-temporal; denominadores y esfuerzo de muestreo explícitos.
Variables principales	Casos, animales en riesgo, pruebas, movilidad, precipitación, distancia, semana.
Método o algoritmo	Regresión espacial/GAM como base; boosting espacial como comparación.
Validación y resultados	AUPRC, calibración, sensibilidad bajo cupo fijo y casos detectados por 20 inspecciones.
Limitación crítica	Validación por región y tiempo; comunicar incertidumbre para evitar estigma.

Actividad sugerida: redactar el objetivo primario, dibujar la estructura de unidades, elaborar el diccionario mínimo, proponer el plan de análisis y anticipar tres decisiones que cambiarían ante resultados distintos.

Caso 13. Resistencia antimicrobiana One Health

Componente	Desarrollo
Población y unidad	Aislamientos animales, humanos y ambientales con metadatos y genómica.
Pregunta científica	¿Qué factores se asocian con resistencia y qué clústeres merecen investigación epidemiológica?
Diseño propuesto	Estudio integrado con muestreo definido y gobernanza intersectorial.
Variables principales	Fenotipo, genes, especie, fuente, fecha, lugar, uso antimicrobiano.
Método o algoritmo	GLMM para asociaciones; filogenia/red para proximidad; validación por linaje.
Validación y resultados	Efectos con IC, desempeño, sensibilidad a muestreo y congruencia genotipo-fenotipo.
Limitación crítica	Cercanía genómica no demuestra transmisión directa.

Actividad sugerida: redactar el objetivo primario, dibujar la estructura de unidades, elaborar el diccionario mínimo, proponer el plan de análisis y anticipar tres decisiones que cambiarían ante resultados distintos.

Caso 14. Detección de estrés en cerdos mediante audio

Componente	Desarrollo
Población y unidad	Corrales con micrófonos y observación conductual.
Pregunta científica	¿Los patrones acústicos anticipan episodios de agresión o estrés térmico?
Diseño propuesto	Cohorte por corral; anotación temporal ciega; separación por granja.
Variables principales	Espectrograma, conducta, THI, densidad, hora, granja.
Método o algoritmo	Clasificador de audio; modelo jerárquico para tasa de episodios.
Validación y resultados	Sensibilidad por episodio, falsas alertas/hora, latencia y robustez al ruido.
Limitación crítica	El corral puede ser unidad experimental; proteger conversaciones de trabajadores.

Actividad sugerida: redactar el objetivo primario, dibujar la estructura de unidades, elaborar el diccionario mínimo, proponer el plan de análisis y anticipar tres decisiones que cambiarían ante resultados distintos.

Caso 15. Historias clínicas y hospitalización

Componente	Desarrollo
Población y unidad	Consultas de urgencias en varias clínicas.
Pregunta científica	¿Qué pacientes requerirán hospitalización usando solo información disponible al cierre del triaje?
Diseño propuesto	Cohorte retrospectiva para desarrollo y cohorte temporal externa para evaluación.
Variables principales	Especie, edad, signos, constantes, pruebas disponibles, hospitalización.
Método o algoritmo	Logística penalizada; random forest/XGBoost como comparación.
Validación y resultados	AUC, AUPRC, calibración, decisión por umbral y subgrupos.
Limitación crítica	Hospitalización refleja gravedad y capacidad local; no es un desenlace puramente biológico.

Actividad sugerida: redactar el objetivo primario, dibujar la estructura de unidades, elaborar el diccionario mínimo, proponer el plan de análisis y anticipar tres decisiones que cambiarían ante resultados distintos.

Caso 16. Efecto de rehabilitación tras cirugía ortopédica

Componente	Desarrollo
Población y unidad	Perros asignados a rehabilitación estructurada o cuidado estándar.
Pregunta científica	¿La rehabilitación mejora función a 12 semanas y reduce tiempo hasta apoyo completo?
Diseño propuesto	Ensayo aleatorizado con medidas repetidas y evaluador enmascarado.
Variables principales	Escala funcional, cinética, dolor, adherencia, rescate, eventos adversos.
Método o algoritmo	Modelo mixto; supervivencia para tiempo hasta apoyo.
Validación y resultados	Diferencias con IC, umbral clínico, adherencia y seguridad.
Limitación crítica	Un predictor de recuperación no identifica por sí mismo qué terapia causará mejora.

Actividad sugerida: redactar el objetivo primario, dibujar la estructura de unidades, elaborar el diccionario mínimo, proponer el plan de análisis y anticipar tres decisiones que cambiarían ante resultados distintos.

Caso 17. Agrupamiento de perfiles metabólicos en vacas

Componente	Desarrollo
Población y unidad	Vacas en transición con biomarcadores y desenlaces posteriores.
Pregunta científica	¿Existen perfiles reproducibles asociados con enfermedad metabólica?
Diseño propuesto	Cohorte multicéntrica; clustering solo con variables basales; validación en otro hato.
Variables principales	Metabolitos, condición corporal, paridad, dieta, enfermedad posterior.
Método o algoritmo	PCA + clustering; estabilidad bootstrap; asociación externa ajustada.
Validación y resultados	Estabilidad, separación, prevalencia por centro y riesgo posterior con IC.
Limitación crítica	No denominar fenotipos hasta demostrar replicación y plausibilidad.

Actividad sugerida: redactar el objetivo primario, dibujar la estructura de unidades, elaborar el diccionario mínimo, proponer el plan de análisis y anticipar tres decisiones que cambiarían ante resultados distintos.

Caso 18. Resumen automático de expediente veterinario

Componente	Desarrollo
Población y unidad	Historias longitudinales de animales con notas, laboratorio y medicación.
Pregunta científica	¿Un sistema genera resúmenes fieles para revisión clínica?
Diseño propuesto	Estudio de desempeño con muestra estratificada y evaluación ciega por veterinarios.
Variables principales	Texto, fechas, negaciones, medicamentos, diagnósticos, evidencia fuente.
Método o algoritmo	Reglas/PLN base; transformer o modelo generativo con recuperación de evidencia.
Validación y resultados	Exactitud factual, omisiones críticas, contradicciones, tiempo ahorrado y acuerdo.
Limitación crítica	No ingresar datos identificables en servicios no autorizados; exigir trazabilidad a la nota fuente.

Actividad sugerida: redactar el objetivo primario, dibujar la estructura de unidades, elaborar el diccionario mínimo, proponer el plan de análisis y anticipar tres decisiones que cambiarían ante resultados distintos.

Anexo 4. Plantillas para protocolos de investigación

Las plantillas están diseñadas para ser copiadas y completadas. Los campos entre corchetes deben sustituirse por información específica; las opciones no aplicables deben eliminarse en la versión final.

4.1 Plantilla maestra de protocolo veterinario

Apartado	Contenido a completar
Título	[Diseño + población/especie + exposición/intervención + desenlace + contexto]
Versión y fecha	[vX.X; día/mes/año; historial de enmiendas]
Equipo y responsabilidades	[Investigador principal, clínico, estadístico, responsable de datos, bienestar animal, tecnología]
Antecedentes y justificación	[Estado de la evidencia, incertidumbre, relevancia clínica/biológica y vacío que el estudio aborda]
Pregunta	[PICO/PECO; o pregunta predictiva con población, momento cero, horizonte, desenlace y uso]
Objetivo primario	[Verbo medible, estimando o desempeño principal]
Objetivos secundarios	[Secundarios y exploratorios claramente separados]
Hipótesis	[Científica/causal o predictiva; dirección solo si está justificada]
Diseño	[Ensayo, cohorte, transversal, caso-control, exactitud, modelado, etc.]
Contexto y periodo	[Clínicas, fincas, laboratorios, región y fechas]
Población objetivo y accesible	[Especie, condición, sistema productivo y marco de muestreo]
Elegibilidad	[Criterios de inclusión, exclusión y retiro]
Unidades	[Experimental, observacional, análisis; niveles de agrupamiento y repetición]

Muestreo y tamaño	[Procedimiento; efecto mínimo/precisión/eventos; supuestos; pérdidas; software]
Intervención o exposición	[Dosis, duración, comparador, adherencia, cointervenciones]
Desenlace primario	[Definición operacional, instrumento, tiempo, adjudicación y validez]
Predictores y covariables	[Definición, disponibilidad temporal, medición y justificación]
Aleatorización y ocultamiento	[Secuencia, unidad, estratos/bloques, responsable y mecanismo]
Enmascaramiento	[Quién, cómo, evaluación del éxito y excepciones]
Recogida de datos	[Calendario, fuente, formularios, entrenamiento y control de calidad]
Gestión de datos	[Diccionario, identificadores, acceso, respaldo, retención, trazabilidad]
Plan estadístico	[Estimando, modelo, estructura de dependencia, supuestos, IC, multiplicidad, faltantes y sensibilidad]
Plan de IA/ML	[Algoritmos base y candidatos, partición, preprocesamiento interno, ajuste, métricas, calibración y validación externa]
Ética y bienestar	[Revisión, consentimiento/autorización, 3R, rescate, eventos adversos y detención]
Reproducibilidad	[Registro, código, semillas, versiones, entorno, disponibilidad responsable]
Comunicación	[Guía de reporte, autoría, publicación, devolución de resultados y conflictos]
Cronograma y recursos	[Hitos, personal, equipos, cómputo, mantenimiento y presupuesto]

4.2 Módulo para ensayo clínico o experimental

Estimando primario: [efecto de asignar intervención A frente a B sobre desenlace Y en tiempo T].

Unidad de aleatorización: [animal/corral/hato/centro]. Justificación: [ ].

Secuencia y ocultamiento: [generador, bloques, estratos, sobres/sistema, responsable].

Comparador y cointervenciones: [estándar de cuidado, placebo cuando sea ético, rescate].

Enmascaramiento: [administrador, evaluador, analista] y procedimiento de desenmascaramiento.

Adherencia, desviaciones y análisis: [intención de tratar/per-protocolo y sensibilidad].

Regla de seguridad: [evento, umbral, responsable y acción].

Guía: [ARRIVE 2.0/REFLECT/PetSORT/CONSORT, según especie y contexto].

4.3 Módulo para estudio observacional

Exposición: [definición, ventana y método de medición].

Desenlace: [definición, cegamiento y tiempo].

Diagrama causal: [confusores mínimos, mediadores, colisionadores y variables proxy].

Estrategia de control: [restricción, ajuste, ponderación, emparejamiento o estandarización].

Sesgos anticipados: [selección, información, confusión, pérdida y medición].

Análisis de sensibilidad: [confusión residual, definiciones alternativas, pérdidas].

Guía: STROBE-Vet.

4.4 Módulo para exactitud diagnóstica

Condición objetivo: [definición clínica/biológica].

Prueba índice: [procedimiento, umbral, versión y operador].

Estándar de referencia: [procedimiento, independencia, adjudicación y limitaciones].

Espectro de casos y muestreo: [consecutivo/aleatorio; prevalencia esperada].

Cegamiento entre prueba y referencia: [ ].

Resultados indeterminados/no evaluables: [manejo y reporte].

Métricas: sensibilidad, especificidad, valores predictivos, razones de verosimilitud, IC y calibración si produce probabilidades.

Guía: STARD o STARD-AI cuando el sistema diagnóstico incorpora IA.

4.5 Módulo para desarrollo o validación de un modelo predictivo

Uso previsto: [usuario, decisión, contexto y acción].

Población, momento cero y horizonte: [ ].

Desenlace: [medición ciega y disponibilidad].

Predictores admisibles: [solo los disponibles en el momento de uso].

Tamaño muestral: [eventos, complejidad, sobreajuste esperado y precisión].

Partición: [por animal, hato, centro y/o tiempo].

Flujo: [imputación, codificación, escalamiento y selección dentro de remuestreo].

Comparadores: [práctica habitual y modelo simple].

Evaluación: discriminación, calibración, error, utilidad, subgrupos y validación externa.

Guía: TRIPOD+AI; riesgo de sesgo con PROBAST+AI.

4.6 Módulo para imágenes y visión computacional

Unidad de partición: [animal/episodio/lámina/centro], nunca imagen aislada si existe dependencia.

Adquisición: [equipo, protocolo, resolución, proyección, compresión y control de calidad].

Anotación: [evaluadores, entrenamiento, independencia, concordancia y adjudicación].

Aumento: [transformaciones anatómica y técnicamente plausibles].

Atajos: [marcadores, fondo, equipo, colimación, texto, centro].

Evaluación: [por caso y subgrupo; calibración, cobertura, mapas solo como auditoría].

Validación externa: [centro/equipo/población no usados].

4.7 Módulo para sensores y series temporales

Frecuencia y sincronización: [sensor, reloj, zona horaria y tolerancia].

Definición de ventana: [historia disponible, origen y horizonte].

Fallos: [batería, pérdidas, desconexión, valores imposibles y mantenimiento].

Referencia: [observación, examen, laboratorio o sistema calibrado].

Partición cronológica y agrupada: [ ].

Métricas operativas: [anticipación, falsas alertas por animal-tiempo, cobertura y tiempo de respuesta].

Plan de deriva: [distribución, equipo, estación, manejo y recalibración].

4.8 Módulo para epidemiología, vigilancia y One Health

Evento y definición de caso: [señal, probable, confirmado].

Denominador y esfuerzo diagnóstico: [población, pruebas, cobertura y retraso].

Unidad espacial y temporal: [resolución y justificación].

Dependencia: [vecindad, redes, linajes, hato y repetición].

Validación: [bloques espaciales, futuro cronológico y separación filogenética].

Capacidad operativa: [número de inspecciones y costo de alarmas].

Gobernanza: [acuerdos animal-humano-ambiental, propósito y comunicación de riesgo].

4.9 Módulo para uso secundario de datos e IA generativa

Finalidad original y nueva finalidad: [compatibilidad y autorización].

Base legal/ética y consentimiento o dispensa: [ ].

Minimización, seudonimización y control de acceso: [ ].

Herramienta generativa: [nombre, versión, fecha, configuración y condiciones de datos].

Prompts y salidas conservadas: [ubicación y acceso].

Verificación: [fuente primaria, revisión estadística, pruebas de código y responsable].

Declaración pública: [tareas asistidas y responsabilidad humana].

4.10 Plantilla de plan de análisis estadístico y predictivo

Componente	Especificación
1. Congelamiento	Versión del protocolo, fecha de cierre y responsables.
2. Poblaciones analíticas	Incluidos, excluidos, intención de tratar, per-protocolo y seguridad.
3. Variables	Definiciones, codificaciones, unidades, derivaciones y ventanas.
4. Descriptivo	Diagrama de flujo, denominadores, distribución, pérdidas y calidad.
5. Estimando/modelo	Contraste primario, enlace, efectos fijos/aleatorios y covariables.
6. Supuestos	Procedimientos diagnósticos y acciones ante incumplimiento.
7. Faltantes	Descripción, supuesto principal, imputación y sensibilidad.
8. Multiplicidad	Desenlace primario, contrastes, jerarquía y ajuste.
9. Predicción	Partición, pipeline, ajuste, modelo base, métricas y calibración.
10. Sensibilidad	Definiciones, valores extremos, centros, mecanismos de ausencia y confusión.

11. Software	Programa, paquetes, versiones, semillas y entorno.
12. Desviaciones	Registro fechado de cambios y clasificación confirmatoria/exploratoria.

Anexo 5. Prompts científicos aplicados a Ciencias Veterinarias

Un prompt científico útil delimita tarea, contexto, datos disponibles, restricciones, forma de salida y procedimiento de verificación. No debe pedir al modelo que invente información faltante ni tratar su respuesta como fuente bibliográfica.

Control obligatorio. No introduzca historias clínicas identificables, datos de tutores, productores o trabajadores en servicios no autorizados. Toda referencia, cálculo, código o afirmación debe comprobarse de manera independiente.

5.1 Plantilla general

Prompt base. Actúa como [rol metodológico]. Objetivo: [pregunta concreta]. Contexto veterinario: [especie, población, lugar, diseño y decisión]. Datos disponibles: [variables, unidades, estructura, fechas, faltantes y tamaño]. Restricciones: no inventes datos ni referencias; distingue hecho, supuesto e incertidumbre; respeta la unidad experimental y evita fuga. Tarea: [acciones]. Salida: [tabla/esquema/código/informe] con [campos]. Verificación: enumera supuestos, pruebas de consistencia, limitaciones y elementos que requieren revisión humana o consulta de la fuente primaria.

Prompt 1. Formulación de pregunta PECO

A partir del tema [tema], formula tres preguntas PECO para [especie/población]. Para cada una especifica población, exposición, comparador, desenlace, periodo, unidad de análisis, diseño viable y principal amenaza de sesgo. No propongas causalidad si el diseño solo permite asociación.

Comprobación mínima: verifique datos de entrada, supuestos, referencias, resultados numéricos y correspondencia entre la salida y la pregunta original.

Prompt 2. Separar objetivo inferencial y predictivo

Para el problema [descripción], redacta un objetivo inferencial y otro predictivo. Define para cada uno estimando o uso previsto, momento cero, horizonte, variables admisibles, diseño, método de evaluación y conclusión que no sería válida.

Comprobación mínima: verifique datos de entrada, supuestos, referencias, resultados numéricos y correspondencia entre la salida y la pregunta original.

Prompt 3. Mapa de evidencia

Construye una estrategia de búsqueda para [pregunta] en PubMed/CAB Abstracts/Scopus. Entrega bloques de términos por población, condición, método y especie; incluye sinónimos y vocabulario controlado. No inventes resultados ni referencias. Señala qué términos deben probarse y documentarse.

Comprobación mínima: verifique datos de entrada, supuestos, referencias, resultados numéricos y correspondencia entre la salida y la pregunta original.

Prompt 4. Cribado bibliográfico

Aplica estos criterios [criterios] a los títulos/resúmenes proporcionados. Devuelve decisión provisional, criterio exacto y fragmento de evidencia. Marca “información insuficiente” cuando corresponda; no completes datos ausentes.

Comprobación mínima: verifique datos de entrada, supuestos, referencias, resultados numéricos y correspondencia entre la salida y la pregunta original.

Prompt 5. Verificación de referencia

Comprueba si la referencia [metadatos] existe y si respalda la afirmación [afirmación]. Informa coincidencia de autores, título, año, revista, DOI y alcance. Si no puedes acceder a la fuente primaria, declara que requiere verificación y no reconstruyas datos.

Comprobación mínima: verifique datos de entrada, supuestos, referencias, resultados numéricos y correspondencia entre la salida y la pregunta original.

Prompt 6. Auditoría de protocolo

Revisa el protocolo adjunto para coherencia entre pregunta, unidad experimental, asignación, desenlace, tamaño muestral y análisis. Clasifica cada observación como crítica, mayor o menor y propone una corrección concreta sin cambiar la pregunta principal.

Comprobación mínima: verifique datos de entrada, supuestos, referencias, resultados numéricos y correspondencia entre la salida y la pregunta original.

Prompt 7. Identificación de pseudorreplicación

Analiza la estructura [descripción]. Distingue unidad experimental, observacional, técnica y de análisis; dibuja los niveles de anidamiento en texto estructurado y explica qué errores estándar serían incorrectos si se trataran todas las filas como independientes.

Comprobación mínima: verifique datos de entrada, supuestos, referencias, resultados numéricos y correspondencia entre la salida y la pregunta original.

Prompt 8. Plan de aleatorización

Diseña una estrategia de aleatorización para [n unidades] en [tratamientos], con bloqueo por [factor]. Explica generación, ocultamiento, conservación del registro y análisis compatible. Entrega pseudocódigo y una lista de comprobación; no generes asignaciones definitivas si faltan identificadores.

Comprobación mínima: verifique datos de entrada, supuestos, referencias, resultados numéricos y correspondencia entre la salida y la pregunta original.

Prompt 9. Tamaño muestral inferencial

Enumera la información necesaria para calcular tamaño muestral para [comparación/modelo]. Separa efecto clínicamente importante, variabilidad, alfa, potencia, correlación intraclúster/repetida, asignación y pérdidas. No inventes valores; muestra la fórmula o procedimiento una vez suministrados.

Comprobación mínima: verifique datos de entrada, supuestos, referencias, resultados numéricos y correspondencia entre la salida y la pregunta original.

Prompt 10. Tamaño muestral predictivo

Evalúa si los datos propuestos permiten desarrollar un modelo para [desenlace]. Considera eventos, prevalencia, parámetros candidatos, sobreajuste, calibración y validación externa. Rechaza reglas universales simplistas y explica qué información falta.

Comprobación mínima: verifique datos de entrada, supuestos, referencias, resultados numéricos y correspondencia entre la salida y la pregunta original.

Prompt 11. Diccionario de datos

Convierte la lista de variables en un diccionario con nombre, etiqueta, definición operacional, tipo, unidad, momento, fuente, rango válido, código de ausencia, nivel de medición y riesgo de fuga. Señala ambigüedades para resolver con el equipo clínico.

Comprobación mínima: verifique datos de entrada, supuestos, referencias, resultados numéricos y correspondencia entre la salida y la pregunta original.

Prompt 12. Reglas de calidad

Para el diccionario adjunto, propone reglas de validez, consistencia longitudinal, duplicados y reconciliación. Separa valores imposibles de extremos clínicamente plausibles y define qué requiere revisión humana.

Comprobación mínima: verifique datos de entrada, supuestos, referencias, resultados numéricos y correspondencia entre la salida y la pregunta original.

Prompt 13. Datos faltantes

Describe el patrón de ausencia por variable, tiempo, clínica y desenlace. Formula hipótesis MCAR/MAR/MNAR, propone gráficos y modelos de ausencia, y diseña análisis principal y de sensibilidad. No afirme que los datos son MAR solo por una prueba estadística.

Comprobación mínima: verifique datos de entrada, supuestos, referencias, resultados numéricos y correspondencia entre la salida y la pregunta original.

Prompt 14. Selección de método estadístico

Con esta pregunta [ ], diseño [ ], unidades [ ], desenlace [ ] y covariables [ ], compara hasta tres métodos. Para cada uno indica estimando, supuestos, estructura de dependencia, interpretación, limitaciones y diagnóstico. Recomienda el método más simple adecuado.

Comprobación mínima: verifique datos de entrada, supuestos, referencias, resultados numéricos y correspondencia entre la salida y la pregunta original.

Prompt 15. Modelo mixto

Propón un modelo mixto para [desenlace] medido en [tiempos] en animales agrupados por [nivel]. Especifica fórmula conceptual, efectos fijos, aleatorios, interacción, estructura temporal, contrastes, supuestos e interpretación clínica.

Comprobación mínima: verifique datos de entrada, supuestos, referencias, resultados numéricos y correspondencia entre la salida y la pregunta original.

Prompt 16. Supervivencia

Diseña el análisis de tiempo hasta [evento]. Define tiempo cero, evento, censura, riesgos competitivos, curvas, modelo, verificación de proporcionalidad, estimaciones absolutas e interpretación de HR sin equipararlo a riesgo relativo.

Comprobación mínima: verifique datos de entrada, supuestos, referencias, resultados numéricos y correspondencia entre la salida y la pregunta original.

Prompt 17. Microbiota/PERMANOVA

Diseña un análisis de microbiota para [intervención]. Incluye preprocesamiento, distancia, ordenación, esquema de permutación, dispersión, covariables, medidas repetidas, multiplicidad y límites de interpretación composicional.

Comprobación mínima: verifique datos de entrada, supuestos, referencias, resultados numéricos y correspondencia entre la salida y la pregunta original.

Prompt 18. Serie temporal

Para predecir [variable] con datos horarios/diarios, define origen, horizonte, ventanas, rezagos, estacionalidad, partición cronológica, línea base, métricas y estrategia frente a fallos de sensor. Identifica fugas temporales posibles.

Comprobación mínima: verifique datos de entrada, supuestos, referencias, resultados numéricos y correspondencia entre la salida y la pregunta original.

Prompt 19. Pipeline de machine learning

Construye un pipeline para [desenlace] con [estructura]. Incluye separación agrupada/temporal, imputación, codificación, escalamiento, selección, ajuste anidado, modelo base, candidatos, calibración y prueba externa. Todo preprocesamiento debe aprenderse dentro del entrenamiento.

Comprobación mínima: verifique datos de entrada, supuestos, referencias, resultados numéricos y correspondencia entre la salida y la pregunta original.

Prompt 20. Comparación justa de modelos

Diseña una comparación entre regresión logística, random forest y XGBoost usando las mismas particiones, variables y métricas. Incluye incertidumbre pareada, calibración, tiempo/costo y regla predefinida para preferir el modelo simple.

Comprobación mínima: verifique datos de entrada, supuestos, referencias, resultados numéricos y correspondencia entre la salida y la pregunta original.

Prompt 21. Radiología/visión computacional

Audita un estudio de imágenes veterinarias. Comprueba partición por animal y centro, referencia, anotadores, aumento, atajos, calidad, espectro, calibración, abstención y validación externa. Devuelve una tabla de riesgo, evidencia y acción correctiva.

Comprobación mínima: verifique datos de entrada, supuestos, referencias, resultados numéricos y correspondencia entre la salida y la pregunta original.

Prompt 22. Evaluación de un modelo de dolor

Propón un protocolo para validar una medición automática de dolor en [especie]. Incluye escala de referencia, sedación, conformación, observadores, tiempos, acuerdo, sensibilidad al cambio, umbral, no evaluables y consecuencias de error.

Comprobación mínima: verifique datos de entrada, supuestos, referencias, resultados numéricos y correspondencia entre la salida y la pregunta original.

Prompt 23. Epidemiología espacial

Para un modelo de riesgo [enfermedad], distingue conteo, incidencia y riesgo. Define denominador, esfuerzo diagnóstico, resolución espacial/temporal, autocorrelación, validación por bloques y comunicación de incertidumbre.

Comprobación mínima: verifique datos de entrada, supuestos, referencias, resultados numéricos y correspondencia entre la salida y la pregunta original.

Prompt 24. Código R reproducible

Escribe código R para [análisis] usando nombres de variables suministrados. Antes del código declara supuestos y estructura esperada. Incluye comprobaciones, semillas, manejo explícito de faltantes, diagnósticos y salida ordenada. No inventes columnas; marca cualquier dato requerido.

Comprobación mínima: verifique datos de entrada, supuestos, referencias, resultados numéricos y correspondencia entre la salida y la pregunta original.

Prompt 25. Depuración de código R

Revisa el código y el mensaje de error adjuntos. Identifica causa probable, reproduce el problema con un ejemplo mínimo, propone corrección y añade pruebas que detecten el fallo. No cambies silenciosamente la pregunta estadística.

Comprobación mínima: verifique datos de entrada, supuestos, referencias, resultados numéricos y correspondencia entre la salida y la pregunta original.

Prompt 26. Interpretación de resultados

Interpreta la tabla proporcionada separando magnitud, IC, valor p, relevancia clínica, importancia biológica y capacidad predictiva. Evita lenguaje causal si el diseño no lo permite y enumera explicaciones alternativas.

Comprobación mínima: verifique datos de entrada, supuestos, referencias, resultados numéricos y correspondencia entre la salida y la pregunta original.

Prompt 27. Auditoría de calibración

A partir de predicciones y desenlaces, especifica cómo evaluar gráfica, intercepto, pendiente y Brier; cómo recalibrar sin reentrenar; y cómo verificar desempeño por centro y subgrupo. No selecciones umbral usando el conjunto de prueba.

Comprobación mínima: verifique datos de entrada, supuestos, referencias, resultados numéricos y correspondencia entre la salida y la pregunta original.

Prompt 28. Revisión crítica de artículo

Evalúa el manuscrito con [TRIPOD+AI/STARD-AI/STROBE-Vet/ARRIVE]. Para cada ítem registra página, evidencia, juicio y recomendación. Distingue falta de reporte de defecto real y no supongas procedimientos no descritos.

Comprobación mínima: verifique datos de entrada, supuestos, referencias, resultados numéricos y correspondencia entre la salida y la pregunta original.

Prompt 29. Resumen científico con trazabilidad

Resume exclusivamente el texto proporcionado en: pregunta, diseño, población, unidad, método, resultados con cifras, limitaciones y aplicabilidad. Vincula cada afirmación a la sección o tabla fuente y marca contradicciones.

Comprobación mínima: verifique datos de entrada, supuestos, referencias, resultados numéricos y correspondencia entre la salida y la pregunta original.

Prompt 30. Declaración de uso de IA

Redacta una declaración transparente del uso de [herramienta, versión y fecha] para [tareas]. Incluye datos introducidos, controles de privacidad, verificación humana, correcciones, conservación de prompts y responsabilidad autoral. No atribuyas autoría a la herramienta.

Comprobación mínima: verifique datos de entrada, supuestos, referencias, resultados numéricos y correspondencia entre la salida y la pregunta original.

Anexo 6. Respuestas orientativas a las preguntas de autoevaluación

Las respuestas sintetizan los criterios esperados. Admiten formulaciones alternativas si conservan rigor metodológico, coherencia veterinaria y una interpretación limitada por el diseño.

Capítulo 1. Fundamentos de inteligencia artificial en Ciencias Veterinarias

1. Defina inteligencia artificial sin atribuir al sistema conciencia o intención.

Respuesta orientativa: La inteligencia artificial es el conjunto de métodos computacionales que ejecuta tareas como clasificación, predicción, generación o apoyo a decisiones a partir de reglas o patrones aprendidos; no presupone conciencia, intención ni comprensión humana.

2. Explique por qué un proyecto de ciencia de datos puede no contener machine learning.

Respuesta orientativa: La ciencia de datos también comprende formulación, limpieza, descripción, visualización, inferencia y comunicación. Un proyecto puede resolver su pregunta mediante estadística descriptiva o modelos clásicos sin entrenar un algoritmo predictivo.

3. Compare machine learning convencional y deep learning mediante un problema veterinario.

Respuesta orientativa: En datos tabulares de mastitis, una regresión o random forest usa variables definidas; una CNN puede aprender representaciones directamente de imágenes ecográficas. El segundo exige más datos, cómputo y control de atajos.

4. ¿Qué diferencia existe entre una herramienta generativa y una fuente científica?

Respuesta orientativa: Una herramienta generativa produce texto o código probabilístico; una fuente científica es evidencia identificable y evaluable. La salida generativa orienta tareas, pero debe verificarse contra fuentes primarias.

5. Formule un ejemplo veterinario de aprendizaje supervisado y otro no supervisado.

Respuesta orientativa: Supervisado: predecir mastitis con casos etiquetados. No supervisado: agrupar perfiles metabólicos de vacas sin clases previas y luego estudiar su estabilidad y significado.

6. ¿Por qué el aprendizaje por refuerzo exige especial cautela cuando las acciones afectan animales?

Respuesta orientativa: Porque aprende mediante consecuencias de acciones y podría explorar conductas dañinas. Antes de interactuar con animales se requieren simulación, restricciones, supervisión, rescate y aprobación ética.

7. Diferencie una pregunta inferencial de una predictiva utilizando un ensayo posquirúrgico.

Respuesta orientativa: Inferencial: estimar la diferencia de dolor entre dos analgésicos con IC. Predictiva: anticipar qué perro necesitará rescate. La primera se centra en un efecto; la segunda en rendimiento fuera de muestra.

8. Explique por qué una variable predictiva no debe interpretarse automáticamente como causa.

Respuesta orientativa: Puede ser un proxy, reflejar confusión o el proceso de atención. Predecir no identifica qué ocurriría al intervenir sobre la variable.

9. Identifique la unidad experimental y la unidad observacional en un estudio con mediciones diarias de vacas agrupadas en hatos.

Respuesta orientativa: La unidad experimental depende de dónde se asigna la exposición; puede ser la vaca o el hato. Las mediciones diarias son unidades observacionales repetidas y no réplicas independientes.

10. ¿Qué partición usaría para evaluar un modelo destinado a una clínica nunca observada?

Respuesta orientativa: Una validación externa por clínica: reservar una o más clínicas completas nunca usadas en entrenamiento ni ajuste.

11. Explique cómo la prevalencia modifica el valor predictivo positivo.

Respuesta orientativa: Con sensibilidad y especificidad constantes, el VPP aumenta al crecer la prevalencia y disminuye cuando la condición es rara, porque cambia la proporción de falsos frente a verdaderos positivos.

12. ¿Qué información aporta la calibración que no aporta el AUC?

Respuesta orientativa: La calibración indica si la magnitud de las probabilidades corresponde a frecuencias observadas; el AUC solo evalúa ordenamiento entre casos y no casos.

13. Proponga dos formas de fuga de información en un estudio radiográfico canino.

Respuesta orientativa: Dividir imágenes del mismo perro entre entrenamiento y prueba; o normalizar/seleccionar imágenes usando información de toda la base, incluida la prueba.

14. ¿Qué riesgos plantea imputar resultados de laboratorio no solicitados?

Respuesta orientativa: La ausencia puede depender de gravedad, criterio del clínico y acceso. Imputar como si fuera aleatoria puede borrar esa información, introducir sesgo y producir predicciones clínicamente incoherentes.

15. ¿Por qué una exactitud de 95 % puede ser inútil en una enfermedad con prevalencia de 5 %?

Respuesta orientativa: Si la prevalencia es 5 %, predecir siempre “sin enfermedad” alcanza 95 % de exactitud y sensibilidad cero; no detecta ningún caso.

16. Distinga explicación algorítmica y explicación causal.

Respuesta orientativa: La explicación algorítmica atribuye una predicción dentro del modelo; la causal explica efectos de intervenciones bajo diseño y supuestos explícitos.

17. ¿Qué variables consideraría para monitorear deriva en un sistema de ordeño?

Respuesta orientativa: Distribuciones de sensores, producción, conductividad, faltantes, tasa de alertas, prevalencia, calibración, desempeño por hato, equipo, estación y versión.

18. Describa una consecuencia ética de un falso negativo y otra de un falso positivo.

Respuesta orientativa: Un falso negativo puede retrasar diagnóstico y tratamiento; un falso positivo puede inducir pruebas, aislamiento, tratamiento o costos innecesarios.

19. ¿Qué evidencia adicional se necesita después de demostrar buen desempeño predictivo?

Respuesta orientativa: Validación externa y prospectiva, calibración, comparación con práctica habitual, estudio de impacto, análisis de daños y monitoreo después del despliegue.

20. Justifique por qué ningún algoritmo corrige un experimento mal diseñado.

Respuesta orientativa: El algoritmo aprende de la información generada por el diseño. No recupera controles ausentes, independencia inexistente, variables no medidas ni asignación confundida.

Capítulo 2. La inteligencia artificial dentro del método científico

1. ¿Cuál es la diferencia entre un tema y un problema de investigación veterinaria?

Respuesta orientativa: El tema delimita un campo; el problema articula una incertidumbre específica, relevante y comprobable dentro de una población y contexto.

2. ¿Por qué una base de datos disponible no basta para justificar una pregunta científica?

Respuesta orientativa: Porque los datos pueden no representar a la población, no medir el constructo o no sostener el objetivo causal o predictivo.

3. Formule una pregunta PECO sobre parasitismo gastrointestinal en ovinos.

Respuesta orientativa: Debe incluir población ovina, exposición, comparador y desenlace medible, además del periodo y contexto cuando sean relevantes.

4. ¿Qué diferencia existe entre una hipótesis científica y una hipótesis predictiva?

Respuesta orientativa: La científica propone una relación sustantiva; la predictiva anticipa desempeño en casos nuevos y requiere validación fuera del ajuste.

5. ¿Por qué la importancia de una variable en random forest no demuestra causalidad?

Respuesta orientativa: Porque la importancia mide contribución al desempeño condicionada por los datos y el algoritmo, no un efecto de intervención libre de confusión.

6. ¿Qué riesgo se produce al dividir imágenes del mismo animal entre entrenamiento y prueba?

Respuesta orientativa: Fuga de información y estimación optimista, pues el modelo reconoce patrones específicos del paciente ya visto.

7. Mencione cuatro elementos que deben registrarse al usar IA para cribado bibliográfico.

Respuesta orientativa: Herramienta/versión, criterios, decisiones de entrenamiento, regla de detención y auditoría de registros excluidos.

8. ¿Cómo se distingue un vacío poblacional de un vacío de validación?

Respuesta orientativa: El primero afecta la representación de grupos; el segundo aparece cuando el desempeño no se ha probado en datos independientes.

9. ¿Qué debe contener un plan de análisis además del nombre de una prueba?

Respuesta orientativa: Estimando, covariables, supuestos, datos faltantes, multiplicidad, sensibilidad, intervalos y visualización.

10. ¿Por qué la ausencia de eventos adversos significativos no demuestra equivalencia de seguridad?

Respuesta orientativa: Porque una muestra pequeña puede carecer de potencia para detectar daños y “no significativo” no equivale a ausencia ni equivalencia.

11. En el caso didáctico, ¿cuál es la unidad experimental y cuáles son las unidades observacionales?

Respuesta orientativa: La perra es la unidad experimental; las puntuaciones repetidas son observaciones correlacionadas dentro del animal.

12. ¿En qué circunstancias sería metodológicamente preferible no entrenar un modelo de aprendizaje automático?

Respuesta orientativa: Cuando la muestra es pequeña, el objetivo es una estimación interpretable, faltan eventos o un modelo simple responde mejor a la pregunta.

Capítulo 3. Ética, integridad científica y gobernanza de datos veterinarios

1. ¿Por qué un estudio con pseudorreplicación constituye también un problema ético?

Respuesta orientativa: Porque trata observaciones dependientes como réplicas, subestima la incertidumbre y puede producir conclusiones inválidas tras usar animales o datos.

2. Diferencie reemplazo, reducción y refinamiento con un ejemplo veterinario para cada principio.

Respuesta orientativa: Reemplazo evita animales vivos; reducción optimiza el número compatible con validez; refinamiento disminuye daño y mejora bienestar.

3. ¿Por qué "retrospectivo" no significa automáticamente "exento de consentimiento"?

Respuesta orientativa: Porque persisten riesgos, finalidades no previstas, datos personales y requisitos de revisión; una dispensa debe estar fundamentada.

4. Explique la diferencia entre anonimización y seudonimización.

Respuesta orientativa: La anonimización busca impedir razonablemente el vínculo; la seudonimización conserva la posibilidad de enlace mediante clave o información adicional.

5. ¿Cómo puede una historia clínica animal contener datos personales protegidos?

Respuesta orientativa: Incluye identificadores y circunstancias del tutor, trabajadores, finca, pagos, imágenes o geolocalización.

6. ¿Por qué FAIR no equivale a acceso público irrestricto?

Respuesta orientativa: La accesibilidad puede ser controlada; FAIR exige condiciones claras y metadatos, no divulgación de datos sensibles.

7. Describa dos formas de fuga de datos en un modelo basado en expedientes veterinarios.

Respuesta orientativa: Visitas del mismo animal en entrenamiento y prueba; variables recogidas después del desenlace disponibles durante el entrenamiento.

8. ¿Qué métricas revisaría por subgrupo para evaluar un clasificador de urgencias?

Respuesta orientativa: Sensibilidad, especificidad, valores predictivos, tasa de falsos negativos y calibración, con intervalos y denominadores.

9. ¿Por qué no siempre es posible satisfacer simultáneamente todas las definiciones de equidad?

Respuesta orientativa: Con prevalencias diferentes, igualar una métrica puede empeorar otra; debe priorizarse según el daño clínico.

10. Enumere cinco elementos mínimos de transparencia antes de adquirir un sistema de IA veterinaria.

Respuesta orientativa: Uso previsto, población, datos, definición de desenlace, validación, métricas, subgrupos, limitaciones, versión y monitoreo.

11. ¿Qué convierte una supervisión humana en meramente simbólica?

Respuesta orientativa: La falta de competencia, tiempo, información o facultad para contradecir la salida.

12. Aplique los cuatro criterios de autoría del ICMJE a un colaborador que solo aportó datos.

Respuesta orientativa: No basta aportar datos; debe participar intelectualmente, aprobar la versión y asumir responsabilidad, o figurar como contribuidor.

13. ¿Dónde declararía una IA usada para extraer variables clínicas y qué información incluiría?

Respuesta orientativa: En métodos, con herramienta y versión, finalidad, entradas, parámetros, validación y revisión humana.

14. ¿Qué riesgos introduce un chatbot cuando se le entrega texto libre de historias clínicas?

Respuesta orientativa: Divulgación o retención de identificadores, transferencia no autorizada, reidentificación y uso secundario por el proveedor.

15. Proponga un criterio de suspensión para un modelo de predicción de hospitalización.

Respuesta orientativa: Por ejemplo, suspender si la sensibilidad cae bajo el umbral predefinido o aparece daño grave atribuible.

Capítulo 4. Diseño de investigaciones veterinarias asistidas por inteligencia artificial

1. En un ensayo de cuatro dietas asignadas a ocho corrales, con 25 cerdos por corral, ¿cuál es la unidad experimental y cuál es el tamaño de replicación por dieta?

Respuesta orientativa: El corral es la unidad experimental; hay dos corrales por dieta. Los 25 cerdos son subunidades observacionales dentro de cada corral.

2. Explique por qué alternar pacientes entre tratamientos A y B no equivale a aleatorizar.

Respuesta orientativa: La alternancia es previsible y permite manipulación consciente o inconsciente; no produce una secuencia probabilística con ocultamiento.

3. ¿Qué diferencia existe entre ocultamiento de la asignación y enmascaramiento del evaluador?

Respuesta orientativa: El ocultamiento evita conocer la próxima asignación antes de incluir al animal; el enmascaramiento evita conocer el tratamiento durante aplicación, evaluación o análisis.

4. Un laboratorio realiza tres PCR por muestra de 12 gatos. ¿Cuántas réplicas biológicas y técnicas existen?

Respuesta orientativa: Doce réplicas biológicas; tres lecturas técnicas por muestra, 36 lecturas en total.

5. ¿Por qué un único hato tratado y un único hato control no permiten estimar separadamente el efecto del tratamiento?

Respuesta orientativa: Tratamiento y hato están completamente confundidos; cualquier diferencia puede atribuirse a manejo, genética, ambiente o intervención.

6. Proponga un factor de bloqueo razonable para un ensayo multicéntrico de analgesia posoperatoria canina y justifique su elección.

Respuesta orientativa: El centro clínico, porque resume diferencias de equipo, operadores y flujo; la aleatorización debe realizarse dentro de cada centro.

7. En un factorial 2 × 2, ¿qué pregunta responde la interacción?

Respuesta orientativa: Si el efecto de un factor cambia según el nivel del otro; por ejemplo, si un apósito funciona de forma distinta bajo dos protocolos analgésicos.

8. ¿Cuándo un conjunto de mediciones en diferentes días corresponde a medidas repetidas y cuándo a grupos independientes?

Respuesta orientativa: Son repetidas si pertenecen al mismo animal; son independientes si el procedimiento terminal obliga a usar animales distintos en cada tiempo.

9. Mencione cuatro componentes necesarios para justificar el tamaño muestral de una comparación confirmatoria de dos medias.

Respuesta orientativa: Efecto mínimo relevante, variabilidad, α, potencia, asignación y pérdidas; se aceptan cuatro correctamente explicados.

10. ¿Por qué la regla de diez eventos por predictor no debe utilizarse como criterio universal para un modelo predictivo?

Respuesta orientativa: Porque ignora prevalencia, complejidad total, señal, sobreajuste, calibración y precisión de predicciones; puede ser insuficiente o innecesariamente restrictiva.

11. Dé un ejemplo veterinario de fuga temporal y otro de fuga por identidad.

Respuesta orientativa: Temporal: usar infección diagnosticada al día 7 para predecir al día 3. Identidad: imágenes del mismo perro en entrenamiento y prueba.

12. Diferencie confusor, mediador y colisionador en términos de la decisión de ajuste.

Respuesta orientativa: El confusor puede requerir control; el mediador está en la vía causal y ajustarlo cambia el estimando; el colisionador no debe condicionarse sin justificación.

13. ¿Qué evidencia solicitaría antes de aceptar un algoritmo de segmentación de heridas como instrumento de medición?

Respuesta orientativa: Anotación de referencia, acuerdo entre evaluadores, validación por animal y externa, error de área, análisis de fallas y desempeño por subgrupos.

14. Explique por qué una AUC alta no demuestra utilidad clínica ni ausencia de sesgo.

Respuesta orientativa: La AUC mide discriminación relativa; no informa por sí sola calibración, valor predictivo en otra prevalencia, consecuencias de errores ni beneficio neto.

15. Enumere cinco tareas en las que la IA puede apoyar responsablemente la planificación y tres decisiones que no deben delegarse.

Respuesta orientativa: Apoyo: simulación, código de aleatorización, auditoría de consistencia, diccionario de datos, detección preliminar de fuga. No delegar: decisión ética, comparador clínico, efecto relevante, aprobación o responsabilidad profesional.

Capítulo 5. Bioestadística, inferencia, causalidad e inteligencia artificial

1. Un estudio registra 1 200 observaciones provenientes de 40 vacas durante 30 días. ¿Cuál es la unidad independiente y por qué una regresión ordinaria sería problemática?

Respuesta orientativa: La vaca es la unidad biológica repetida; las mediciones dentro de una vaca están correlacionadas. Se requiere un modelo longitudinal o mixto que represente esa dependencia.

2. Explique la diferencia entre desviación estándar y error estándar mediante un ejemplo de peso canino.

Respuesta orientativa: La desviación estándar describe diferencias de peso entre perros; el error estándar describe la precisión con que la muestra estima la media poblacional.

3. Un tratamiento produce p = 0,03 y una diferencia media de 0,2 puntos de dolor con IC 95 % de 0,02 a 0,38. ¿Qué información falta para juzgar relevancia clínica?

Respuesta orientativa: Falta el umbral de cambio clínicamente importante, además del contexto de seguridad, escala, seguimiento y comparador.

4. ¿Por qué p = 0,06 no demuestra ausencia de efecto? Formule una interpretación basada en un intervalo de confianza.

Respuesta orientativa: Indica que, bajo el modelo nulo, el resultado no cruzó el umbral elegido. El IC muestra qué efectos siguen siendo compatibles y si descarta o no magnitudes relevantes.

5. ¿Cuándo utilizaría t pareada y cuándo un modelo mixto en mediciones antes y después?

Respuesta orientativa: La t pareada sirve para dos mediciones completas por sujeto y una comparación simple. Un mixto maneja varios tiempos, datos desequilibrados, covariables y estructuras de correlación.

6. En un ANOVA con cuatro dietas, ¿por qué no es apropiado ejecutar seis pruebas t sin ajuste?

Respuesta orientativa: Porque el error familiar aumenta. Deben definirse contrastes o aplicar un procedimiento como Tukey o Holm según el objetivo.

7. Describa una situación en la que ajustar una covariable posterior al tratamiento introduzca sesgo.

Respuesta orientativa: Ajustar por inflamación medida después de una terapia al estimar su efecto total sobre recuperación puede bloquear mediación e introducir sesgo.

8. ¿Qué prueba no paramétrica usaría para tres grupos independientes con desenlace ordinal y qué limita la interpretación?

Respuesta orientativa: Kruskal-Wallis. Compara distribuciones mediante rangos y no prueba automáticamente igualdad de medianas.

9. Interprete un OR de 2,0 sin afirmar que el riesgo se duplicó.

Respuesta orientativa: Las odds del evento son dos veces mayores; para comunicar riesgo se necesitan el riesgo basal o probabilidades predichas.

10. ¿Qué evidencia sugeriría sobredispersión en un modelo Poisson de conteos parasitarios?

Respuesta orientativa: Varianza condicional muy superior a la media, residuos incompatibles y mayor frecuencia de conteos extremos que la esperada.

11. ¿Por qué medir 100 aves en un solo galpón por tratamiento constituye pseudorreplicación?

Respuesta orientativa: El tratamiento se asignó al galpón; las aves comparten exposición y no son réplicas independientes del tratamiento.

12. Defina censura e indique cuándo podría ser informativa en un estudio oncológico.

Respuesta orientativa: Es seguimiento incompleto sin evento observado. Puede ser informativa si animales que empeoran abandonan con más frecuencia.

13. Interprete un HR de 0,48 con IC 95 % de 0,26 a 0,89.

Respuesta orientativa: La tasa instantánea ajustada se estima 52 % menor; los datos son compatibles con reducciones entre 11 % y 74 %. No equivale a riesgo relativo ni prueba causalidad.

14. ¿Qué debe examinarse antes de atribuir una PERMANOVA significativa a diferencias entre centroides?

Respuesta orientativa: La homogeneidad de dispersiones y el esquema de permutación compatible con el diseño.

15. ¿Qué aporta un GAM frente a un polinomio de alto grado en una relación temperatura-producción?

Respuesta orientativa: Permite una función suave penalizada y evita imponer oscilaciones globales; aun así requiere seleccionar base, penalización y validación.

16. Proponga un predictor de hospitalización que pueda ser útil sin ser causal.

Respuesta orientativa: La clínica de origen puede anticipar ingreso por diferencias operativas, sin ser una causa biológica directa de enfermedad.

17. ¿Por qué una explicación SHAP no demuestra el efecto de intervenir sobre una variable?

Respuesta orientativa: SHAP atribuye una predicción dentro de un modelo y distribución de datos; no identifica el resultado de una intervención.

18. Distinga validación cruzada, validación temporal y validación externa.

Respuesta orientativa: La cruzada reutiliza particiones dentro del desarrollo; la temporal prueba un periodo posterior; la externa usa otra población o centro.

19. ¿Qué significa positividad en una comparación causal de tratamientos?

Respuesta orientativa: Que exista probabilidad positiva de recibir cada estrategia comparada para todos los perfiles de covariables relevantes.

20. Enumere cinco elementos estadísticos que deberían aparecer en el informe de un estudio veterinario reproducible.

Respuesta orientativa: Diseño y unidad experimental, estimando, modelo y supuestos, manejo de datos faltantes, estimación con IC, software y versión, y análisis de sensibilidad.

Capítulo 6. Gestión, calidad y preparación de datos veterinarios

1. ¿Por qué una historia clínica electrónica no puede interpretarse automáticamente como una base epidemiológica representativa?

Respuesta orientativa: Porque se genera durante la atención, depende del acceso y de decisiones clínicas, y contiene medición y codificación selectivas.

2. Diferencie paciente, visita, episodio y muestra mediante un ejemplo clínico.

Respuesta orientativa: El perro es paciente; cada consulta es visita; varias consultas por la misma dermatitis forman un episodio; cada tubo remitido es una muestra.

3. ¿Qué campos mínimos debe incluir un resultado de laboratorio para ser reutilizable?

Respuesta orientativa: Animal, muestra, tiempo de toma, analito, método, valor, unidad, límites y estado de validación.

4. Explique por qué varias radiografías del mismo perro no deben repartirse entre entrenamiento y prueba.

Respuesta orientativa: Comparten paciente, adquisición y contexto; dividirlas filtra información y sobrestima generalización.

5. Diseñe cinco campos de un diccionario para estudiar mastitis subclínica bovina.

Respuesta orientativa: Por ejemplo: cow_id, fecha de muestreo, recuento celular, cuarto mamario y método diagnóstico, con unidad y dominio.

6. ¿Qué diferencia existe entre validez, consistencia y exactitud?

Respuesta orientativa: Validez es cumplir el dominio; consistencia es no contradecir otros datos; exactitud es proximidad al valor verdadero.

7. Proponga una regla para detectar duplicados de visitas sin fusionar animales homónimos.

Respuesta orientativa: Combinar identificador estable o microchip, hora, clínica y evidencia de episodio, con revisión de conflictos.

8. ¿Cuándo un valor extremo de creatinina debe conservarse?

Respuesta orientativa: Cuando es verificable, clínicamente plausible y no proviene de unidad o transcripción errónea.

9. Compare MCAR, MAR y MNAR con ejemplos veterinarios.

Respuesta orientativa: MCAR no depende de información; MAR depende de observados; MNAR conserva dependencia del no observado tras condicionar.

10. ¿Por qué imputar antes de dividir entrenamiento y prueba constituye fuga de información?

Respuesta orientativa: Porque los parámetros de prueba influyen en la transformación aprendida por el modelo.

11. ¿Cuándo elegiría escalamiento robusto en lugar de estandarización z?

Respuesta orientativa: Ante distribuciones con extremos legítimos que afectarían fuertemente media y desviación.

12. ¿Qué problema produce codificar razas con números enteros consecutivos?

Respuesta orientativa: Introduce orden y distancias artificiales entre categorías nominales.

13. Mencione tres características derivadas que podrían anticipar mastitis sin usar información futura.

Respuesta orientativa: Descenso de producción previo, cambio de conductividad y reducción de rumia en una ventana anterior.

14. ¿Por qué 98% de exactitud puede ser inútil en una enfermedad con 2% de prevalencia?

Respuesta orientativa: Puede corresponder a predecir siempre la clase mayoritaria y tener sensibilidad nula.

15. ¿Qué diferencia existe entre interoperabilidad técnica y semántica?

Respuesta orientativa: La primera permite transmitir; la segunda conserva significado.

16. ¿Qué elementos permiten reconstruir una matriz analítica seis meses después?

Respuesta orientativa: Fuentes congeladas, diccionario, código, versiones, parámetros, registro de decisiones y sumas de verificación.

17. Identifique cuatro variables con fuga potencial en un modelo de hospitalización.

Respuesta orientativa: Código de cama, orden de ingreso, diagnóstico de alta y medicación posterior.

18. ¿Cómo evaluaría si la ausencia de temperatura difiere por clínica y gravedad?

Respuesta orientativa: Comparando porcentajes y modelos de ausencia por clínica, signos basales, horario y desenlace, sin afirmar que ello prueba MAR.

19. ¿Qué daños podría causar aplicar SMOTE antes de separar la prueba?

Respuesta orientativa: Generaría sintéticos relacionados con prueba y produciría una estimación optimista; además alteraría su prevalencia.

20. Explique por qué FAIR no implica publicar historias clínicas identificables.

Respuesta orientativa: FAIR admite autenticación y restricciones; la reutilización debe respetar ética, confidencialidad y gobernanza.

Capítulo 7. Aprendizaje automático supervisado en investigación veterinaria

1. ¿Qué elementos mínimos debe contener una pregunta predictiva veterinaria?

Respuesta orientativa: Población, momento cero, desenlace verificable, horizonte, predictores disponibles, contexto, usuario y acción prevista.

2. ¿Por qué el momento cero determina qué variables pueden utilizarse?

Respuesta orientativa: Porque simula el instante de uso; cualquier variable posterior introduce información que todavía no existiría.

3. Diferencie parámetro de hiperparámetro con un ejemplo de regresión y otro de random forest.

Respuesta orientativa: Un coeficiente logístico es parámetro; la intensidad de penalización es hiperparámetro. En un bosque, los valores de hojas son parámetros y mtry es hiperparámetro.

4. ¿Por qué la regresión logística debe incluirse como línea base en muchos problemas tabulares?

Respuesta orientativa: Porque proporciona una referencia transparente, regularizable y frecuentemente bien calibrada.

5. Explique cómo random forest reduce la varianza de un árbol individual.

Respuesta orientativa: Promedia árboles entrenados con muestras bootstrap y subconjuntos de variables, disminuyendo correlación e inestabilidad.

6. ¿Qué relación existe entre tasa de aprendizaje y número de árboles en gradient boosting?

Respuesta orientativa: Una tasa menor suele requerir más árboles; ambos se ajustan conjuntamente.

7. ¿Por qué SVM y kNN requieren escalamiento de predictores?

Respuesta orientativa: Porque sus cálculos dependen de productos internos o distancias y las unidades pueden dominar el resultado.

8. ¿Qué partición usaría si cada perro aporta cinco radiografías y por qué?

Respuesta orientativa: Partición por perro, manteniendo todas sus imágenes juntas para impedir memorización del paciente.

9. Diferencie validación cruzada agrupada de validación externa.

Respuesta orientativa: La agrupada remuestrea la misma fuente respetando grupos; la externa prueba otra fuente, centro, región o periodo.

10. ¿Cómo puede SMOTE producir fuga si se aplica antes de dividir los datos?

Respuesta orientativa: Las observaciones sintéticas pueden usar vecinos que luego quedan en validación, transfiriendo información.

11. ¿Qué patrón de rendimiento sugiere sobreajuste?

Respuesta orientativa: Rendimiento alto en entrenamiento y claramente inferior en validación o prueba.

12. ¿Por qué un AUC alto no garantiza probabilidades útiles?

Respuesta orientativa: Porque AUC mide ordenamiento, no correspondencia entre probabilidad y frecuencia ni utilidad de un umbral.

13. ¿Cómo cambia el valor predictivo positivo cuando cambia la prevalencia?

Respuesta orientativa: En general aumenta con prevalencia, aun si sensibilidad y especificidad se mantienen.

14. ¿Qué información aporta la pendiente de calibración?

Respuesta orientativa: Indica si las probabilidades son demasiado extremas o demasiado moderadas; el ideal es uno.

15. ¿Por qué una variable muy importante para XGBoost no debe interpretarse como causa?

Respuesta orientativa: La importancia describe dependencia predictiva dentro del modelo y puede reflejar confusión, proxy o proceso de atención.

16. ¿Cuándo elegiría un modelo más simple aunque tenga una AUC ligeramente menor?

Respuesta orientativa: Cuando la diferencia es incierta o pequeña y el modelo simple ofrece mejor calibración, estabilidad o implementación.

17. ¿Qué grupos examinaría para evaluar desempeño diferencial en un hospital de pequeños animales?

Respuesta orientativa: Especie, edad, sexo, raza o conformación, centro, gravedad, turno y disponibilidad de pruebas, según pertinencia y ética.

18. ¿Qué debe permanecer congelado durante la evaluación final?

Respuesta orientativa: El conjunto de prueba, la definición de métricas y el flujo seleccionado antes de la evaluación final.

19. ¿Qué diferencia existe entre predicción de hospitalización y efecto causal de hospitalizar?

Respuesta orientativa: La primera anticipa una decisión observada; la segunda pregunta qué ocurriría bajo una intervención y exige inferencia causal.

20. Mencione cuatro elementos que reportaría siguiendo TRIPOD+AI.

Respuesta orientativa: Fuentes y fechas, participantes y exclusiones, desenlace y predictores, faltantes, partición, ajuste, métricas con intervalos, calibración y disponibilidad.

Capítulo 8. Aprendizaje no supervisado, series temporales y detección de patrones

1. ¿Por qué un análisis sin variable objetivo no está libre de decisiones subjetivas?

Respuesta orientativa: Porque selección de variables, escala, métrica, algoritmo, hiperparámetros y unidad analítica incorporan supuestos.

2. ¿Cuándo conviene estandarizar variables antes de un PCA y qué información podría alterarse?

Respuesta orientativa: Cuando las magnitudes o unidades no son comparables; puede disminuir el peso de variables cuya varianza absoluta tiene significado biológico.

3. ¿Qué diferencia práctica existe entre enlace simple, completo y Ward?

Respuesta orientativa: Simple usa la pareja más cercana y encadena; completo usa la más distante y compacta; Ward minimiza el aumento de variación intragrupo.

4. ¿Por qué el valor máximo de silueta no basta para elegir el número de grupos?

Respuesta orientativa: Porque mide estructura interna, no estabilidad, significado clínico, generalización ni utilidad.

5. ¿Qué precauciones exige interpretar una proyección t-SNE o UMAP?

Respuesta orientativa: No interpretar ejes, distancias globales, tamaños o islas como categorías naturales sin análisis de sensibilidad.

6. Distinga anomalía puntual, contextual y colectiva con un ejemplo veterinario.

Respuesta orientativa: Puntual: temperatura imposible; contextual: baja rumia a una hora inusual; colectiva: descenso sostenido multivariable.

7. ¿Por qué exactitud global puede ser una métrica inadecuada para alertas sanitarias raras?

Respuesta orientativa: La clase normal domina; se necesitan sensibilidad por evento, valor predictivo positivo y falsas alarmas por animal-tiempo.

8. Explique soporte, confianza y lift en una regla de asociación.

Respuesta orientativa: Soporte mide frecuencia conjunta, confianza la probabilidad condicional observada y lift la compara con la frecuencia basal.

9. ¿Qué es fuga temporal y cómo podría ocurrir al predecir producción lechera?

Respuesta orientativa: Usar datos posteriores al origen del pronóstico, o ajustar preprocesamiento con todo el periodo.

10. ¿Qué diferencia hay entre predecir nuevos días de vacas conocidas y generalizar a vacas nuevas?

Respuesta orientativa: La primera permite información histórica individual; la segunda exige partición por animal y evalúa transportabilidad.

11. ¿Por qué un mapa de casos sin denominador puede inducir conclusiones erróneas?

Respuesta orientativa: Puede reflejar mayor densidad de animales, pruebas o notificación en vez de mayor riesgo.

12. ¿Qué evidencia adicional se necesita antes de denominar fenotipo a un conglomerado?

Respuesta orientativa: Estabilidad, replicación, asociación externa, plausibilidad y utilidad, sin confundirlas con causalidad.

13. ¿Cómo distinguiría una caída de actividad por enfermedad de un fallo del collar?

Respuesta orientativa: Auditar pérdidas, batería y posición; contrastar otros sensores, observación y examen clínico.

14. ¿Por qué un modelo de alta precisión no estima necesariamente el efecto de una intervención?

Respuesta orientativa: La predicción explota asociaciones; una intervención requiere comparador y control de confusión o aleatorización.

15. Proponga una validación externa para un sistema de detección temprana de cojera.

Respuesta orientativa: Congelar el modelo y evaluarlo prospectivamente en fincas, animales y dispositivos no usados, con confirmación clínica ciega.

Capítulo 9. Deep learning, visión computacional y diagnóstico por imágenes

1. ¿Por qué una red neuronal puede alcanzar alta exactitud sin haber aprendido la lesión?

Respuesta orientativa: Porque puede explotar atajos correlacionados con la etiqueta, como equipo, marcador, fondo, colimación o calidad. La validación por centro, el enmascaramiento y el análisis de activación ayudan a detectarlos.

2. ¿Cuál debe ser la unidad de partición cuando un animal aporta varias proyecciones?

Respuesta orientativa: El animal o, según la pregunta, el episodio clínico completo. Todas las imágenes correlacionadas deben permanecer en el mismo subconjunto.

3. ¿Qué diferencia existe entre detección y segmentación?

Respuesta orientativa: La detección localiza objetos mediante cajas; la segmentación clasifica píxeles y permite delinear bordes, áreas o volúmenes.

4. ¿Por qué AUROC puede ser insuficiente en enfermedad infrecuente?

Respuesta orientativa: Puede parecer alta aunque existan muchos falsos positivos. La curva precisión-recall, valores predictivos y carga de alertas reflejan mejor el problema bajo baja prevalencia.

5. ¿Qué aporta la calibración?

Respuesta orientativa: Evalúa si probabilidades predichas corresponden a frecuencias observadas; es necesaria para interpretar riesgo y elegir umbrales.

6. ¿Por qué los parches histológicos no son réplicas biológicas?

Respuesta orientativa: Porque comparten lámina y animal; la independencia y el tamaño muestral se determinan en niveles superiores.

7. ¿Cuándo una transformación de aumento es inapropiada?

Respuesta orientativa: Cuando crea anatomía, orientación, color o adquisición sin plausibilidad o destruye información diagnóstica.

8. ¿Cómo debe manejarse el desacuerdo entre evaluadores?

Respuesta orientativa: Con lectura independiente, medición de concordancia, reglas de adjudicación y categorías de incertidumbre, no mediante una etiqueta forzada sin registro.

9. ¿Qué es el sesgo de espectro?

Respuesta orientativa: La distorsión que aparece cuando la muestra no incluye la diversidad y dificultad de casos del uso previsto.

10. ¿Por qué una validación aleatoria interna no demuestra generalización?

Respuesta orientativa: Porque comparte centro, equipos, protocolos y población; puede conservar las mismas señales espurias.

11. ¿Qué prueba es prioritaria para un modelo multicéntrico?

Respuesta orientativa: Una evaluación en centros no usados para desarrollo, con métricas, calibración y análisis de subgrupos preespecificados.

12. ¿Qué demuestra un mapa Grad-CAM sobre causalidad?

Respuesta orientativa: No demuestra causalidad. Solo muestra sensibilidad espacial aproximada de la predicción bajo una implementación particular.

13. ¿Qué debe reportarse cuando el modelo puede abstenerse?

Respuesta orientativa: Cobertura, motivos de abstención, rendimiento entre casos aceptados y distribución de los excluidos.

14. ¿Cuál es la limitación de transferir un modelo de heridas humanas a animales?

Respuesta orientativa: Existe cambio de dominio por piel, pelo, pigmentación, anatomía y contexto; se requiere ajuste y validación veterinaria.

15. ¿Qué secuencia de evaluación precede a la adopción clínica?

Respuesta orientativa: Desarrollo, validación interna, validación externa, evaluación prospectiva silenciosa, estudio con usuarios y monitoreo posterior.

Capítulo 10. Procesamiento del lenguaje natural e inteligencia artificial generativa

1. ¿Por qué una búsqueda por palabras puede fallar ante negación y temporalidad?

Respuesta orientativa: Porque compartir vocabulario no determina si un signo está presente, negado, es histórico o pertenece a otro sujeto.

2. ¿Qué diferencia existe entre un embedding estático y uno contextual?

Respuesta orientativa: El estático asigna un vector estable a una palabra; el contextual cambia según la secuencia.

3. ¿Por qué las notas del mismo animal no deben dividirse entre entrenamiento y prueba?

Respuesta orientativa: Porque el estilo y contenido repetido generan fuga y una estimación optimista.

4. ¿Cuándo sería preferible una regla a un transformer?

Respuesta orientativa: Cuando la expresión es estable, la tarea es pequeña y la transparencia supera la ganancia esperada de complejidad.

5. ¿Qué dimensiones deben evaluarse en un resumen clínico?

Respuesta orientativa: Factualidad, completitud, relevancia, cronología, contradicciones y utilidad clínica.

6. ¿Por qué una puntuación alta en un examen no valida un modelo para decisiones clínicas?

Respuesta orientativa: Porque el entorno clínico incluye información incompleta, examen, consecuencias y distribución distinta.

7. ¿Cómo se verifica que una referencia respalda una afirmación?

Respuesta orientativa: Comprobando metadatos y leyendo la fuente primaria para vincularla con la proposición exacta.

8. ¿Qué diferencia hay entre seudonimización y anonimización?

Respuesta orientativa: La seudonimización conserva una clave o posibilidad de vinculación; la anonimización busca impedir identificación razonable.

9. ¿Qué elementos mínimos debe contener un prompt de extracción?

Respuesta orientativa: Objetivo, contexto, definiciones, restricciones, esquema, manejo de incertidumbre y evidencia textual.

10. ¿Por qué el F1 global puede ocultar un problema clínico?

Respuesta orientativa: Porque las clases frecuentes dominan el promedio y pueden ocultar baja sensibilidad en un signo raro.

11. ¿Qué información debe incluir una declaración de uso de IA?

Respuesta orientativa: Herramienta/versión, fecha, tarea, datos, prompts, parámetros, validación y responsabilidad humana.

12. ¿Qué controles aplicaría al código R generado por un asistente?

Respuesta orientativa: Inspección, pruebas con datos conocidos, verificación de paquetes, supuestos, semillas y comparación independiente.

Capítulo 11. Inteligencia artificial en medicina clínica, cirugía y especialidades

1. ¿Qué diferencia metodológica existe entre un modelo diagnóstico y uno pronóstico?

Respuesta orientativa: El diagnóstico estima una condición presente; el pronóstico estima un evento futuro y requiere un tiempo cero y horizonte definidos.

2. ¿Por qué una AUC alta no demuestra que usar el algoritmo mejore los desenlaces?

Respuesta orientativa: La AUC mide ordenamiento entre casos y no casos. No mide calibración, daños, carga, adopción ni efecto de una intervención basada en la salida.

3. ¿Qué fuga de información aparecería si se usa un vasopresor administrado después del inicio de hipotensión para predecirla?

Respuesta orientativa: El vasopresor es consecuencia del episodio o de su sospecha; incluirlo permite al modelo leer información posterior al momento de predicción.

4. ¿Por qué todos los fotogramas de un animal deben permanecer en una misma partición?

Respuesta orientativa: Las observaciones del mismo animal están correlacionadas. Separarlas produce contaminación entre entrenamiento y prueba y sobreestima rendimiento.

5. ¿Qué aporta la calibración que no aporta la sensibilidad?

Respuesta orientativa: La calibración determina si una probabilidad, por ejemplo 20 %, corresponde aproximadamente a una frecuencia observada de 20 %.

6. ¿Cómo cambia el valor predictivo positivo entre consulta primaria y hospital de referencia?

Respuesta orientativa: Cambia con la prevalencia y selección de casos; suele ser mayor en poblaciones con más enfermedad, aun con sensibilidad y especificidad similares.

7. ¿Qué condiciones pueden confundir un sistema de evaluación facial del dolor?

Respuesta orientativa: Conformación facial, enfermedad ocular, sedación, iluminación, ángulo, estrés, visibilidad y tipo de dolor.

8. ¿Por qué un modelo que asocia tratamiento oncológico con mortalidad no estima necesariamente el efecto del tratamiento?

Respuesta orientativa: Existe confusión por indicación: los pacientes más graves pueden recibir terapias más intensas y tener peor pronóstico por su enfermedad basal.

9. ¿Qué diferencia existe entre validar una medición automática y validar una decisión clínica?

Respuesta orientativa: La primera compara la salida con una referencia; la segunda demuestra que usarla cambia conducta y mejora beneficios sin daños desproporcionados.

10. ¿Cuándo podría una regresión logística ser preferible a gradient boosting?

Respuesta orientativa: Cuando el tamaño muestral es limitado, se necesita interpretabilidad y calibración estable, y la relación puede representarse adecuadamente.

11. ¿Qué métricas usaría para evaluar una alerta continua además de la AUC?

Respuesta orientativa: Tiempo de anticipación, alarmas falsas por hora o paciente, proporción de alertas accionables, fallos de señal y tiempo de respuesta.

12. ¿Qué información debería mostrar una interfaz clínica antes de que el veterinario acepte la salida?

Respuesta orientativa: Población de validación, horizonte, probabilidad, calidad de entrada, datos faltantes, incertidumbre, variables influyentes y límites de uso.

13. ¿Cómo evaluaría un modelo multimodal cuando falta una de las modalidades?

Respuesta orientativa: Con casos incompletos reales, comparando rutas por disponibilidad y reportando rendimiento para cada patrón de ausencia.

14. ¿Qué significa que un modelo emita una salida no evaluable?

Respuesta orientativa: Que la entrada o incertidumbre no permite una estimación segura; es una conducta de seguridad, no un fallo que deba ocultarse.

15. ¿Qué diseño permitiría estudiar si la implementación reduce complicaciones y no solo si predice bien?

Respuesta orientativa: Un estudio prospectivo de impacto, idealmente controlado o escalonado, con desenlaces clínicos, recursos, carga y eventos adversos.

Capítulo 12. Ganadería de precisión, producción, reproducción y bienestar animal

1. Explique la diferencia entre monitorización, detección, predicción y control con un ejemplo pecuario para cada función.

Respuesta orientativa: Monitorizar registra producción diaria; detectar identifica un episodio actual de cojera; predecir anticipa mastitis; controlar ajusta ventilación según una regla. Son funciones y evidencias distintas.

2. ¿Por qué la validación de un sensor debe preceder a la validación del algoritmo que utiliza sus datos?

Respuesta orientativa: Si la señal no mide con exactitud y estabilidad el fenómeno, el algoritmo aprenderá error o artefactos. Primero se establece precisión, sesgo, repetibilidad y condiciones de fallo del sensor.

3. Una alerta de mastitis tiene sensibilidad de 85 % y valor predictivo positivo de 12 %. ¿Qué significa cada cifra para el trabajo diario?

Respuesta orientativa: La sensibilidad de 85 % indica que detecta 85 de cada 100 episodios verdaderos. VPP de 12 % significa que solo 12 de cada 100 alertas corresponden a mastitis; la carga de revisión será alta.

4. ¿Qué fuga de información se produce si los días de una misma vaca aparecen en entrenamiento y prueba?

Respuesta orientativa: Existe fuga por identidad y autocorrelación: el modelo reconoce el patrón individual visto y la evaluación sobreestima su capacidad para vacas nuevas.

5. Describa tres factores que pueden reducir la transferencia de un detector de celo entre granjas.

Respuesta orientativa: Diferencias de raza y manejo; dispositivo/colocación y calidad de señal; definición o verificación del celo. También influyen clima, alojamiento y prevalencia.

6. ¿Cuándo sería preferible un modelo mixto frente a un algoritmo de boosting en un estudio de producción de leche?

Respuesta orientativa: Cuando el objetivo es estimar efectos y trayectorias interpretables con repetición por vaca y hato, especialmente con muestra moderada. Boosting puede considerarse para predicción compleja validada.

7. ¿Por qué el corral puede ser la unidad experimental aunque se registren cientos de animales?

Respuesta orientativa: Si el tratamiento se aplica al corral, todos los animales comparten la misma asignación y ambiente. El número de réplicas es el de corrales, no el de animales.

8. Proponga un estándar de referencia para validar alimentación, agresión y cojera.

Respuesta orientativa: Alimentación: observación directa sincronizada o sistema validado; agresión: etograma por observadores ciegos; cojera: examen locomotor y diagnóstico clínico con acuerdo documentado.

9. Explique por qué exactitud global es insuficiente para una conducta rara como agresión.

Respuesta orientativa: La conducta normal domina la exactitud. Deben informarse sensibilidad por episodio, VPP, AUPRC y falsas alertas por unidad de tiempo.

10. ¿Qué riesgos de bienestar pueden surgir por dependencia excesiva de sistemas automatizados?

Respuesta orientativa: Omisión de animales no detectados, retraso de inspección manual, intervenciones erróneas, estrés por dispositivos y normalización de fallos o inequidades entre grupos.

11. ¿Cómo evaluaría si una tecnología reduce realmente la huella ambiental y no solo la intensidad por unidad de producto?

Respuesta orientativa: Medir emisiones y recursos totales, producción total, rebotes, fabricación y vida útil del sistema mediante límites de análisis explícitos; no solo emisiones por litro o kilogramo.

12. Diseñe un protocolo de respuesta ante una alerta de caída de consumo en cerdos.

Respuesta orientativa: Confirmar integridad del sensor; revisar agua, alimento, ambiente y salud del lote; inspeccionar animales; registrar acción y resultado; escalar al veterinario; auditar falsos positivos y ajustar sin medicar automáticamente.

Capítulo 13. Epidemiología, salud pública veterinaria y enfoque One Health

1. ¿Por qué un mapa de conteos de casos no representa necesariamente un mapa de riesgo?

Respuesta orientativa: Porque el conteo depende de población, muestreo y acceso diagnóstico.

2. Diferencie una señal, un caso probable y un caso confirmado dentro de un sistema de vigilancia.

Respuesta orientativa: La señal requiere verificación; el caso probable cumple criterios parciales; el confirmado cumple el estándar establecido.

3. ¿Qué fuga de información puede ocurrir al calcular variables climáticas o incidencias móviles?

Respuesta orientativa: Usar información posterior a la fecha de predicción o calcular transformaciones con toda la serie.

4. ¿Por qué una partición aleatoria es inadecuada cuando existen celdas vecinas o aislamientos clonales?

Respuesta orientativa: Porque comparte estructura casi idéntica entre entrenamiento y prueba y sobreestima generalización.

5. ¿Qué diferencia existe entre discriminación y calibración?

Respuesta orientativa: La discriminación ordena unidades; la calibración compara probabilidades con frecuencias observadas.

6. ¿Por qué AUPRC puede ser más informativa que exactitud en brotes raros?

Respuesta orientativa: Porque se concentra en la clase positiva y su precisión frente a una prevalencia baja.

7. ¿Qué variables permiten distinguir aumento de enfermedad de aumento de esfuerzo diagnóstico?

Respuesta orientativa: Número de pruebas, cobertura, distancia al laboratorio, notificadores activos y retraso.

8. ¿Qué condiciones deben cumplirse antes de integrar datos humanos y animales?

Respuesta orientativa: Objetivo común, definiciones compatibles, base legal, protección, procedencia y plan de acción.

9. ¿Por qué la cercanía genómica no demuestra transmisión directa?

Respuesta orientativa: Puede existir una fuente intermedia no muestreada y la resolución depende del patógeno.

10. ¿Qué aporta una red de contactos frente a un análisis tabular convencional?

Respuesta orientativa: Representa rutas y nodos puente que no aparecen en variables independientes.

11. ¿Cómo evaluaría el cambio de dominio entre producción intensiva y familiar?

Respuesta orientativa: Con validación externa estratificada y análisis de cobertura, calibración y errores.

12. ¿Qué daño puede producir una comunicación imprecisa de "alto riesgo"?

Respuesta orientativa: Estigma, pérdidas económicas, acciones desproporcionadas y pérdida de confianza.

13. ¿Cuándo sería preferible un modelo de regresión o GAM frente a una red neuronal?

Respuesta orientativa: Cuando la muestra es limitada, se requiere interpretabilidad o la estructura epidemiológica puede especificarse.

14. ¿Qué métricas utilizaría para un sistema de alerta con capacidad limitada de inspección?

Respuesta orientativa: Sensibilidad, valor predictivo positivo, retraso, AUPRC, calibración y número de inspecciones.

15. ¿Qué evidencia adicional se requiere para pasar de asociación predictiva a recomendación causal?

Respuesta orientativa: Diseño causal, temporalidad, control de confusión, supuestos explícitos y análisis de intervención.

Capítulo 14. Validación, interpretación y evaluación crítica de modelos

1. Explique por qué una división aleatoria por imagen puede producir fuga de información cuando cada animal aporta varias radiografías.

Respuesta orientativa: Las imágenes del mismo animal comparten anatomía, técnica y artefactos; deben mantenerse en un solo pliegue.

2. Diferencie validación interna, temporal y geográfica mediante un ejemplo de mastitis bovina.

Respuesta orientativa: Interna: misma fuente con remuestreo; temporal: periodo posterior; geográfica: otra finca o región.

3. ¿Por qué el VPP cambia cuando se aplica el mismo modelo en una población con menor prevalencia?

Respuesta orientativa: Porque aumenta la proporción de falsos positivos respecto a verdaderos positivos aun si sensibilidad y especificidad se conservan.

4. ¿Qué interpreta realmente una AUC de 0,82 y qué no permite concluir?

Respuesta orientativa: Expresa ordenamiento entre un caso y un no caso; no es exactitud, calibración ni utilidad.

5. ¿Cuándo resulta más informativa una curva precision-recall que una curva ROC?

Respuesta orientativa: Cuando la clase positiva es rara y el costo de falsas alertas es central.

6. Compare MAE y RMSE en una predicción de producción de leche.

Respuesta orientativa: MAE resume error típico lineal; RMSE penaliza más los errores grandes.

7. ¿Cómo se interpretan un intercepto de calibración negativo y una pendiente menor que 1?

Respuesta orientativa: Sobreestimación global y predicciones excesivamente extremas, respectivamente.

8. ¿Por qué el umbral 0,50 puede ser inadecuado para una alerta de sepsis?

Respuesta orientativa: Porque el costo relativo de errores y la prevalencia pueden exigir un umbral mucho menor o mayor.

9. Diferencie importancia predictiva de efecto causal.

Respuesta orientativa: La primera describe contribución a predicción; la segunda requiere un estimando y supuestos causales.

10. Mencione tres limitaciones de SHAP para justificar una decisión clínica.

Respuesta orientativa: Dependencia de referencia, correlación entre predictores y ausencia de interpretación causal.

11. Diseñe una prueba de robustez para un clasificador de radiografías caninas.

Respuesta orientativa: Variar proyección, exposición, compresión, equipo y artefactos plausibles, con evaluación por subgrupo.

12. ¿Qué condiciones debe cumplir una comparación justa entre IA y radiólogos veterinarios?

Respuesta orientativa: Mismos casos, información, tiempo, estándar independiente y análisis pareado con efectos de lector y caso.

13. Diferencie validez estadística, clínica y biológica.

Respuesta orientativa: Rendimiento cuantitativo; beneficio en decisiones; coherencia con fisiopatología y medición.

14. ¿Qué aporta un estudio de impacto que no aporta una validación externa retrospectiva?

Respuesta orientativa: Mide cambios en decisiones, resultados, flujo, daño y adopción al usar el sistema.

15. Proponga cuatro indicadores para detectar deterioro de un modelo en operación.

Respuesta orientativa: Distribución de entradas, faltantes, tasa de alertas, calibración, discriminación y errores por subgrupo.

Capítulo 15. Reproducibilidad, comunicación científica y futuro de la IA veterinaria

1. Diferencie repetibilidad, reproducibilidad, replicabilidad y generalización mediante un ejemplo veterinario.

Respuesta orientativa: Repetibilidad: mismo equipo y análisis obtiene igual resultado; reproducibilidad: terceros reconstruyen el resultado con datos/código; replicabilidad: nuevo estudio confirma; generalización: el modelo mantiene desempeño en otra población.

2. Explique por qué un conjunto de datos FAIR no necesariamente es abierto.

Respuesta orientativa: FAIR exige localización, metadatos, condiciones de acceso, interoperabilidad y reutilización. Los datos clínicos sensibles pueden ser FAIR bajo acceso controlado sin hacerse públicos.

3. Proponga una estructura de proyecto para un estudio de supervivencia en oncología canina.

Respuesta orientativa: Carpetas separadas para datos originales de solo lectura, datos procesados, código, resultados, informes, documentación y entorno; identificadores estables y flujo ejecutable desde importación hasta tablas.

4. ¿Qué información del entorno de R debe conservarse para reconstruir un análisis?

Respuesta orientativa: Versión de R, sistema operativo, paquetes y versiones, dependencias, semillas, configuración regional, comandos de instalación o archivo de entorno y fecha de ejecución.

5. Compare ventajas y riesgos de scripts, cuadernos y reportes dinámicos.

Respuesta orientativa: Scripts son auditables pero pueden separar narrativa; cuadernos integran explicación y código pero favorecen ejecución fuera de orden; reportes dinámicos regeneran productos pero requieren entorno estable y control de dependencias.

6. ¿Qué debe contener una ficha de modelo veterinario?

Respuesta orientativa: Uso previsto, población, versión, datos, desenlace, predictores, arquitectura, preprocesamiento, métricas con IC, calibración, subgrupos, límites, no evaluables, validaciones, responsable y monitoreo.

7. Justifique por qué la partición por imágenes puede producir fuga en radiología veterinaria.

Respuesta orientativa: Las imágenes del mismo animal comparten anatomía, adquisición y artefactos; separarlas permite memorización y produce una prueba optimista.

8. Seleccione una guía de reporte para un estudio de exactitud de una CNN y explique su elección.

Respuesta orientativa: STARD-AI junto con STARD, porque se trata de exactitud diagnóstica de un sistema de IA; debe complementarse con información veterinaria y estructura por animal/centro.

9. ¿Por qué un AUC alto no garantiza utilidad clínica?

Respuesta orientativa: Porque el AUC no informa calibración, prevalencia, umbral, carga de falsas alertas, costos, flujo ni beneficio de actuar sobre la predicción.

10. Diseñe una secuencia de validación local, evaluación prospectiva y vigilancia.

Respuesta orientativa: Primero validar localmente un modelo congelado; luego realizar evaluación prospectiva silenciosa; después estudio de impacto con usuarios; finalmente desplegar con control de versión, deriva, incidentes y suspensión.

11. Enumere cinco costos ocultos de implementar un sistema de IA en una clínica o finca.

Respuesta orientativa: Integración con sistemas, anotación y limpieza, infraestructura, ciberseguridad, capacitación, revisión de alertas, mantenimiento, recalibración, soporte, licencias y gestión de incidentes.

12. Analice dos riesgos científicos de los modelos multimodales.

Respuesta orientativa: La desalineación o ausencia selectiva de modalidades puede sesgar; una modalidad dominante puede ocultar atajos o hacer que la aparente ganancia no se transporte. También aumenta complejidad y tamaño requerido.

13. ¿Qué diferencia un gemelo digital de un tablero de monitoreo?

Respuesta orientativa: El tablero muestra estados o indicadores; un gemelo digital representa procesos dinámicos y permite simular evolución o intervenciones bajo supuestos explícitos.

14. Proponga tres prioridades de investigación en IA veterinaria para Ecuador o América Latina.

Respuesta orientativa: Bases multicéntricas representativas de sistemas locales; validación de soluciones de bajo costo para producción y clínica; vigilancia One Health de zoonosis y resistencia con gobernanza y capacidad regional.

15. Explique cómo documentaría el uso de un modelo generativo durante la redacción de un artículo.

Respuesta orientativa: Registrar herramienta, versión, fecha, tarea, prompts relevantes, datos compartidos, salidas, verificaciones, errores corregidos y responsable humano; declarar el uso según política editorial sin atribuir autoría.

Acerca de los autores



Víctor Alejandro Bosquez Barcenes

Víctor Alejandro Bosquez Barcenes es docente investigador en la Universidad Estatal de Bolívar, ubicada en Guaranda, Ecuador. Posee una Licenciatura en Ciencias de la Educación mención Informática Educativa, una Especialidad en Redes de Comunicación de Datos, una Maestría en Informática Empresarial, una Maestría en Estadística Aplicada y un Doctorado en Educación. Cuenta con más de 15 años de experiencia docente en niveles de pregrado y posgrado, habiendo impartido módulos en diversas universidades ecuatorianas como la UPSE, UNEMI, UNACH, UTEQ y la Universidad Metropolitana, en maestrías relacionadas con educación, tecnología, inteligencia artificial, ciencias veterinarias y salud pública. Ha dirigido y participado en proyectos de investigación sobre accesibilidad web, inteligencia artificial aplicada a la educación y preservación digital, entre otros.



Joscelito Bolívar Solano Gaibor

Profesor e investigador de la Universidad Estatal de Bolívar, con una destacada trayectoria en gestión universitaria, educación superior e investigación. Ha ejercido como presidente de la Comisión de Intervención de la Universidad Técnica Luis Vargas Torres de Esmeraldas; administrador temporal del Consejo de Educación Superior (CES); funcionario de la Secretaría de Educación Superior, Ciencia, Tecnología e Innovación (SENESCYT); director de Relaciones Internacionales de la Universidad Estatal de Bolívar; y coordinador de la Unidad de Integración Curricular de la carrera de Medicina Veterinaria. Actualmente se desempeña como director de Posgrado de la Universidad Estatal de Bolívar.

Ha participado en proyectos Erasmus en colaboración con la Universidad de Valladolid y la Universidad de Porto. Y Proyectos AECID con la U de Almería. Es autor de artículos científicos publicados en revistas de alto impacto y coautor de libros sobre propedéutica, bioestadística y diseño experimental. Además, es autor de un libro dedicado a la crianza y producción de avestruces, gallinas de Guinea y faisanes.



Danilo Fabian Yanes Silva

El Dr. Danilo Fabián Yánez Silva es Médico Veterinario Zootecnista, graduado de la Universidad Estatal de Bolívar, y Magíster en Gerencia Educativa, con un Diplomado Superior en Planificación y Gestión Educativa. Cuenta con una amplia trayectoria profesional y académica vinculada a la medicina veterinaria, la docencia universitaria y la gestión institucional. Ha ejercido funciones como Coordinador de la Carrera de Medicina Veterinaria y Zootecnia y Vicedecano de la Facultad de Ciencias Agropecuarias de la Universidad Estatal de Bolívar. Su experiencia docente comprende asignaturas como Parasitología y Enfermedades Parasitarias, Producción y Salud Porcina, Propedéutica y Semiología, Epidemiología e Higiene de los Alimentos e Inspección Sanitaria. Asimismo, ha participado en procesos de capacitación, vinculación con la colectividad e investigación científica, con publicaciones relacionadas con sanidad animal, medicina veterinaria, epidemiología, producción agropecuaria y salud pública. Su trayectoria integra la formación profesional veterinaria, docencia, gestión académica e investigación.



Jonnathan Vinicio Arias Real

El Médico Veterinario Zootecnista Jonnathan Vinicio Arias Real, graduado de la Universidad Estatal de Bolívar, cuenta con aproximadamente diez años de experiencia profesional en medicina y cirugía de pequeñas especies. Es Magíster en Medicina Veterinaria, con mención en Clínica y Cirugía de Pequeñas Especies, por la Universidad Católica de Cuenca, y posee formación especializada en cirugía, anestesia veterinaria, ecografía e imagenología. Su experiencia comprende cirugía de tejidos blandos, cirugía reconstructiva, medicina de emergencias, anestesiología, diagnóstico por imagen y odontología veterinaria. En el ámbito académico, ha ejercido la docencia en el Instituto Superior Tecnológico San Gabriel de Riobamba y actualmente integra la Escuela de Medicina Veterinaria de la Universidad Estatal de Bolívar. Complementa su actividad profesional y docente con participación en investigación científica y producción académica, orientada al fortalecimiento de las ciencias veterinarias, la formación profesional y el bienestar animal.