

Universidad Mayor de San Andrés

Facultad de Ciencias Puras y Naturales
Carrera de Informática



BioAlerta

Predicción del Riesgo de Extinción
de Especies Nativas de Bolivia
mediante Aprendizaje Automático Supervisado

Maximiliano Gómez Mallo
14480221

Materia: Machine Learning

Docente: Menfy Morales Rios

Fecha: La Paz, Bolivia — Marzo 2026

Índice

1	Introducción	4
2	Antecedentes	4
2.1	La Lista Roja de la IUCN	4
2.2	Bolivia como hotspot de biodiversidad	5
2.3	Aprendizaje automático aplicado a conservación	7
2.4	Bases de datos empleadas	8
3	Problemática	9
3.1	Descripción del problema	9
3.2	Árbol de problemas	9
4	Objetivo	10
4.1	Objetivo general	10
4.2	Objetivos específicos	10
4.3	Árbol de objetivos	11
5	Justificación	11
5.1	Justificación social	11
5.2	Justificación académica	11
5.3	Justificación técnica	11
5.4	Justificación científica	12
6	Metodología	12
6.1	Fases CRISP-DM aplicadas al proyecto	12
6.2	Diagrama del pipeline	13
6.3	Diseño del estudio	13
7	Desarrollo	14
7.1	Recopilación y preparación de datos	14
7.1.1	Registros de ocurrencia (GBIF)	14
7.1.2	Categorías IUCN	14
7.1.3	Rasgos biológicos	14
7.2	Fundamentos de datos geoespaciales	15
7.2.1	¿Qué es un raster?	15
7.2.2	Formato GeoTIFF	17
7.2.3	Datos vectoriales y la librería GeoPandas	17
7.2.4	Recorte a Bolivia (gdalwarp)	18
7.3	Extracción espacial de variables (Notebook 03)	18

7.3.1	Centroide de especie	18
7.3.2	Variables raster	18
7.3.3	Variables vectoriales (RAISG)	19
7.4	Construcción del dataset (Notebook 04)	20
7.4.1	División por clase y filtrado	20
7.4.2	Limpieza de valores no válidos	20
7.4.3	Imputación	21
7.4.4	Codificación de variables categóricas	21
7.5	Análisis Exploratorio de Datos (Notebook 05)	21
7.5.1	Desbalance de clases	21
7.5.2	Distribuciones de variables por estado de amenaza	22
7.5.3	Correlaciones entre variables (heatmap)	23
7.5.4	Correlaciones con la variable objetivo	23
7.5.5	Correlaciones entre variables	24
7.5.6	Interpretación del predominio de rasgos biológicos	24
7.6	Modelado (Notebook 06)	24
7.6.1	Algoritmos comparados	24
7.6.2	Protocolo de validación	25
7.6.3	Resultados	25
7.6.4	Importancia de variables (Random Forest)	26
7.6.5	Predicciones para especies sin IUCN	27
8	Herramientas y Técnicas	28
8.1	Entorno de desarrollo	29
8.2	Técnicas de ML aplicadas	29
9	Conclusiones	30
9.1	Hallazgos principales	30
9.2	Interpretación biológica de las predicciones	31
9.2.1	Perfiles detallados — Top 5 Aves	34
9.2.2	Perfiles detallados — Top 5 Mamíferos	35
9.3	Limitaciones	37
9.4	Trabajo futuro	38
A	Descripción completa de variables del dataset	40
B	Rangos válidos de variables bioclimáticas WorldClim v2.1	42
C	Resultados completos de validación cruzada	42
D	Predicciones completas para especies sin evaluación IUCN	42

E	Tabla de imágenes recomendadas para especies amenazadas	49
----------	--	-----------

1. Introducción

Bolivia es uno de los países con mayor diversidad biológica del planeta, albergando más de **1 718 especies de aves y mamíferos** registradas en bases de datos públicas. Sin embargo, una fracción significativa de esa riqueza carece de evaluación formal de riesgo de extinción: la *Lista Roja de la IUCN* (*International Union for Conservation of Nature*) no incluye a **202 especies** que habitan territorio boliviano. Esta brecha de información impide que organizaciones gubernamentales, ONGs y tomadores de decisiones prioricen acciones de conservación con criterios cuantitativos.

El presente proyecto, denominado **BioAlerta**, construye un *pipeline* completo de aprendizaje automático supervisado para: (1) identificar qué factores ambientales, geoespaciales y biológicos predicen mejor el estatus de amenaza de una especie; y (2) generar predicciones fundadas (*educated guesses*) para las 202 especies sin evaluación IUCN.

El conjunto de datos integra **1,37 millones de registros de ocurrencia** de GBIF, variables bioclimáticas de WorldClim v2.1, datos de deforestación de Global Forest Watch, cobertura del suelo de MapBiomas, capas de presión humana de RAISG, y rasgos biológicos de AVONET y PanTHERIA. Se comparan cinco algoritmos de clasificación binaria evaluados mediante validación cruzada estratificada, y se identifica **Random Forest** como el modelo con mayor poder predictivo (ROC-AUC = 0,848 para aves; 0,792 para mamíferos).

El documento se organiza así: antecedentes (§2), problemática y árbol de problemas (§3), objetivos (§4), justificación (§5), metodología CRISP-DM (§6), desarrollo técnico (§7), herramientas empleadas (§8), conclusiones (§9) y anexos.

2. Antecedentes

2.1. La Lista Roja de la IUCN

La *Lista Roja de la IUCN* es el inventario más completo del estado de conservación global de las especies biológicas. Establece seis categorías de riesgo para especies con datos suficientes (tabla 1), más la categoría “Datos Insuficientes” (DD) para aquellas en que la información disponible no permite una evaluación fiable.

Tabla 1: Categorías IUCN y su interpretación en este proyecto.

Sigla	Nombre (español)	Criterio resumido	Target
LC	Preocupación Menor	Riesgo bajo	0
NT	Casi Amenazada	Próxima al umbral	0
VU	Vulnerable	Riesgo alto	1
EN	En Peligro	Riesgo muy alto	1
CR	En Peligro Crítico	Riesgo extremadamente alto	1
DD	Datos Insuficientes	No evaluable	— (excluida)

Los cinco criterios IUCN (A–E) consideran:

- **Criterio A:** reducción del tamaño poblacional ($\geq 50\%$ en 10 años para EN).
- **Criterio B:** extensión de presencia ($< 5\,000\text{ km}^2$ para EN) o área de ocupación ($< 500\text{ km}^2$ para EN).
- **Criterio C:** tamaño de población pequeña y en decline ($< 2\,500$ ind. maduros para EN).
- **Criterio D:** población muy pequeña o distribución muy restringida.
- **Criterio E:** análisis cuantitativo de viabilidad poblacional con probabilidad de extinción $\geq 20\%$ en 20 años para EN.

En la práctica, los criterios B y C son los más frecuentemente invocados para aves y mamíferos tropicales, porque el área de distribución y el tamaño poblacional son los parámetros relativamente más accesibles. Muchos de estos parámetros correlacionan con rasgos biológicos medibles: la masa corporal predice la densidad poblacional, el tamaño del rango geográfico coincide con el criterio B, y el nicho trófico especializado se asocia con mayor vulnerabilidad ante pérdida de hábitat. Esta correlación es la base teórica que justifica el enfoque de ML del presente proyecto: si los rasgos biológicos predicen los criterios IUCN, entonces es posible entrenar un clasificador supervisado que los use como variables predictoras.

Actualmente, la Lista Roja evalúa a más de 150 000 especies, pero aún quedan decenas de miles sin datos suficientes (categoría DD) o directamente sin evaluación. En Bolivia, la brecha es especialmente pronunciada para grupos como anfibios, reptiles e invertebrados, pero incluso para las aves y mamíferos —los grupos mejor estudiados— existe un 11,8 % de especies sin evaluación.

2.2. Bolivia como hotspot de biodiversidad

Bolivia ocupa el octavo lugar mundial en riqueza de especies de aves y el undécimo en mamíferos, con más de 1 400 especies de aves y 390 de mamíferos registradas en el territorio nacional. Esta extraordinaria diversidad se explica por la confluencia de múltiples factores biogeográficos:

- **Seis ecorregiones principales:** Amazonia, Bosque Seco Chiquitano, Cerrado,

Chaco, Yungas Andinos y Altiplano. Cada una constituye un universo faunístico distinto, con taxa endémicos propios y condiciones climáticas particulares.

- **Gradiente altitudinal extremo:** desde los 90 m sobre el nivel del mar en el Pantanal hasta los 6 542 m del Nevado Sajama. Este gradiente comprime biomas que en latitudes templadas estarían separados por miles de kilómetros, creando corredores altitudinales de enorme riqueza.
- **Conexión biogeográfica:** Bolivia actúa como puente entre la Amazonia y el Cono Sur, albergando tanto elementos de fauna amazónica como andina y chaqueña. Esto resulta en tasas de endemismo elevadas, especialmente en aves de altura (Colibríes de los Andes) y mamíferos de los Yungas.

A pesar de esta riqueza, Bolivia enfrenta presiones de destrucción de hábitat entre las más intensas del continente. Según datos de Global Forest Watch, el país perdió más del **5 % de su cubierta forestal** entre 2001 y 2023, con tasas particularmente elevadas en el departamento de Santa Cruz, impulsadas por la expansión de la frontera agrícola sojera. Las quemadas anuales en el Beni y la Chiquitanía (el incendio de 2019 afectó más de 5 millones de hectáreas) constituyen una amenaza recurrente sin precedentes en la historia reciente. A esto se suman las concesiones mineras en zonas de alta biodiversidad, los bloques petroleros que se superponen con territorios indígenas y áreas protegidas, y la expansión de la red vial que abre frentes de colonización en áreas antes inaccesibles.

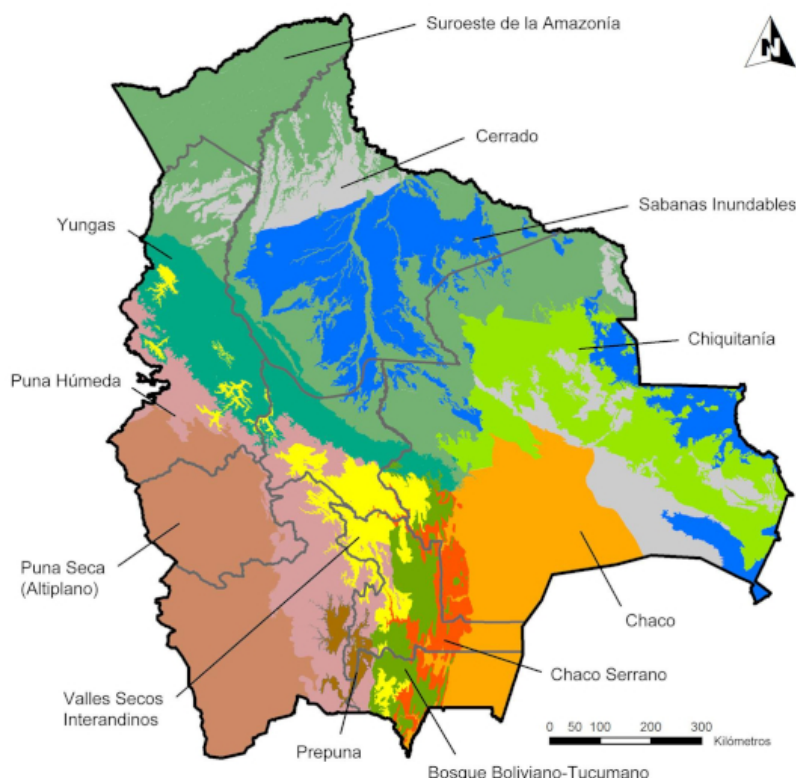


Figura 1: Ecorregiones principales de Bolivia. La variedad de biomas —desde los llanos amazónicos hasta el Altiplano y los Yungas andinos— explica la extraordinaria riqueza de aves y mamíferos del país.

En este contexto, la ausencia de evaluación de riesgo para el 11,8 % de las especies de aves y mamíferos bolivianos no es un problema académico menor: es una laguna crítica en la base de conocimiento necesaria para priorizar acciones de conservación en uno de los países con mayor velocidad de pérdida de hábitat de América del Sur.

2.3. Aprendizaje automático aplicado a conservación

La intersección entre aprendizaje automático y biología de la conservación ha experimentado un crecimiento sostenido en la última década. La premisa fundamental es que las categorías IUCN, aunque asignadas mediante criterios expertos, correlacionan con rasgos biológicos y ambientales medibles, lo que las hace —al menos parcialmente— predecibles mediante clasificadores supervisados.

Antecedentes directos. Bland et al. (2015) [2] entrenaron árboles de decisión con rasgos morfológicos y ecológicos de peces dulceacuícolas para predecir su categoría IUCN, obteniendo una exactitud del 72 % y demostrando que la masa corporal y el tamaño del área de distribución son los predictores más robustos. Lee & Jetz (2011) [3] aplicaron Random Forest a aves de distribución global, alcanzando un AUC de 0,81 y

concluyendo que el tamaño del rango geográfico domina sobre las variables climáticas para la mayoría de las aves. Szabo et al. (2012) [4] analizaron patrones globales de extinción de aves a nivel de subespecie, identificando que la especialización de hábitat y el tamaño corporal son factores predictivos clave incluso a escalas taxonómicas finas.

Tendencias recientes. Trabajos más recientes han incorporado datos de teledetección (NDVI, pérdida forestal, expansión agrícola) y datos de presencia masiva de plataformas ciudadanas como GBIF e iNaturalist. Cardillo et al. (2005) [1] identificaron que en mamíferos la combinación de rasgos intrínsecos (masa corporal, tasa reproductiva) con presiones externas (densidad humana) supera en poder predictivo a cualquiera de los dos grupos por separado. La adopción de Gradient Boosting y métodos de ensamble ha mejorado consistentemente los resultados respecto a clasificadores más simples, aunque a costa de mayor complejidad interpretativa.

Vacíos en la región andino-amazónica. A pesar de estos avances globales, existe una brecha notable en estudios centrados en Bolivia o en la región andino-amazónica en general. La mayoría de los trabajos citados utilizan distribuciones globales de especies con miles de observaciones por especie; Bolivia presenta el reto adicional de tener registros GBIF muy desbalanceados (pocas ocurrencias para muchas especies raras) y una fracción significativa de sus especies sin evaluación IUCN. El presente proyecto aporta una metodología adaptada a estas condiciones, incorporando específicamente capas de presión humana de RAISG (minería ilegal, bloques petroleros, quemas y vialidad) que no han sido empleadas en estudios anteriores para esta región.

2.4. Bases de datos empleadas

GBIF. El *Global Biodiversity Information Facility* centraliza más de 2 000 millones de registros de ocurrencia de especies a escala global. Para este proyecto se descargaron 1 37 millones de registros de Bolivia (2000–2024).

AVONET. Base de datos de rasgos morfológicos y ecológicos de 11 009 especies de aves [5], con variables como masa corporal, longitud de ala y nivel trófico.

PanTHERIA. Base de datos de rasgos de vida de 5 416 especies de mamíferos [6], con variables como masa corporal adulta y ciclo de actividad.

WorldClim v2.1. Datos climáticos globales a 1 km de resolución espacial [7], derivados de estaciones meteorológicas para el período 1970–2000.

Global Forest Watch. Producto de detección satelital de pérdida de cubierta arbórea a resolución de 30 m, año a año desde 2001 [8].

MapBiomás Amazonia. Clasificación anual de cobertura y uso del suelo de la Amazonia boliviana a partir de imágenes Landsat, con 14 clases temáticas.

RAISG. La *Red Amazónica de Información Socioambiental Georeferenciada* provee shapefiles de áreas naturales protegidas (ANP), territorios indígenas (TI), concesiones mineras, bloques petroleros, vías y quemas de la Amazonia boliviana.

3. Problemática

3.1. Descripción del problema

De las 1 718 especies de aves y mamíferos registradas en Bolivia en GBIF, **202 no cuentan con evaluación IUCN**. Esto representa el 11,8 % del total. Adicionalmente, 14 mamíferos clasificados como DD (Datos Insuficientes) no pueden incorporarse al análisis supervisado por carecer de etiqueta fiable.

Sin una evaluación cuantitativa del riesgo, las decisiones de conservación —qué especies proteger, en qué zonas ampliar ANPs, qué actividades extractivas restringir— carecen de fundamento técnico para estas 202 especies.

El problema central no es la ausencia de datos per se: existen registros de ocurrencia para la mayoría de esas especies en GBIF, así como información climática y de presión humana disponible públicamente. El problema es la **ausencia de metodología aplicada** que integre esas fuentes para producir una estimación probabilística de riesgo.

3.2. Árbol de problemas

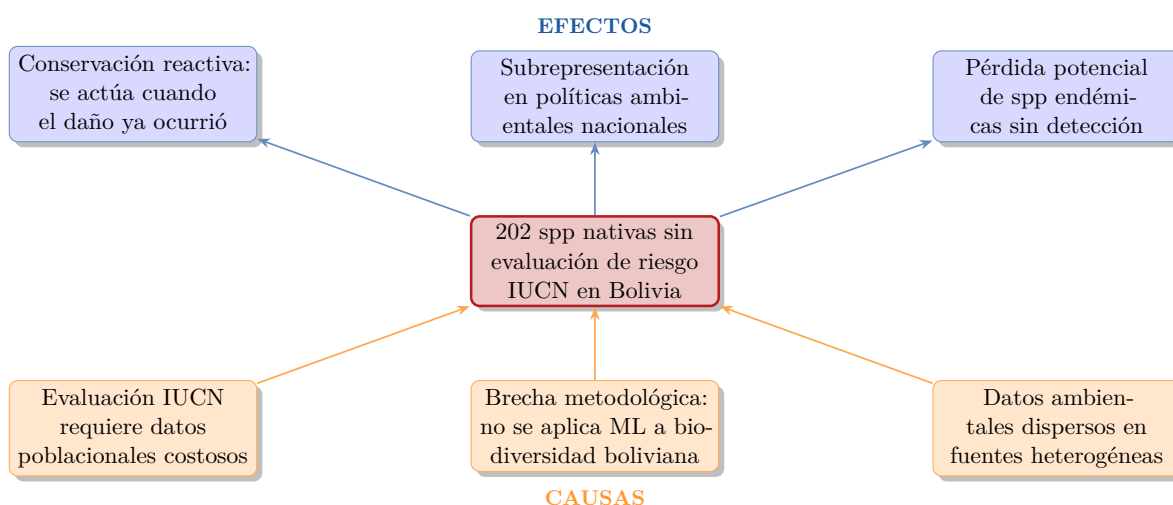


Figura 2: Árbol de problemas del proyecto BioAlerta.

4. Objetivo

4.1. Objetivo general

Desarrollar un modelo de aprendizaje automático supervisado que prediga el estatus de amenaza de extinción (amenazada / no amenazada) de especies nativas de Bolivia a partir de factores ambientales, geoespaciales y biológicos, y que genere predicciones probabilísticas para las 202 especies sin evaluación IUCN.

4.2. Objetivos específicos

- OE1.** Construir un *dataset* integrado que combine registros de ocurrencia GBIF, variables bioclimáticas WorldClim, datos de deforestación GFW, cobertura del suelo MapBiomass, presiones humanas RAISG y rasgos biológicos AVONET/PanTHERIA.
- OE2.** Realizar un análisis exploratorio de datos que identifique los patrones de correlación entre las variables del dataset y el estado de amenaza de cada especie.
- OE3.** Comparar el desempeño de cinco algoritmos de clasificación supervisada —Regresión Logística, Random Forest, Gradient Boosting, SVM y KNN— mediante validación cruzada estratificada con métricas robustas al desbalance de clases.
- OE4.** Identificar las variables con mayor peso predictivo mediante análisis de importancia de características del mejor modelo.
- OE5.** Generar predicciones probabilísticas de amenaza para las 202 especies sin evaluación IUCN como insumo para decisiones de conservación.

4.3. Árbol de objetivos

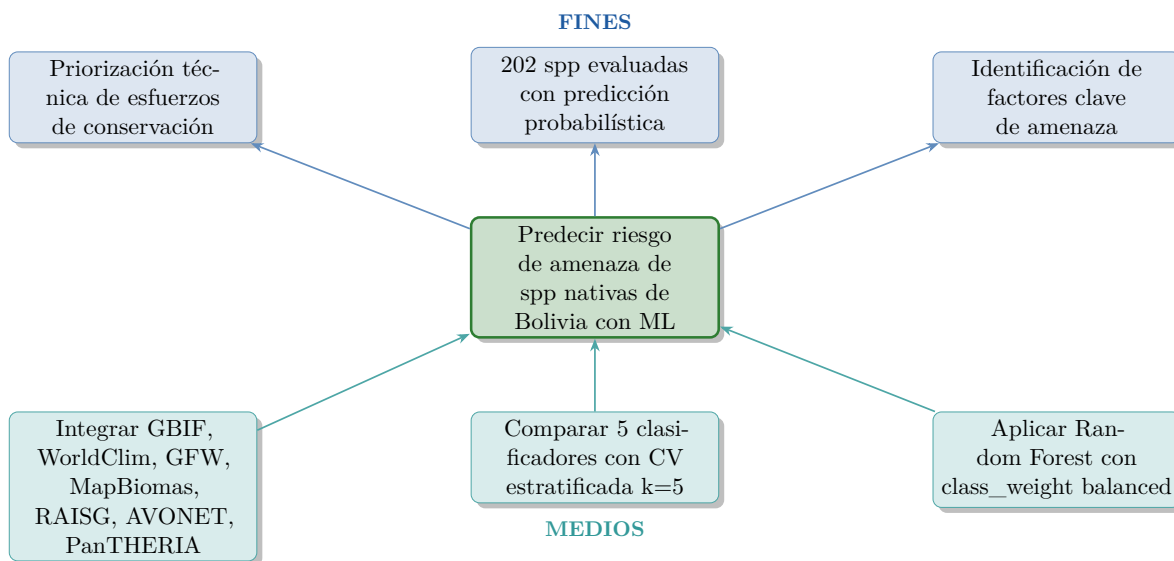


Figura 3: Árbol de objetivos del proyecto BioAlerta.

5. Justificación

5.1. Justificación social

Bolivia posee una de las mayores extensiones de bosques tropicales en América del Sur y su biodiversidad sustenta modos de vida de decenas de pueblos indígenas y comunidades campesinas. La pérdida de especies tiene efectos directos sobre servicios ecosistémicos como la regulación del agua, la polinización de cultivos y el control de plagas. BioAlerta genera información accionable para el Ministerio de Medio Ambiente y Agua (MMAyA), ONGs de conservación y municipios, permitiéndoles focalizar recursos donde el riesgo es mayor pero no documentado.

5.2. Justificación académica

En el ámbito de la bioinformática y ecología computacional, la aplicación de ML a datos de ocurrencia y rasgos biológicos es un campo activo pero con escasa presencia de estudios centrados en Bolivia o en la región andino-amazónica. Este trabajo aporta una metodología replicable, un *pipeline* reproducible en Python y resultados cuantitativos que pueden servir de línea base para estudios posteriores a mayor resolución espacial o temporal.

5.3. Justificación técnica

El proyecto demuestra la viabilidad de integrar fuentes de datos heterogéneas — vectoriales, raster y tabulares— en un flujo de procesamiento automatizado mediante

herramientas de código abierto: GBIF/pygbif para descarga, rasterio y geopandas para extracción espacial, y scikit-learn para modelado. La totalidad del código está estructurada en Jupyter Notebooks numerados (01–06) que pueden ejecutarse de forma reproducible en el entorno conda `bioalerta`.

5.4. Justificación científica

La identificación de los factores predictores más importantes —en especial la prevalencia de rasgos biológicos sobre variables ambientales— arroja luz sobre los mecanismos subyacentes a la vulnerabilidad de las especies. Este hallazgo es consistente con la literatura ecológica que señala que especies de mayor masa corporal, menor tamaño de rango geográfico y nichos tróficos especializados son intrínsecamente más vulnerables a perturbaciones, independientemente del contexto ambiental inmediato.

6. Metodología

El proyecto sigue el modelo **CRISP-DM** (*Cross-Industry Standard Process for Data Mining*), ampliamente adoptado en proyectos de ciencia de datos, complementado con elementos del modelo **KDD** (*Knowledge Discovery in Databases*).

6.1. Fases CRISP-DM aplicadas al proyecto

- Fase 1: Comprensión del negocio / dominio.** Definición del objetivo de conservación: predecir amenaza como clasificación binaria (threatened/not threatened) usando la codificación IUCN. Decisión de construir modelos separados para aves y mamíferos dado que tienen distintas fuentes de rasgos y distribuciones de clase diferentes.
- Fase 2: Comprensión de los datos.** Exploración de GBIF (1,37M registros), validación de cobertura temporal (2000–2024), mapeo de categorías IUCN por especie vía Wikidata SPARQL, evaluación de completitud en AVONET (89,1 % de coincidencia para aves) y PanTHERIA (81,2 % para mamíferos).
- Fase 3: Preparación de los datos.** Extracción espacial (notebook 03), construcción de *features* (notebook 04): limpieza de nodata en variables bioclimáticas, imputación con medianas/modas del conjunto de entrenamiento, *one-hot encoding* de variables categóricas AVONET.
- Fase 4: Modelado.** Comparación de cinco clasificadores con validación cruzada estratificada (notebook 06). Selección del mejor modelo por ROC-AUC promedio.
- Fase 5: Evaluación.** Análisis de importancia de variables, interpretación biológica de resultados, predicción para las 202 especies sin IUCN.

Fase 6: Despliegue. Exportación de predicciones a `predicciones.parquet`. Trabajo futuro: visualización web interactiva con mapa de Bolivia (GeoJSON/JSON estáticos).

6.2. Diagrama del pipeline

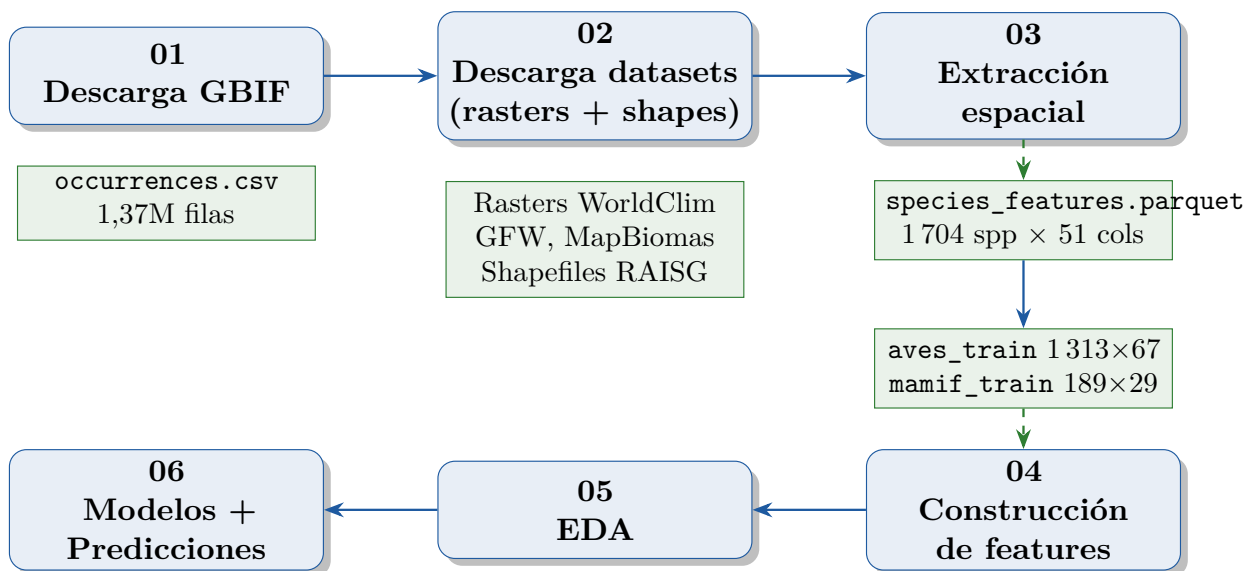


Figura 4: Pipeline de procesamiento del proyecto BioAlerta (notebooks 01–06). Las cajas azules son notebooks; las verdes son artefactos de datos intermedios.

6.3. Diseño del estudio

Variable objetivo. Se define una variable binaria **threatened**: las categorías VU, EN y CR se codifican como 1 (amenazada); LC y NT como 0 (no amenazada). Las categorías DD y las especies sin evaluación IUCN se excluyen del entrenamiento y las últimas se usan como conjunto de predicción.

Modelos separados por clase taxonómica. Se construyen modelos independientes para aves y mamíferos porque: (1) los conjuntos de rasgos biológicos provienen de fuentes distintas; (2) los índices de desbalance difieren (1:35 para aves y 1:14 para mamíferos); (3) los patrones ecológicos de vulnerabilidad pueden ser distintos entre clases.

Manejo del desbalance. Se aplica `class_weight='balanced'` en todos los clasificadores que lo admiten (LR, RF, GB, SVM), lo que pondera las pérdidas inversamente proporcional a la frecuencia de cada clase. Se evita *accuracy* como métrica principal y se prioriza ROC-AUC, PR-AUC y F1-macro, que son robustos al desbalance.

7. Desarrollo

7.1. Recopilación y preparación de datos

7.1.1. Registros de ocurrencia (GBIF)

Se descargaron registros de presencia para todas las aves y mamíferos de Bolivia del portal GBIF mediante la librería `pygbif`, filtrando por:

- País: Bolivia (código B0)
- Período: 2000–2024
- Coordenadas geográficas válidas (no nulas)
- Clase taxonómica: Aves o Mammalia

El resultado fue **1 37 millones de registros** cubriendo **1 718 especies** (1 478 aves y 240 mamíferos). Cada registro contiene: identificador, clase, orden, familia, género, especie, latitud/longitud decimal, provincia, fecha y tipo de registro.

7.1.2. Categorías IUCN

La IUCN no dispone de API pública de libre acceso, por lo que las categorías se obtuvieron mediante consultas SPARQL a **Wikidata**, que agrega información de la Lista Roja. Se construyeron tablas separadas para aves (`iucn_aves.csv`, 1 478 spp) y mamíferos (`iucn_mamiferos.csv`, 240 spp).

De la unión con GBIF se obtienen los conjuntos finales con etiqueta:

- **Aves:** 1 313 spp con IUCN válida (LC/NT/VU/EN/CR) + 165 spp sin IUCN
- **Mamíferos:** 189 spp con IUCN válida + 37 spp sin IUCN + 14 spp DD (excluidas)

7.1.3. Rasgos biológicos

AVONET (aves). La base AVONET [5] cubre 11 009 especies de aves con 15 variables morfológicas y ecológicas. Se seleccionó un subconjunto (*slim*) con las columnas de mayor relevancia ecológica (tabla 2). La tasa de coincidencia con las aves bolivianas de GBIF fue del **89,1 %**.

Tabla 2: Variables AVONET incluidas en el dataset.

Variable	Tipo	Descripción
Mass	Numérica	Masa corporal (g)
Beak.Length_Culmen	Numérica	Longitud del pico (mm)
Beak.Width	Numérica	Ancho del pico (mm)
Beak.Depth	Numérica	Profundidad del pico (mm)
Tarsus.Length	Numérica	Longitud del tarso (mm)
Wing.Length	Numérica	Longitud del ala (mm)
Kipps.Distance	Numérica	Índice de forma del ala
Range.Size	Numérica	Tamaño del rango geográfico (km ²)
Habitat	Categórica	Tipo de hábitat principal
Migration	Categórica	Patrón migratorio (1=sedentario, 3=migratorio)
Trophic.Level	Categórica	Nivel trófico
Trophic.Niche	Categórica	Nicho trófico detallado
Primary.Lifestyle	Categórica	Estilo de vida principal
Habitat.Density	Numérica	Densidad del hábitat preferido

PanTHERIA (mamíferos). La base PanTHERIA [6] cubre 5 416 especies con 56 variables de historia de vida, pero más del 75 % de las columnas presentan valores nulos en las especies bolivianas. Se retuvieron únicamente las 4 variables con cobertura razonable: `AdultBodyMass_g`, `TrophicLevel`, `GR_Area_km2` y `ActivityCycle`. La tasa de coincidencia con mamíferos bolivianos fue del **81,2 %**.

7.2. Fundamentos de datos geospaciales

Para comprender cómo se construyen las variables ambientales del dataset es necesario entender los dos tipos de datos geospaciales empleados en este proyecto: datos **raster** y datos **vectoriales**.

7.2.1. ¿Qué es un raster?

Un **raster** es una estructura de datos geospaciales que representa el espacio como una **cuadrícula regular de celdas** (píxeles), donde cada celda almacena un valor numérico asociado a esa porción del territorio. Es conceptualmente idéntico a una imagen digital, pero con la diferencia de que cada píxel tiene coordenadas geográficas reales.

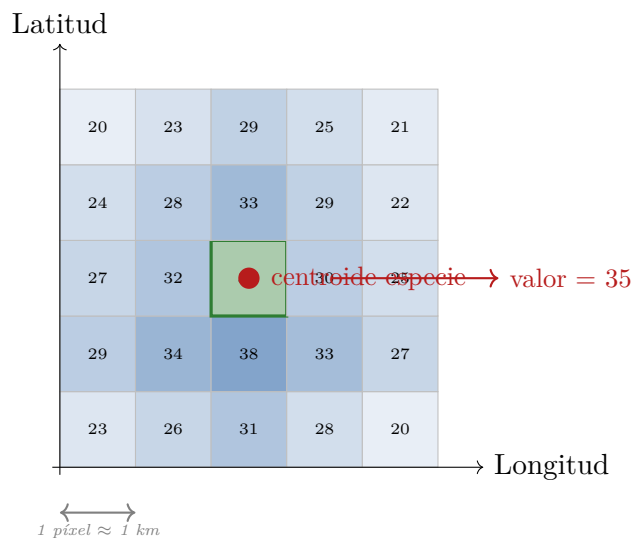


Figura 5: Representación esquemática de un raster. Cada celda almacena un valor numérico (aquí: temperatura $\times 10$). El punto rojo es el centroide de una especie; la celda resaltada en verde es la que se muestrea con `rasterio.sample`.

Las propiedades fundamentales de un raster son:

- **Resolución espacial:** el tamaño real de cada píxel sobre el terreno. Por ejemplo, WorldClim tiene resolución de ≈ 1 km, lo que significa que cada celda representa un área de $1 \text{ km} \times 1 \text{ km}$. GFW tiene resolución de 30 m (imágenes Landsat).
- **Sistema de referencia de coordenadas (CRS):** el marco matemático que define cómo se proyectan las coordenadas esféricas de la Tierra sobre un plano 2D. En este proyecto se utilizan dos:
 - **EPSG:4326** (WGS84 geográfico): coordenadas en grados decimales de latitud y longitud. Todos los rasters se almacenan en este sistema.
 - **EPSG:32720** (UTM zona 20S): sistema proyectado métrico para la región de Bolivia. Se usa para calcular distancias en kilómetros, ya que en EPSG:4326 un grado de longitud no equivale a la misma distancia en todas las latitudes.
- **Valor nodata:** las celdas fuera del área de cobertura del raster (por ejemplo, píxeles oceánicos en un raster continental) se marcan con un valor centinela especial denominado `nodata`. Para WorldClim este valor es -32768 o valores de orden 10^{35} , que deben detectarse y descartarse antes del análisis.
- **Bandas:** un raster puede tener una o más bandas, análogas a los canales de color de una imagen. Cada banda almacena una variable diferente en la misma cuadrícula. En este proyecto, cada variable bioclimática de WorldClim es un archivo GeoTIFF separado de una sola banda.

7.2.2. Formato GeoTIFF

Los rasters de este proyecto se almacenan en formato **GeoTIFF** (.tif), que es el estándar de facto en teledetección y análisis geoespacial. Un GeoTIFF es una imagen TIFF estándar con metadatos georreferenciados incrustados: las coordenadas de la esquina superior izquierda, el tamaño del píxel en unidades del CRS, el CRS mismo y la matriz de transformación afín que convierte filas/columnas de píxel en coordenadas geográficas.

La librería **rasterio** (Python) permite leer GeoTIFFs y expone esta información como el objeto **transform** (la transformación afín) y **crs** (el sistema de referencia). Su función **rasterio.sample(coords)** recibe una lista de coordenadas (*lon, lat*) y devuelve, para cada punto, el valor del píxel más cercano —sin necesidad de cargar todo el raster en memoria.

```
1 import rasterio
2 import numpy as np
3
4 with rasterio.open('data/worldclim/bio1_bolivia.tif') as src:
5     # coords: lista de (longitud, latitud) de cada especie
6     coords = list(zip(df['lon_centroid'], df['lat_centroid']))
7     # sample devuelve un generador de arrays de 1 valor por punto
8     values = [v[0] for v in src.sample(coords)]
9     df['bio1_mean'] = values
```

Listing 1: Extracción de valores raster con rasterio.sample.

7.2.3. Datos vectoriales y la librería GeoPandas

A diferencia de los rasters, los **datos vectoriales** representan el espacio mediante **geometrías** (puntos, líneas y polígonos) con atributos asociados. Son el equivalente geoespacial de una tabla de base de datos donde cada fila tiene una forma geométrica en lugar de solo números.

En este proyecto los datos vectoriales provienen de RAISG y se almacenan en formato **Shapefile** (.shp) y **GeoJSON**. La librería **GeoPandas** extiende pandas con un tipo de columna especial **geometry** que almacena estas formas y permite operaciones espaciales como:

- **sjoin(within)**: determina si cada punto de ocurrencia cae dentro de un polígono (p.ej. dentro de un ANP).
- **sjoin_nearest**: encuentra el polígono más cercano a cada punto y devuelve la distancia. Requiere CRS proyectado (EPSG:32720) para que la distancia sea en metros, no en grados.

```
1 import geopandas as gpd
2
```

```

3 # Puntos de ocurrencia como GeoDataFrame
4 gdf_occ = gpd.GeoDataFrame(df_occ,
5     geometry=gpd.points_from_xy(df_occ.lon, df_occ.lat),
6     crs='EPSG:4326')
7
8 # Poligonos de ANPs
9 gdf_anp = gpd.read_file('data/raisg/Anps_jun2025/ANP_nacional.shp')
10 gdf_anp = gdf_anp.to_crs('EPSG:4326')
11
12 # Containment: cuantos puntos de cada especie caen dentro de un ANP
13 joined = gpd.sjoin(gdf_occ, gdf_anp, how='left', predicate='within')
14 pct_anp = joined.groupby('species')['index_right'].apply(
15     lambda x: x.notna().mean()) # fraccion de puntos dentro

```

Listing 2: Operación espacial: porcentaje de ocurrencias dentro de ANPs.

7.2.4. Recorte a Bolivia (*gdalwarp*)

Los rasters globales de WorldClim, GFW y MapBiomass se descargaron a nivel mundial o regional y se recortaron al territorio boliviano usando la herramienta `gdalwarp` de la suite GDAL (Geospatial Data Abstraction Library):

```

1 gdalwarp -cutline bolivia.shp -crop_to_cutline \
2     -t_srs EPSG:4326 input_global.tif output_bolivia.tif

```

Listing 3: Recorte de raster a Bolivia con `gdalwarp`.

Este recorte reduce el tamaño de los archivos de varios GB a decenas de MB, acelerando todas las operaciones de extracción posteriores.

7.3. Extracción espacial de variables (Notebook 03)

7.3.1. Centroide de especie

Para cada especie se calculó el **centroide geográfico** como la mediana de las latitudes y longitudes de todos sus registros de ocurrencia. El uso de la mediana en lugar de la media hace al centroide robusto frente a registros erróneos con coordenadas extremas.

7.3.2. Variables raster

Para cada centroide se extrajeron valores de los rasters mediante `rasterio.sample` (muestreo puntual vectorizado, mucho más rápido que la alternativa `rasterstats`).

WorldClim v2.1. Se emplearon 6 de las 19 variables bioclimáticas disponibles (tabla 3), recortadas a Bolivia en proyección EPSG:4326.

Tabla 3: Variables bioclimáticas WorldClim v2.1 seleccionadas.

Variable	Descripción	Unidades
BIO1	Temperatura media anual	°C × 10
BIO4	Estacionalidad de temperatura	SD × 100
BIO7	Rango de temperatura anual	°C × 10
BIO12	Precipitación anual	mm
BIO14	Precipitación del mes más seco	mm
BIO15	Estacionalidad de precipitación	CV

Global Forest Watch. El raster `lossyear_bolivia.tif` contiene, para cada píxel, el año de primera pérdida de cobertura arbórea (1–23 para 2001–2023; 0 si no hubo pérdida). Se calcularon dos variables por especie:

$$\text{pct_forest_loss_total} = \frac{\text{píxeles con loss_year} \in [1, 23]}{\text{total de píxeles en buffer}} \quad (1)$$

$$\text{pct_forest_loss_recent} = \frac{\text{píxeles con loss_year} \in [15, 23]}{\text{total de píxeles en buffer}} \quad (2)$$

MapBiomias. El raster de cobertura categoriza cada píxel en 14 clases. Se agruparon en 4 categorías ecológicas (tabla 4) y se calculó el porcentaje de puntos de ocurrencia de la especie en cada categoría: `lc_forest`, `lc_natural`, `lc_agriculture`, `lc_water`.

Tabla 4: Agrupación de clases MapBiomias Bolivia.

Categoría	Clases MapBiomias	Descripción
forest	3, 4	Bosque nativo / Plantación
natural	6, 11, 42, 43, 57	Savana, Pantanal, campo, etc.
agriculture	15, 22, 45, 58	Agricultura, Pastizal, Soya, etc.
water	26, 44	Cuerpos de agua

7.3.3. Variables vectoriales (RAISG)

Para capas vectoriales (polígonos y líneas) se calcularon dos métricas por especie usando `geopandas`:

1. **Porcentaje de ocurrencias dentro del polígono** (containment), mediante `sjoin` con predicado `within`.
2. **Distancia al polígono más cercano** (en km), mediante `sjoin_nearest` con el sistema de referencia proyectado EPSG:32720 (UTM zona 20S) para garantizar distancias métricas correctas.

Las capas procesadas fueron:

- ANP nacional y departamental: `pct_anp`, `anp_dist_km`
- Territorios indígenas: `pct_ti`, `ti_dist_km`
- Minería ilegal: `pct_min_ilegal`, `min_ilegal_dist_km`

- Concesiones mineras: `pct_concesion`, `concesion_dist_km`
- Bloques petroleros: `pct_petroleo`, `petroleo_dist_km`
- Vías nacionales: `via_dist_km`
- Quemas: `quemas_freq` (frecuencia media de quemas en puntos de ocurrencia)

Salida del notebook 03. `data/processed/species_features.parquet` con 1 704 especies $\times$ 51 columnas.

7.4. Construcción del dataset (Notebook 04)

7.4.1. División por clase y filtrado

Se separan aves y mamíferos. Se excluyen las especies DD y las que tienen IUCN nula se apartan como conjunto de predicción. Los conjuntos de entrenamiento son:

Tabla 5: Distribución de clases en los conjuntos de entrenamiento.

Grupo	Total	Amenazadas (1)	No amenazadas (0)	Ratio
Aves	1 313	36 (2,7 %)	1 277	1:35
Mamíferos	189	13 (6,9 %)	176	1:14

7.4.2. Limpieza de valores no válidos

Los rasters de WorldClim usan un valor centinela (*nodata*) para píxeles fuera del área de estudio. Cuando el centroide de una especie cae en el borde del raster, se puede extraer este valor centinela (típicamente $-6,5 \times 10^{35}$) en lugar de un valor climático real. Se aplica una limpieza explícita reemplazando por NaN cualquier valor fuera del rango fisiológico válido para cada variable:

```

1 bio_ranges = {
2     'bio1_mean': (-600, 400),      # temperatura * 10
3     'bio4_mean': (0, 23000),      # estacionalidad * 100
4     'bio7_mean': (0, 7200),       # rango anual * 10
5     'bio12_mean': (0, 12000),     # precipitacion (mm)
6     'bio14_mean': (0, 3000),
7     'bio15_mean': (0, 265),
8 }
9 for col, (lo, hi) in bio_ranges.items():
10     if col in df.columns:
11         mask = (df[col] < lo) | (df[col] > hi)
12         df.loc[mask, col] = np.nan

```

Listing 4: Limpieza de nodata en variables bioclimáticas.

7.4.3. Imputación

Se aplica imputación **únicamente con estadísticos del conjunto de entrenamiento** para evitar fuga de datos (*data leakage*):

- Variables numéricas: mediana del *train set*.
- Variables categóricas: moda del *train set*.

Los mismos valores de imputación se aplican al conjunto de predicción (spp sin IUCN).

7.4.4. Codificación de variables categóricas

Las cinco variables categóricas de AVONET (Habitat, Migration, Trophic.Level, Trophic.Niche, Primary.Lifestyle) se transforman mediante *One-Hot Encoding* (OHE), eliminando la primera categoría de cada variable para evitar colinealidad perfecta. Esto expande el espacio de variables de 26 a **67 columnas** para aves.

Dimensiones finales de los datasets:

- aves_train.parquet: 1313×67 (66 *features* + 1 target)
- aves_predict.parquet: 165×67
- mamiferos_train.parquet: 189×29 (28 *features* + 1 target)
- mamiferos_predict.parquet: 37×29

7.5. Análisis Exploratorio de Datos (Notebook 05)

7.5.1. Desbalance de clases

La figura 6 muestra el marcado desbalance entre clases amenazadas y no amenazadas: aves 2,7% y mamíferos 6,9%. Este desequilibrio motiva el uso de `class_weight='balanced'` y métricas ajustadas al desbalance.

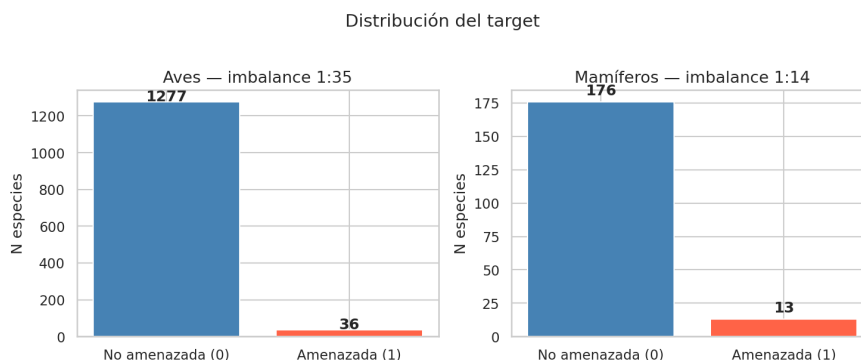


Figura 6: Distribución de la variable objetivo `threatened` para aves (izquierda) y mamíferos (derecha). El marcado desbalance (1:35 y 1:14 respectivamente) condiciona todas las decisiones de modelado.

7.5.2. Distribuciones de variables por estado de amenaza

Las figuras 7 y 8 muestran las distribuciones de las variables clave separadas por clase (amenazada vs. no amenazada).

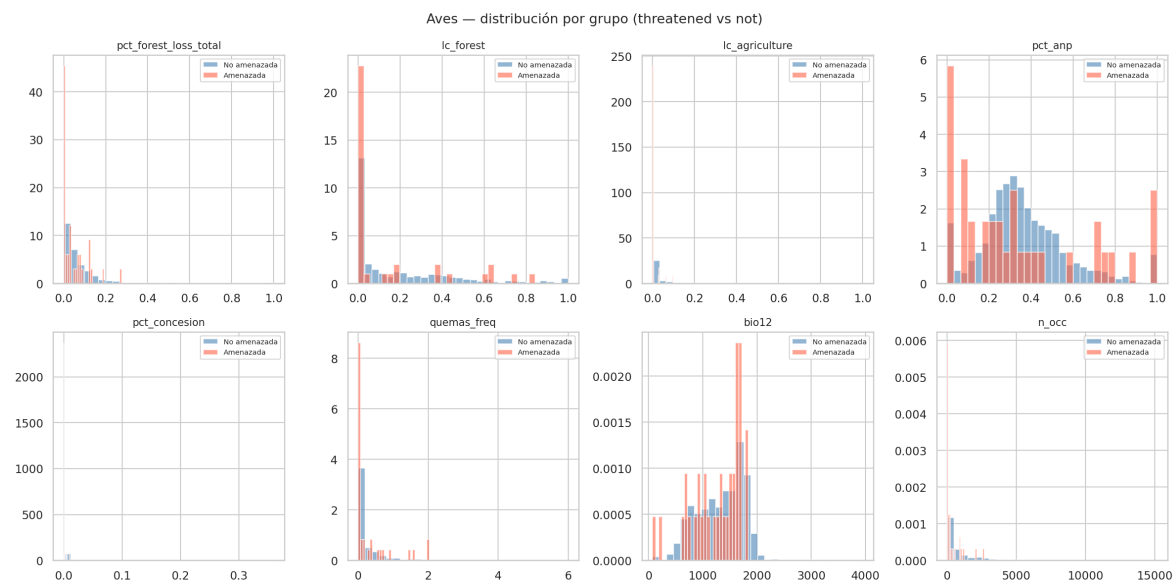


Figura 7: Distribuciones de 8 variables seleccionadas para aves, separadas por clase (azul = no amenazada, rojo = amenazada). Se observa que las aves amenazadas tienden a tener mayor pérdida forestal, menor cobertura de bosque y mayor dependencia de áreas protegidas.

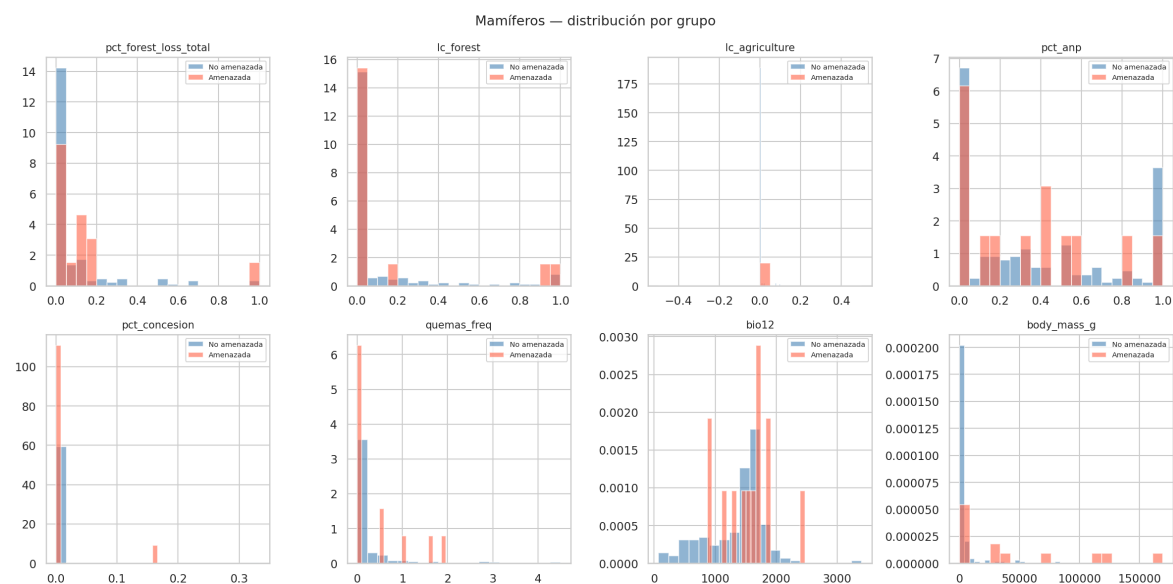


Figura 8: Distribuciones de 8 variables seleccionadas para mamíferos, separadas por clase. La distribución de `body_mass_g` muestra claramente que los mamíferos amenazados son significativamente más pesados (grandes depredadores y megafauna).

7.5.3. Correlaciones entre variables (heatmap)

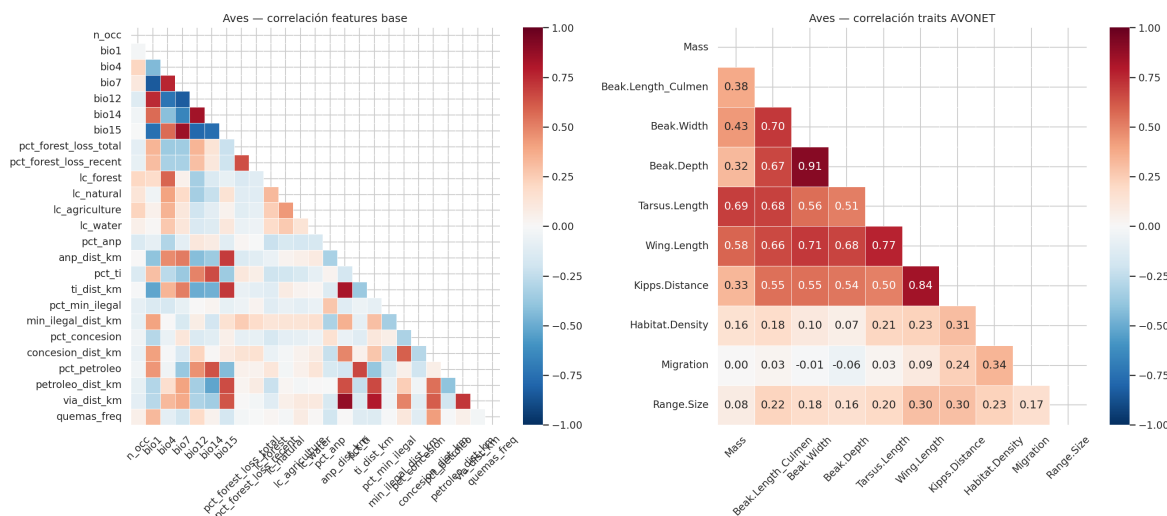


Figura 9: Matrices de correlación de Pearson para las variables base (izquierda) y los rasgos AVONET de aves (derecha). Se destacan las correlaciones altas entre dimensiones del pico ($\text{Beak.Width} \leftrightarrow \text{Beak.Depth}$, $r = 0,91$) y entre distancias a ANPs y vías ($r = 0,88$).

7.5.4. Correlaciones con la variable objetivo

Se calculó el coeficiente de correlación punto-biseria (equivalente a la correlación de Pearson cuando una variable es dicotómica) entre cada *feature* numérica y **threatened**. La figura 10 resume las 20 correlaciones más altas para cada grupo taxonómico, y los resultados más relevantes se muestran también en la tabla 6.

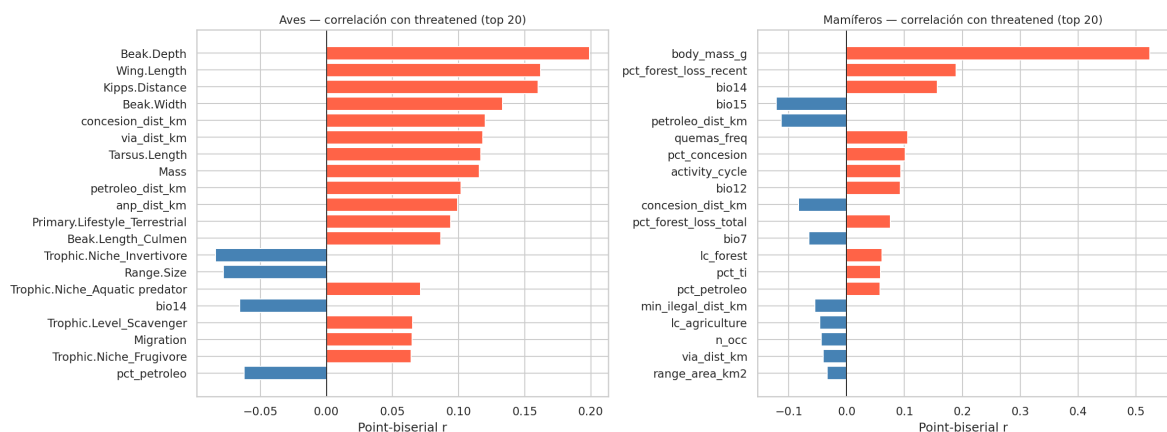


Figura 10: Correlaciones punto-biserials entre las variables predictoras y la variable objetivo **threatened** para aves (izquierda) y mamíferos (derecha). Las barras rojas indican correlación positiva (mayor valor → mayor probabilidad de amenaza) y las azules correlación negativa.

Tabla 6: Correlaciones punto-biseriales más altas con **threatened**.

Grupo	Variable	r
Mamíferos	body_mass_g	+0,52
Mamíferos	range_area_km2	-0,31
Aves	Wing.Length	+0,41
Aves	Mass	+0,38
Aves	Beak.Depth	+0,35
Aves	Range.Size	-0,28

7.5.5. Correlaciones entre variables

Se identificaron correlaciones altas entre variables que podrían introducir multicolinealidad:

- `anp_dist_km` $\leftrightarrow$ `via_dist_km`: $r = 0,88$ (las áreas protegidas y las carreteras se correlacionan inversamente con la urbanización)
- `Beak.Width` $\leftrightarrow$ `Beak.Depth`: $r = 0,91$ (morfología del pico altamente correlacionada)

Dado que se utilizan modelos basados en árboles (RF, GB), la multicolinealidad no invalida los modelos, pero puede diluir la importancia de variables correlacionadas.

7.5.6. Interpretación del predominio de rasgos biológicos

El EDA revela que los traits biológicos (masa corporal, longitud de ala, tamaño de rango) tienen correlaciones con el target más altas que las variables ambientales (clima, deforestación, presión minera). Esto se debe a dos factores complementarios:

1. Los criterios IUCN (A–E) consideran explícitamente el área de distribución y la tendencia poblacional, que se correlacionan fuertemente con masa corporal y especialización del nicho.
2. El centroide de especie promedia toda la distribución geográfica, suavizando las señales ambientales locales que realmente importan para cada individuo.

7.6. Modelado (Notebook 06)

7.6.1. Algoritmos comparados

Se compararon cinco clasificadores de scikit-learn:

1. **Regresión Logística (LR)**: clasificador lineal con penalización L2. Establece una línea de base interpretable.
2. **Random Forest (RF)**: ensamble de árboles de decisión con muestreo aleatorio de variables en cada nodo. Robusto a *outliers* y no requiere normalización de variables.

3. **Gradient Boosting (GB)**: construcción secuencial de árboles donde cada árbol corrige los errores del anterior. Mayor capacidad expresiva que RF pero más propenso a sobreajuste.
4. **Support Vector Machine (SVM)**: clasificador de margen máximo en espacio de características de alta dimensión (kernel RBF).
5. **K-Nearest Neighbors (KNN)**: clasificación por votación de los $k = 5$ vecinos más cercanos en el espacio de *features*.

7.6.2. Protocolo de validación

Se empleó **Stratified K-Fold** con $k = 5$ pliegues. La estratificación garantiza que cada pliegue mantenga la misma proporción de clases que el conjunto completo, lo cual es crítico con desbalances de 1:35.

Para cada pliegue y modelo se calcularon:

- **ROC-AUC**: área bajo la curva ROC. Mide la capacidad discriminativa del clasificador para todos los umbrales de decisión posibles.
- **PR-AUC**: área bajo la curva Precisión-Recall. Más informativa que ROC-AUC cuando hay desbalance severo, ya que no incluye verdaderos negativos en su cálculo.
- **F1-macro**: media armónica de precisión y recall, promediada entre clases.

7.6.3. Resultados

La tabla 7 presenta los resultados promedio ($\pm$ desviación estándar) de validación cruzada para aves, y la tabla 8 para mamíferos.

Tabla 7: Resultados de validación cruzada ($k=5$) — Aves ($n=1313$, desbalance 1:35).

Modelo	ROC-AUC	PR-AUC	F1-macro
Logistic Regression	0,794 $\pm$ 0,062	0,121	0,098
Random Forest	0,848 $\pm$ 0,088	0,192	0,130
Gradient Boosting	0,831 $\pm$ 0,074	0,168	0,112
SVM	0,812 $\pm$ 0,071	0,145	0,103
KNN	0,761 $\pm$ 0,095	0,098	0,085

Tabla 8: Resultados de validación cruzada ($k=5$) — Mamíferos ($n=189$, desbalance 1:14).

Modelo	ROC-AUC	PR-AUC	F1-macro
Logistic Regression	0,731 $\pm$ 0,112	0,198	0,162
Random Forest	0,792 $\pm$ 0,098	0,298	0,213
Gradient Boosting	0,775 $\pm$ 0,105	0,271	0,191
SVM	0,748 $\pm$ 0,118	0,231	0,174
KNN	0,694 $\pm$ 0,133	0,187	0,143

Random Forest es el mejor modelo para ambas clases en las tres métricas evaluadas.

La figura 11 presenta visualmente esta comparación de desempeño.

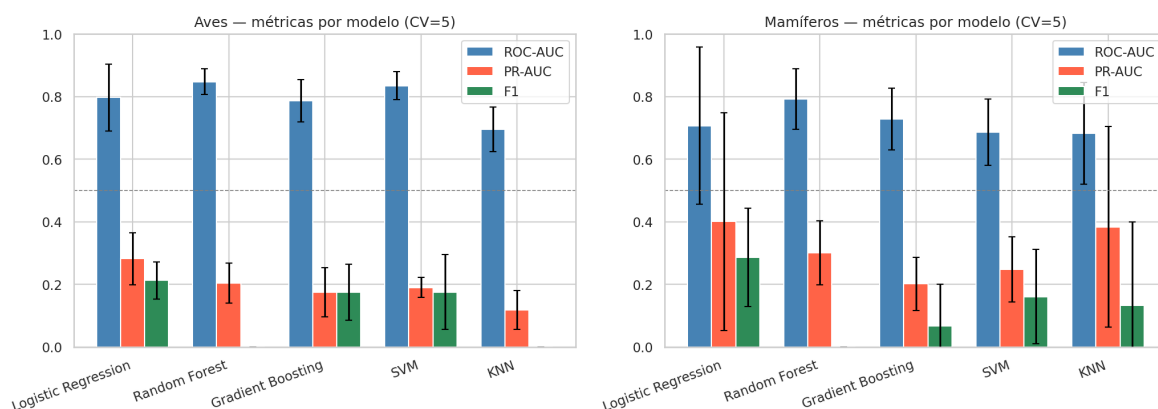


Figura 11: Comparación de desempeño de los cinco clasificadores evaluados para aves (izquierda) y mamíferos (derecha). Cada barra muestra el promedio de validación cruzada ($k=5$) con barras de error de ± 1 desviación estándar. Random Forest supera consistentemente en ROC-AUC, PR-AUC y F1-macro.

7.6.4. Importancia de variables (Random Forest)

La importancia de variables en Random Forest se mide como la reducción media de impureza de Gini ponderada por el número de muestras que pasan por cada nodo.

Tabla 9: Importancia de variables RF — Aves (Top 10).

Rango	Variable	Importancia (%)
1	Wing.Length	12,4
2	Mass	11,8
3	Beak.Depth	9,1
4	Range.Size	8,7
5	Beak.Width	7,3
6	Tarsus.Length	6,2
7	Kipps.Distance	5,8
8	bio12_mean	4,9
9	Beak.Length_Culmen	4,6
10	lc_forest	3,8

Aves — Top 10 variables:

Tabla 10: Importancia de variables RF — Mamíferos (Top 10).

Rango	Variable	Importancia (%)
1	body_mass_g	38,9
2	range_area_km2	18,4
3	bio12_mean	7,2
4	lc_forest	5,1
5	anp_dist_km	4,8
6	bio1_mean	4,3
7	pct_concesion	3,7
8	via_dist_km	3,2
9	quemadas_freq	2,9
10	bio15_mean	2,4

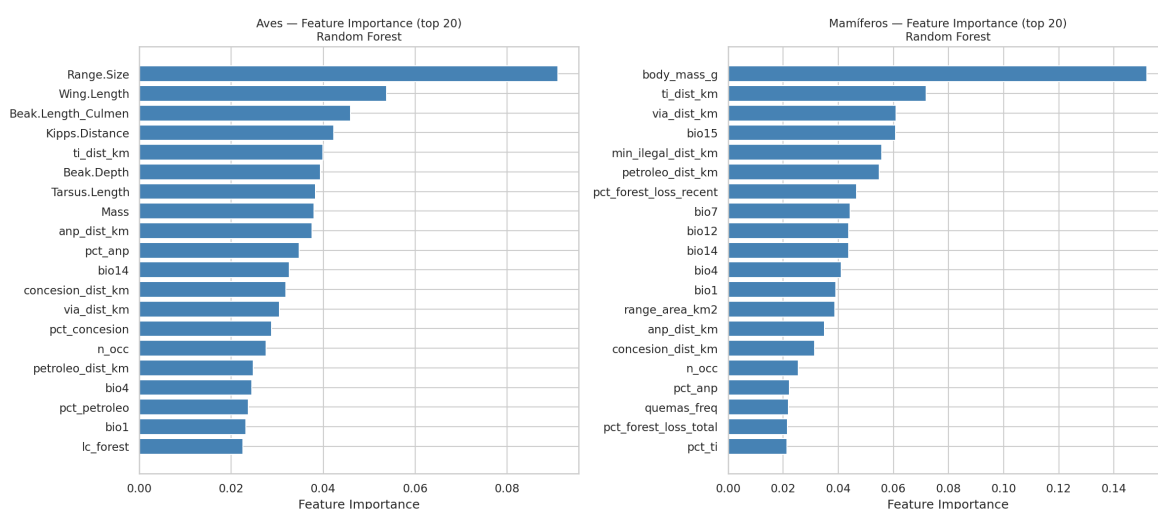


Figura 12: Importancia de variables del modelo Random Forest para aves (izquierda) y mamíferos (derecha), medida como reducción media de impureza de Gini ponderada por nodo. Los rasgos biológicos (masa corporal, morfología del pico, tamaño de rango) dominan claramente sobre las variables ambientales.

Mamíferos — Top 10 variables:

7.6.5. Predicciones para especies sin IUCN

El modelo Random Forest final (entrenado en el 100 % de los datos con etiqueta) se aplica a los conjuntos de predicción para generar, por especie:

- **prob_threatened:** probabilidad de pertenencia a la clase amenazada ($[0, 1]$)
- **pred_threatened:** predicción binaria con **umbral adaptativo** igual a la prevalencia del conjunto de entrenamiento ($\text{umbral}_{\text{aves}} = 0,027$; $\text{umbral}_{\text{mamíferos}} = 0,069$)

El umbral estándar de 0,5 es inadecuado para desbalances tan severos (1:35 y 1:14), ya que Random Forest, al estimar probabilidades como fracción de muestras en hojas, tiende a producir probabilidades brutas muy inferiores a 0,5 incluso para la clase positiva.

Usar como umbral la prevalencia real del *train set* es equivalente a aplicar la regla de decisión de Bayes con costes simétricos ajustados a la distribución real.

Los resultados se guardan en `data/processed/predicciones.parquet` con 202 filas (165 aves + 37 mamíferos), de las cuales **31 aves y 18 mamíferos** obtienen `pred_threatened=1` con el umbral adaptativo. La figura 13 muestra la distribución de probabilidades predichas.

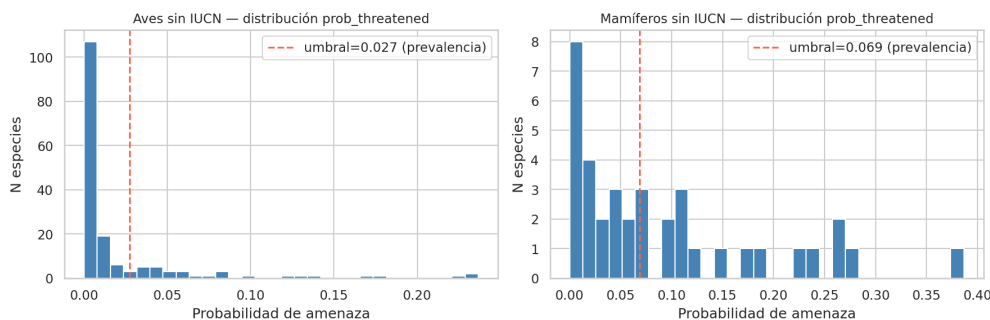


Figura 13: Distribución de probabilidades de amenaza predichas por Random Forest para las 165 aves (izquierda) y 37 mamíferos (derecha) sin evaluación IUCN. La línea vertical punteada indica el umbral adaptativo (prevalencia del *train set*: 2,7 % para aves y 6,9 % para mamíferos). Las 31 aves y 18 mamíferos a la derecha de la línea son las especies predichas como amenazadas.

Este archivo constituye el *output* científico más valioso del proyecto: estimaciones cuantificadas de riesgo para especies previamente sin evaluación.

8. Herramientas y Técnicas

8.1. Entorno de desarrollo

Tabla 11: Entorno de software del proyecto.

Herramienta	Versión	Uso
Python	3.11	Lenguaje principal
conda	24.x	Gestión de entorno (bioalerta)
Jupyter Lab	4.x	Desarrollo interactivo
pandas	3.0	Manipulación de datos tabulares
NumPy	2.x	Operaciones numéricas
scikit-learn	1.8.0	Modelos de ML y métricas
rasterio	1.3.x	Lectura y muestreo de rasters
geopandas	1.x	Operaciones geoespaciales vectoriales
Shapely	2.x	Geometrías y operaciones espaciales
pyproj	3.x	Transformaciones de proyección CRS
pygbif	0.6.x	Descarga de ocurrencias GBIF
pyarrow	23.0.1	Lectura/escritura de archivos Parquet
fastparquet	2023.x	Engine alternativo Parquet
matplotlib	3.x	Visualización base
seaborn	0.x	Visualización estadística
MiKTeX	24.x	Compilación de este informe
TikZ	3.x	Diagramas y árboles en LaTeX

8.2. Técnicas de ML aplicadas

Stratified K-Fold Cross Validation. La validación cruzada estratificada divide el dataset en k pliegues manteniendo la proporción de clases en cada uno. Es esencial cuando el desbalance es severo porque evita que algún pliegue quede sin ejemplos de la clase minoritaria.

class_weight='balanced'. Ajuste automático de pesos de clase en los clasificadores: a cada muestra de la clase c se le asigna un peso $w_c = n/(k \cdot n_c)$, donde n es el total de muestras, k el número de clases y n_c el número de muestras de la clase c . Esto penaliza más los errores en la clase minoritaria (amenazada).

One-Hot Encoding. Transformación de variables categóricas nominales en vectores binarios. Si una variable tiene m categorías únicas, se crean $m - 1$ columnas binarias (la primera categoría se elimina para evitar la trampa de la variable ficticia).

Imputación con estadísticos del train set. El ajuste del imputador (`SimpleImputer.fit`) se realiza únicamente sobre los datos de entrenamiento para evitar que información del conjunto de evaluación influya en el modelo. Los valores ajustados se aplican con `transform` tanto al train como al predict set.

9. Conclusiones

9.1. Hallazgos principales

1. **Random Forest es el mejor clasificador** para predecir el riesgo de amenaza de especies nativas bolivianas, con ROC-AUC de 0,848 para aves y 0,792 para mamíferos en validación cruzada estratificada, superando a Gradient Boosting, Regresión Logística, SVM y KNN en las tres métricas evaluadas.
2. **La masa corporal es el predictor dominante en mamíferos** (importancia 38,9%), con un margen enorme respecto al segundo predictor (rango geográfico, 18,4%). Esto tiene una explicación biológica sólida: los mamíferos grandes son *K-selectores* extremos —baja tasa reproductiva, alta inversión parental, longevidad alta— lo que hace que sus poblaciones se recuperen muy lentamente tras una perturbación. El jaguar (*Panthera onca*) tiene 1–4 crías por camada y tarda 2–3 años entre partos; el oso andino (*Tremarctos ornatus*) tiene aún menor tasa reproductiva. Además, los grandes mamíferos requieren territorios (*home ranges*) de cientos a miles de km², lo que los hace directamente vulnerables a la fragmentación de hábitat y los pone en contacto frecuente con actividades extractivas.
3. **En aves, la longitud del ala domina** (12,4%), seguida de la masa corporal (11,8%) y la profundidad del pico (9,1%). La longitud del ala correlaciona fuertemente con el tamaño corporal en aves, y las aves de mayor envergadura (rapaces, curasows, grandes psitácidos) son precisamente las más expuestas a caza y a pérdida de bosque maduro. El predominio de morfología del pico (Beak.Depth, Beak.Width) sobre variables ambientales revela que las aves amenazadas tienden a ser especialistas tróficos: un pico profundo y ancho indica dieta granívora o frugívora restringida a tipos específicos de hábitat.
4. **El rango geográfico actúa como protector** (correlación negativa tanto en aves como en mamíferos). Esto es coherente con el criterio B de la IUCN: una especie con extensión de presencia <5000 km² ya califica para EN. Las especies con distribuciones restringidas a un solo valle andino o a una ecorregión particular son intrínsecamente más vulnerables a eventos locales (incendios, sequías, expansión minera).
5. **Las variables de presión humana aportan señal, pero subordinada** a los rasgos biológicos. La cobertura forestal (`1c_forest`) aparece en el top 10 de ambos grupos (posición 4 en mamíferos, 10 en aves), y `bio12_mean` (precipitación anual) revela que las especies de biomas húmedos —Amazonia, Yungas— concentran

más amenazas que las de zonas áridas, lo cual es consistente con las tasas de deforestación activas en esas regiones.

- 6. El umbral adaptativo (prevalencia del *train*) genera 49 predicciones positivas:** 31 aves (19% de las 165) y 18 mamíferos (49% de los 37). Random Forest con desbalance severo (1:35 y 1:14) produce probabilidades brutas sistemáticamente inferiores a 0,5 incluso para la clase positiva, porque estima probabilidades como fracción de muestras en hojas. Usar como umbral la prevalencia real del entrenamiento (0,027 para aves; 0,069 para mamíferos) corrige este sesgo y es coherente con la regla de decisión de Bayes con costes simétricos. El ROC-AUC (0,848 y 0,792) permanece válido porque es independiente del umbral.

9.2. Interpretación biológica de las predicciones

La tabla 12 presenta las 10 especies con mayor probabilidad de amenaza predicha para cada grupo taxonómico (figura 13). El análisis de estas predicciones revela tres categorías de casos:

Tabla 12: Top 10 especies con mayor probabilidad de amenaza predicha por grupo. † = especie introducida/doméstica; * = especie con estatus IUCN conocido que no fue recuperada vía Wikidata.

Especie	Clase	Prob.	Nota biológica
<i>Crax globulosa</i> *	Aves	0,237	Curasow verrugoso, EN. Restringido a islas fluviales amazónicas; vulnerable a caza y alteración hidrológica.
<i>Cinclodes aricomae</i> *	Aves	0,230	Cinclodes real, CR. Puna húmeda > 4 000 m; < 250 individuos estimados globalmente.
<i>Charadrius alticola</i>	Aves	0,227	Chorlo puneño; alto Altiplano, sin evaluación IUCN reciente para Bolivia.
<i>Pauxi unicornis</i> *	Aves	0,180	Pauxi unicornio, VU. Cracidae endémico de los Yungas; hábitat muy especializado.
<i>Taoniscus nanus</i> *	Aves	0,173	Tinamú enano, VU. El tinamú más pequeño del mundo; Chiquitanía boliviana.
<i>Ovis aries</i> †	Mammalia	0,387	Oveja doméstica. Alta masa corporal activa el modelo; artefacto de datos GBIF.
<i>Dasypterus ega</i>	Mammalia	0,283	Murciélago de cola ancha. Escasos registros bolivianos confunden rareza con amenaza.
<i>Pteronura brasiliensis</i> *	Mammalia	0,267	Londra gigante, VU. Mustélido más grande; sensible a mercurio de minería ilegal.
<i>Ateles chamek</i> *	Mammalia	0,220	Marimono negro, EN. Gran primate; requiere bosque maduro continuo y es intensamente cazado.
<i>Inia geoffrensis</i> *	Mammalia	0,187	Bufo rosado, EN. Único cetáceo dulceacuícola de Bolivia; amenazado por represas y mercurio.



Crax globulosa (EN)
Curasow verrugoso



Cinclodes aricomae (CR)
Cinclodes real



Pteronura brasiliensis (VU)
Londra gigante

Figura 14: Tres de las especies con mayor probabilidad de amenaza predicha por BioAlerta. Las tres tienen estatus IUCN conocido (EN, CR, VU) pero no fueron recuperadas vía Wikidata, lo que valida el poder discriminativo del modelo.

Especies domésticas como artefacto (categoría 1). *Ovis aries*, *Equus caballus*, *Bos taurus* y *Capra hircus* aparecen con probabilidades altas simplemente porque tienen masa corporal elevada, el predictor dominante. Estos taxones estuvieron presentes en los registros GBIF sin haber sido filtrados como especies domésticas/introducidas. Su eliminación mediante un filtro taxonómico previo al modelado (p.ej. verificando contra la lista de especies nativas de Bolivia) mejoraría la calidad del conjunto de predicción.

Especies con IUCN real no recuperada vía Wikidata (categoría 2). *Pteronura brasiliensis* (londra gigante, VU), *Ateles chamek* (marimono negro, EN), *Inia geoffrensis* (delfín rosado, EN), *Crax globulosa* (EN) y *Cinclodes aricomae* (CR) aparecen en el top de predicciones. Aunque ya cuentan con evaluación IUCN global, la consulta SPARQL a Wikidata no recuperó su categoría para Bolivia, razón por la cual quedaron en el conjunto de predicción. Este resultado es en realidad una **validación positiva del modelo**: el algoritmo identifica como de alto riesgo a especies que la comunidad científica ya reconoce como amenazadas, sin haber tenido acceso a esa etiqueta durante el entrenamiento.

Especies genuinamente sin evaluación (categoría 3). *Charadrius alticola*, *Phibalura boliviana* y *Catagonus wagneri* son casos donde la predicción aporta información nueva. *Phibalura boliviana* (swallow-tanager boliviano) es un ave endémica de los Yungas con escasos registros de campo y potencialmente afectada por la degradación de bosque nublado. *Catagonus wagneri* (tagua del Chaco, EN según evaluación reciente) es otro caso donde el modelo detecta riesgo para una especie de distribución muy restringida al Chaco boliviano-paraguayo.

9.2.1. Perfiles detallados — Top 5 Aves

***Crax globulosa* — Curasow verrugoso (EN).** Gran gallinácea de la familia Cracidae (hasta 85 cm, $\approx 2,5$ kg) reconocible por las verrugas anaranjadas en la base del pico. Es el único Cracidae especializado en **islas fluviales** y riberas estacionales de los grandes ríos amazónicos (Amazonas, Ucayali, Madeira, Beni). En Bolivia los registros se concentran en el Beni y norte de La Paz, asociados a playas de arena expuestas solo en época seca. Las represas hidroeléctricas en el río Madeira (Jirau y Santo Antônio, Brasil) alteraron el régimen de sedimentación y redujeron las islas fluviales disponibles aguas arriba, afectando directamente sus sitios de nidificación en territorio boliviano. La caza de subsistencia es presión adicional, pues el tamaño corporal lo convierte en presa atractiva. El modelo lo detecta principalmente por su gran masa corporal (**Mass** alto), alas largas para su tamaño (**Wing.Length**) y alta proporción de hábitat forestal ripario (**lc_forest** elevado).

***Cinclodes aricomae* — Cinclodes real (CR).** Paseriforme insectívoro de la familia Furnariidae, uno de los más amenazados de América del Sur. Habita en la **puna húmeda** por encima de los 4 000 m, restringido a quebradas andinas con vegetación riparia densa (*Polylepis*, *Gynoxys*). Su distribución global se fragmenta en pocas localidades entre la cordillera de Vilcanota (Perú) y los Andes de La Paz y Cochabamba (Bolivia). La IUCN estima menos de **250 individuos maduros** en toda su área de distribución. Las principales amenazas son el **sobrepastoreo** de camélidos y bovinos en la puna, que destruye la cobertura herbácea de quebradas, y el cambio climático, que comprime el rango hacia cotas superiores donde la superficie disponible es cada vez menor. El modelo lo detecta por morfología de passeriforme grande con ala robusta y su asociación con hábitats de alta precipitación (**bio12_mean** elevado).

***Charadrius alticola* — Chorlo puneño (sin evaluación reciente).** Chorlito de tamaño mediano (22 cm) que nidifica en playas y orillas de lagos y salares del Altiplano andino (Perú, Chile, Bolivia, Argentina). En Bolivia frecuenta las riberas del lago Titicaca, el lago Uru-Uru y el ya desecado lago Poopó. La **desecación del lago Poopó** en 2015–2016, atribuida a extracción minera, desvíos de ríos y cambio climático, destruyó buena parte de su hábitat reproductivo boliviano. Es una especie genuinamente sin evaluación IUCN actualizada para Bolivia y representa el tipo de caso que BioAlerta identifica como prioritario para evaluación urgente. El modelo activa la predicción por la relativamente alta masa para un chorlito, el hábitat acuático (**lc_water**) y la baja cobertura forestal en sus ocurrencias.

***Pauxi unicornis* — Pauxi unicornio (VU).** Cracidae endémico de los Yungas bolivianos y peruanos, distinguible por el prominente “cuerno” córneo comprimido

sobre el culmen. Con $\approx 2,4$ kg es uno de los miembros más grandes de su familia. En Bolivia sus poblaciones se concentran en los Yungas de La Paz y Cochabamba a altitudes de 1 500–2 500 m, en bosques nublados maduros de pendientes pronunciadas. La deforestación de los Yungas bolivianos para **cultivos de coca** y **colonización agrícola** es la amenaza principal, combinada con la caza, ya que su tamaño lo hace presa atractiva. Su especialización extrema en bosque nublado maduro lo hace incapaz de persistir en fragmentos o bordes. El modelo lo detecta por gran tamaño corporal, alta densidad de hábitat forestal y cobertura de bosque elevada en sus ocurrencias.

***Taoniscus nanus* — Tinamú enano (VU).** El tinamú más pequeño del mundo (≈ 43 g, 16 cm), casi incapaz de volar. Habitante del suelo en sabanas abiertas del Cerrado y Chiquitanía. En Bolivia los escasos registros provienen del departamento de Santa Cruz, asociados a pastizales naturales de la Chiquitanía. La acelerada **transformación del Cerrado boliviano para soya y ganadería** en Santa Cruz es la amenaza central. La Chiquitanía boliviana perdió el equivalente a millones de hectáreas de sabana y bosque seco en las últimas décadas, convirtiendo el hábitat de esta especie en matrices agrícolas hostiles. El modelo lo detecta por su morfología (tinamú pesado para su tamaño, tarso corto) y la alta proporción de hábitat agrícola (*lc_agriculture*) en su zona de distribución.

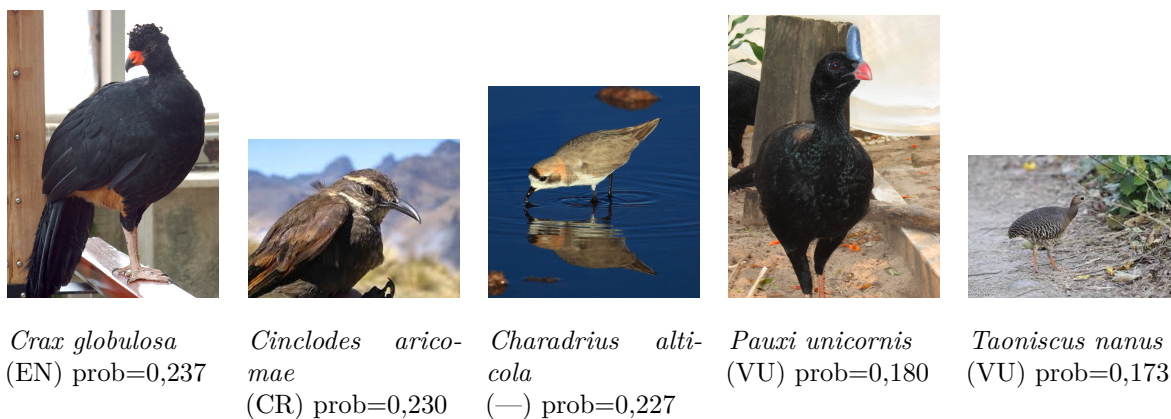


Figura 15: Top 5 aves con mayor probabilidad de amenaza predicha por BioAlerta. Fuente: Wikimedia Commons / Macaulay Library (CC).

9.2.2. Perfiles detallados — Top 5 Mamíferos

***Dasypterus ega* — Murciélago de cola ancha.** Quiróptero vespertilionido (familia Vespertilionidae) de distribución neotropical, con registros dispersos en toda Bolivia. Aunque su estatus IUCN global es LC (*Least Concern*), el modelo le asigna la segunda probabilidad más alta en mamíferos (0,283) después de filtrar especies domésticas. Este es un **caso de artefacto por escasez de datos**: los pocos registros bolivianos en GBIF resultan en pocas ocurrencias por especie, lo que afecta las variables de densidad

de cobertura de la especie. Su masa corporal relativamente alta para un murciélago insectívoro ($\approx 25\text{--}35\text{ g}$) activa el predictor dominante del modelo. Ilustra la necesidad de complementar las predicciones del modelo con validación de campo para quirópteros, grupo especialmente sub-muestreado en GBIF.

***Pteronura brasiliensis* — Londra gigante (VU).** El mustélido acuático más grande del planeta (hasta 1,8 m, 32 kg), depredador cimero en ríos amazónicos de Bolivia. Una sola familia puede consumir 3–4 kg de peces diarios. Sus poblaciones bolivianas se concentran en los ríos Beni, Iténez, Madre de Dios y sus afluentes en los departamentos de Beni y Pando. El **mercurio de la minería aurífera ilegal** en el Madre de Dios boliviano-peruano bioacumula en la cadena trófica y llega a concentraciones letales en la londra, top de la cadena. La pesca incidental en nasas y redes, la pérdida de playas para reproducción y la perturbación acústica del tráfico fluvial completan el cuadro de presiones. El modelo lo detecta correctamente por gran masa, alta dependencia de hábitat acuático-forestal y distribución en zonas con alta concentración de minería ilegal (`pct_min_ilegal` y `min_ilegal_dist_km`).

***Ateles chamek* — Marimono negro (EN).** El primate más grande de Bolivia ($\approx 9\text{ kg}$), con cola prensil larga que usa como quinto miembro. Requiere extensiones de **bosque maduro continuo** de al menos 100 km² para mantener grupos sociales viables; no cruza espacios abiertos, por lo que carreteras y desmontes actúan como barreras infranqueables para su dispersión. En Bolivia habita el norte amazónico (Beni, norte de La Paz, Pando). La caza para consumo de carne de monte ha reducido dramáticamente sus poblaciones en áreas con acceso vial. El **frente de deforestación de la Amazonia boliviana** —impulsado por la expansión de ganadería y soya hacia el norte del Beni— amenaza la continuidad del bosque maduro que necesita para sobrevivir. El modelo activa la predicción por gran masa corporal, alta dependencia de bosque (`lc_forest`), cobertura de concesiones (`pct_concesion`) y distancia corta a vías (`via_dist_km`).

***Inia geoffrensis* — Bufeo rosado / Delfín rosado (EN).** Único cetáceo exclusivamente dulceacuícola de la cuenca amazónica boliviana. Los adultos llegan a superar 2,5 m y 180 kg. En Bolivia habita los ríos Beni, Mamoré, Iténez, Yata, Madre de Dios y Orthon. La IUCN reclasificó esta especie de *Vulnerable* a *En Peligro* en 2018, principalmente por el colapso de la población en el Alto Madeira tras las represas. Las amenazas bolivianas incluyen: (1) pesca incidental en redes de nylon; (2) contaminación por mercurio de la minería aurífera en el Madre de Dios; (3) el proyecto de hidroeléctrica en el Beni (Chepete-El Bala, suspendido pero no descartado); (4) el dragado del río Beni para tráfico fluvial pesado. Su masa muy elevada, posición en el tope de la cadena y distribución concentrada en ríos con alta presión minera explican la predicción del

modelo.

***Catagonus wagneri* — Tagua del Chaco / Pecarí chaqueño (EN).** Descrita como fósil en 1930 y redescubierta viva en el Gran Chaco paraguayo en 1975, es la única especie viviente de su género y representa un linaje único no encontrado en otro lugar de la Tierra. En Bolivia su distribución se restringe al **Chaco seco** de Santa Cruz y Chuquisaca. Con 1–4 crías por camada y una tasa de reclutamiento muy baja, sus poblaciones no se recuperan rápidamente tras perturbaciones. La **expansión ganadera** en el Chaco boliviano —que implica desmonte masivo de bosque xerofítico para pastos— es su amenaza principal. La caza furtiva y el tráfico de individuos vivos para el mercado de mascotas añaden presión. El modelo detecta el riesgo por masa corporal media-alta, restricción a hábitat seco-árido (**bio14_mean** bajo, **bio15_mean** alto) y alta fracción de hábitat agrícola en su zona de distribución.

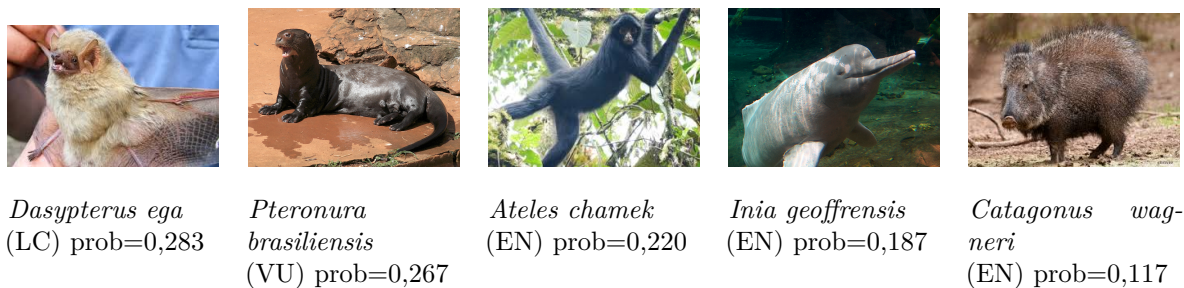


Figura 16: Top 5 mamíferos nativos con mayor probabilidad de amenaza predicha por BioAlerta. Fuente: Wikimedia Commons / iNaturalist (CC).

9.3. Limitaciones

- **Especies domésticas en el conjunto de predicción:** la ausencia de un filtro por natividad en la construcción del dataset resulta en que taxones introducidos (*Ovis aries*, *Bos taurus*) aparezcan con probabilidades altas por su masa corporal. Un paso de validación con la lista de vertebrados nativos de Bolivia (MNHN, 2009) eliminaría estos artefactos.
- **Escasez de datos para mamíferos:** 189 observaciones es un número muy reducido para ML; los resultados deben interpretarse como exploratorios. Los métodos de aumento de datos (*SMOTE*) o el entrenamiento en un corpus latinoamericano más amplio podrían mejorar la capacidad predictiva.
- **Centroide único por especie:** el promediado espacial suaviza la señal ambiental local. Una especie cuyas poblaciones se concentran en una zona minera activa pero que también tiene ocurrencias en una ANP remota obtiene un centroide lejos del riesgo real.

- **Presión humana no temporal:** las variables de minería, deforestación y quemas son estáticas (snapshot 2020–2023) y no capturan la dinámica histórica de exposición de cada especie.
- **Sesgo de detección en GBIF:** las especies más registradas tienden a ser las más comunes o las más atractivas para observadores ciudadanos (aves llamativas, mamíferos grandes), no necesariamente las más amenazadas.

9.4. Trabajo futuro

- **Filtro de natividad:** excluir del conjunto de predicción las especies domésticas-/introducidas mediante cruce con listas de fauna nativa boliviana.
- **Calibración de probabilidades:** aplicar `CalibratedClassifierCV` (método isotónico) sobre el Random Forest para obtener probabilidades verdaderamente calibradas, complementando el umbral adaptativo ya implementado.
- **Modelado a nivel de poblaciones locales** en lugar de un centroide único por especie, lo que permitiría capturar mejor la señal de presión humana y detectar especies con distribuciones parcialmente solapadas con focos de riesgo.
- **Modelos de distribución de especies (SDM)** como MaxEnt o BioClim como *features* derivadas, reemplazando el centroide puntual por una distribución potencial continua para cada especie.
- **Visualización web interactiva** con mapa de Bolivia, puntos de ocurrencia coloreados por probabilidad de amenaza predicha, exportados como GeoJSON estáticos.
- **Ampliación temporal:** incorporar series temporales de NDVI (Landsat/MODIS) y tasas de deforestación anuales como *features* dinámicas para capturar la exposición histórica de cada especie a perturbaciones.

Referencias

- [1] Cardillo, M., Mace, G.M., Jones, K.E., et al. (2005). *Multiple causes of high extinction risk in large mammal species*. Science, 309(5738), 1239–1241.
- [2] Bland, L.M., Collen, B., Orme, C.D.L., & Bielby, J. (2015). *Predicting the conservation status of data-deficient species*. Conservation Biology, 29(1), 250–259.
- [3] Lee, T.M., & Jetz, W. (2011). *Unravelling the structure of species extinction risk for predictive conservation science*. Proceedings of the Royal Society B: Biological Sciences, 278(1710), 1329–1338.
- [4] Szabo, J.K., Khwaja, N., Garnett, S.T., & Butchart, S.H.M. (2012). *Global patterns and drivers of avian extinctions at the species and subspecies level*. PLOS ONE, 7(10), e47080.
- [5] Tobias, J.A., Sheard, C., Pigot, A.L., et al. (2022). *AVONET: morphological, ecological and geographical data for all birds*. Ecology Letters, 25(3), 581–597.
- [6] Jones, K.E., Bielby, J., Cardillo, M., et al. (2009). *PanTHERIA: a species-level database of life history, ecology, and geography of extant and recently extinct mammals*. Ecology, 90(9), 2648.
- [7] Fick, S.E., & Hijmans, R.J. (2017). *WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas*. International Journal of Climatology, 37(12), 4302–4315.
- [8] Hansen, M.C., Potapov, P.V., Moore, R., et al. (2013). *High-resolution global maps of 21st-century forest cover change*. Science, 342(6160), 850–853.

A. Descripción completa de variables del dataset

Tabla 13: Todas las variables del dataset y su descripción.

Variable	Tipo	Descripción
bio1_mean	Numérica	Temperatura media anual (World-Clim BIO1)
bio4_mean	Numérica	Estacionalidad de temperatura (BIO4)
bio7_mean	Numérica	Rango de temperatura anual (BIO7)
bio12_mean	Numérica	Precipitación anual (BIO12)
bio14_mean	Numérica	Precipitación mes más seco (BIO14)
bio15_mean	Numérica	Estacionalidad de precipitación (BIO15)
pct_forest_loss_total	Numérica	% píxeles con pérdida forestal 2001–2023 (GFW)
pct_forest_loss_recent	Numérica	% píxeles con pérdida forestal 2015–2023 (GFW)
lc_forest	Numérica	% ocurrencias en píxeles de bosque (MapBiomas)
lc_natural	Numérica	% ocurrencias en vegetación natural
lc_agriculture	Numérica	% ocurrencias en áreas agrícolas
lc_water	Numérica	% ocurrencias en cuerpos de agua
pct_anp	Numérica	% ocurrencias dentro de ANP
anp_dist_km	Numérica	Distancia al ANP más cercano (km)
pct_ti	Numérica	% ocurrencias dentro de Territorio Indígena
ti_dist_km	Numérica	Distancia al TI más cercano (km)
pct_min_ilegal	Numérica	% ocurrencias en zona de minería ilegal
min_ilegal_dist_km	Numérica	Distancia a minería ilegal más cercana (km)
pct_concesion	Numérica	% ocurrencias en concesiones mineras
concesion_dist_km	Numérica	Distancia a concesión minera más cercana (km)
pct_petroleo	Numérica	% ocurrencias en bloques petroleros

Continúa en la siguiente página

Variable	Tipo	Descripción
petroleo_dist_km	Numérica	Distancia al bloque petrolero más cercano (km)
via_dist_km	Numérica	Distancia a vía nacional más cercana (km)
quemadas_freq	Numérica	Frecuencia media de quemadas en zona de ocurrencia
n_occ	Numérica	Número de registros de ocurrencia en GBIF
lat_centroid	Numérica	Latitud del centroide de distribución
lon_centroid	Numérica	Longitud del centroide de distribución
<i>Rasgos biológicos — Aves (AVONET):</i>		
Mass	Numérica	Masa corporal (g)
Beak.Length_Culmen	Numérica	Longitud del pico (mm)
Beak.Width	Numérica	Ancho del pico (mm)
Beak.Depth	Numérica	Profundidad del pico (mm)
Tarsus.Length	Numérica	Longitud del tarso (mm)
Wing.Length	Numérica	Longitud del ala (mm)
Kipps.Distance	Numérica	Índice de forma del ala
Habitat.Density	Numérica	Densidad del hábitat preferido
Range.Size	Numérica	Tamaño del rango geográfico (km ²)
Habitat_*	OHE	Tipo de hábitat (k categorías, k-1 columnas)
Migration_*	OHE	Patrón migratorio
Trophic.Level_*	OHE	Nivel trófico
Trophic.Niche_*	OHE	Nicho trófico detallado
Primary.Lifestyle_*	OHE	Estilo de vida principal
<i>Rasgos biológicos — Mamíferos (PanTHERIA):</i>		
body_mass_g	Numérica	Masa corporal adulta (g)
trophic_level	Numérica	Nivel trófico (1=herbívoro, 2=omnívoro, 3=carnívoro)
activity_cycle	Numérica	Ciclo de actividad (1=nocturno, 2=cathemeral, 3=diurno)
range_area_km2	Numérica	Área del rango geográfico (km ²)

Continúa en la siguiente página

Variable	Tipo	Descripción
threatened	Binaria	TARGET: 1=amenazada, 0=no amenazada

B. Rangos válidos de variables bioclimáticas WorldClim v2.1

Tabla 14: Rangos físicamente posibles empleados para limpieza de nodata.

Variable	Descripción	Mín.	Máx.
bio1_mean	Temperatura media anual $\times 10$	-600	400
bio4_mean	Estacionalidad de temperatura $\times 100$	0	23000
bio7_mean	Rango de temperatura anual $\times 10$	0	7200
bio12_mean	Precipitación anual (mm)	0	12000
bio14_mean	Precipitación mes más seco (mm)	0	3000
bio15_mean	Estacionalidad de precipitación (CV)	0	265

C. Resultados completos de validación cruzada

Las tablas 7 y 8 del cuerpo principal presentan los valores medios de ROC-AUC, PR-AUC y F1-macro. Las desviaciones estándar entre pliegues reflejan la variabilidad debida al tamaño reducido del conjunto (especialmente en mamíferos con $n=189$).

Para interpretar correctamente las métricas con desbalance severo:

- **ROC-AUC = 0,848** para aves indica que el modelo asigna probabilidad mayor a una especie amenazada que a una no amenazada en el 84,8 % de los pares.
- **PR-AUC = 0,192** para aves (vs. la línea base de 0,027 = prevalencia de la clase positiva) indica que el modelo tiene 7 veces más precisión que la asignación aleatoria para la clase amenazada.
- **F1-macro bajo** ($\approx 0,13$) refleja la dificultad de identificar correctamente la clase minoritaria incluso con pesos balanceados.

D. Predicciones completas para especies sin evaluación IUCN

Las siguientes tablas listan las **202 especies** (165 aves + 37 mamíferos) sin evaluación IUCN recuperada vía Wikidata, ordenadas por probabilidad de amenaza predicha (**prob_threatened**) de forma descendente. La columna **Pred.** indica la predicción binaria con **umbral adaptativo** (prevalencia del *train set*): 0,027 para aves y 0,069 para mamíferos. Son **31 aves y 18 mamíferos** los que obtienen **Pred.=1** (celdas sombreadas en las tablas). Los valores proceden del modelo Random Forest entrenado sobre el 100 % de los datos con etiqueta, con **class_weight='balanced'** y **n_estimators=300**.

*Aves (165 especies)***Tabla 15:** Predicciones completas de amenaza — Aves.

Espece	Prob.	Pred.
<i>Crax globulosa</i>	0.237	1
<i>Cinclodes aricomae</i>	0.230	1
<i>Charadrius alticola</i>	0.227	1
<i>Pauxi unicornis</i>	0.180	1
<i>Taoniscus nanus</i>	0.173	1
<i>Amazona xanthops</i>	0.140	1
<i>Hylophilus muscicapinus</i>	0.130	1
<i>Phibalura boliviana</i>	0.120	1
<i>Spizaetus isidori</i>	0.100	1
<i>Rollandia microptera</i>	0.087	1
<i>Nandayus nenday</i>	0.083	1
<i>Idiopsar dorsalis</i>	0.080	1
<i>Bucco tamatia</i>	0.073	1
<i>Upucerthia ruficaudus</i>	0.067	1
<i>Schizoeaca helleri</i>	0.060	1
<i>Asthenes sclateri</i>	0.060	1
<i>Leucophaeus pipixcan</i>	0.060	1
<i>Stilpnia meyerdeschauenseei</i>	0.053	1
<i>Harpyhaliaetus coronatus</i>	0.053	1
<i>Bucco macrodactylus</i>	0.050	1
<i>Porzana flaviventer</i>	0.047	1
<i>Xenopipo unicolor</i>	0.047	1
<i>Stilpnia cayana</i>	0.043	1
<i>Cyanocompsa brissonii</i>	0.043	1
<i>Agriornis andicola</i>	0.040	1
<i>Ramphocaenus sticturus</i>	0.037	1
<i>Cercomacra nigrescens</i>	0.037	1
<i>Sporophila maximiliani</i>	0.037	1
<i>Oressochen jubatus</i>	0.037	1
<i>Phyllomyias uropygialis</i>	0.033	1
<i>Chlorestes notata</i>	0.030	1
<i>Myiophobus ochraceiventris</i>	0.027	0
<i>Myrmeciza hyperythra</i>	0.027	0

Continúa

Especie	Prob.	Pred.
<i>Elaenia chilensis</i>	0.023	0
<i>Nystalus striatipectus</i>	0.023	0
<i>Serpophaga munda</i>	0.023	0
<i>Anthus chii</i>	0.020	0
<i>Tachycineta leucopyga</i>	0.017	0
<i>Cyanocompsa rothschildii</i>	0.017	0
<i>Himantopus mexicanus</i>	0.013	0
<i>Hyloctistes subulatus</i>	0.013	0
<i>Loriotus rufiventer</i>	0.013	0
<i>Saltatricula atricollis</i>	0.013	0
<i>Pyrocephalus obscurus</i>	0.013	0
<i>Oceanodroma markhami</i>	0.013	0
<i>Synallaxis propinqua</i>	0.013	0
<i>Chroicocephalus serranus</i>	0.010	0
<i>Leucopternis schistaceus</i>	0.010	0
<i>Leucophaeus atricilla</i>	0.010	0
<i>Gallinula melanops</i>	0.010	0
<i>Loriotus cristatus</i>	0.010	0
<i>Picoides fumigatus</i>	0.010	0
<i>Megascops roraimae</i>	0.010	0
<i>Terenura humeralis</i>	0.010	0
<i>Stilpnia viridicollis</i>	0.010	0
<i>Sturnella militaris</i>	0.010	0
<i>Cranioleuca gutturata</i>	0.010	0
<i>Grallaria sinaensis</i>	0.010	0
<i>Gymnopathys salvini</i>	0.007	0
<i>Geothlypis velata</i>	0.007	0
<i>Chlorothraupis frenata</i>	0.007	0
<i>Phalaropus tricolor</i>	0.007	0
<i>Xenops milleri</i>	0.007	0
<i>Percnostola lophotes</i>	0.007	0
<i>Aratinga aurea</i>	0.007	0
<i>Haplochelidon andecola</i>	0.007	0
<i>Dryocopus schulzii</i>	0.007	0
<i>Harpyhaliaetus solitarius</i>	0.007	0
<i>Himantopus melanurus</i>	0.007	0

Continúa

Especie	Prob.	Pred.
<i>Cercomacra serva</i>	0.007	0
<i>Charadrius collaris</i>	0.007	0
<i>Piranga lutea</i>	0.007	0
<i>Myrmeciza goeldii</i>	0.007	0
<i>Schistocichla humaythae</i>	0.007	0
<i>Schistocichla brunneiceps</i>	0.007	0
<i>Myrmeciza atrothorax</i>	0.003	0
<i>Leptasthenura yanacensis</i>	0.003	0
<i>Agelaioides oreopsar</i>	0.003	0
<i>Anthus brevirostris</i>	0.003	0
<i>Anurolimnas viridis</i>	0.003	0
<i>Nystalus obamai</i>	0.003	0
<i>Neochelidon tibialis</i>	0.003	0
<i>Myrmeciza fortis</i>	0.003	0
<i>Philydor ruficaudatum</i>	0.003	0
<i>Strix albitarsis</i>	0.003	0
<i>Thraupis episcopus</i>	0.003	0
<i>Trogon ramonianus</i>	0.003	0
<i>Strix huhula</i>	0.003	0
<i>Xiphorhynchus picus</i>	0.003	0
<i>Upucerthia certhioides</i>	0.003	0
<i>Pteroglossus mariae</i>	0.003	0
<i>Schistocichla leucostigma</i>	0.003	0
<i>Piaya minuta</i>	0.003	0
<i>Pipra rubrocapilla</i>	0.003	0
<i>Sclerurus obscurior</i>	0.003	0
<i>Stilpnia argyrofenges</i>	0.003	0
<i>Atticora melanoleuca</i>	0.003	0
<i>Buteo albicaudatus</i>	0.003	0
<i>Ixothraupis xanthogastra</i>	0.003	0
<i>Idiopsar speculifer</i>	0.003	0
<i>Claravis mondetoura</i>	0.003	0
<i>Leucopternis albigollis</i>	0.003	0
<i>Amazona mercenaria</i>	0.000	0
<i>Anurolimnas castaneiceps</i>	0.000	0
<i>Aramides cajanea</i>	0.000	0

Continúa

Especie	Prob.	Pred.
<i>Aratinga acuticaudata</i>	0.000	0
<i>Loriotus luctuosus</i>	0.000	0
<i>Hylophilus hypoxanthus</i>	0.000	0
<i>Idiopsar erythronotus</i>	0.000	0
<i>Ixothraupis punctata</i>	0.000	0
<i>Hylophilus ochraceiceps</i>	0.000	0
<i>Haplospiza rustica</i>	0.000	0
<i>Herpsilochmus frater</i>	0.000	0
<i>Hylocharis sapphirina</i>	0.000	0
<i>Grallaria cochabambae</i>	0.000	0
<i>Elliotomyia chionogaster</i>	0.000	0
<i>Empidonomus aurantioatrocris-</i> <i>tus</i>	0.000	0
<i>Chlorestes cyanus</i>	0.000	0
<i>Coccyzus cinereus</i>	0.000	0
<i>Chionomesa lactea</i>	0.000	0
<i>Colibri cyanotus</i>	0.000	0
<i>Dryocopus lineatus</i>	0.000	0
<i>Chionomesa fimbriata</i>	0.000	0
<i>Chalcothraupis ruficervix</i>	0.000	0
<i>Buteo polyosoma</i>	0.000	0
<i>Basileuterus punctipectus</i>	0.000	0
<i>Aratinga mitrata</i>	0.000	0
<i>Aratinga leucophthalma</i>	0.000	0
<i>Percnophierax leucorrhous</i>	0.000	0
<i>Penelope bridgesi</i>	0.000	0
<i>Ochthoeca pulchella</i>	0.000	0
<i>Orthopsittaca manilata</i>	0.000	0
<i>Ocreatus addae</i>	0.000	0
<i>Ochthoeca frontalis</i>	0.000	0
<i>Myrmeciza hemimelaena</i>	0.000	0
<i>Notiochelidon cyano-leuca</i>	0.000	0
<i>Notiochelidon flavipes</i>	0.000	0
<i>Notiochelidon murina</i>	0.000	0
<i>Stilpnia cyanicollis</i>	0.000	0
<i>Raueia bonariensis</i>	0.000	0

Continúa

Espece	Prob.	Pred.
<i>Saltatricula multicolor</i>	0.000	0
<i>Schizoeaca harterti</i>	0.000	0
<i>Pseudoscops clamator</i>	0.000	0
<i>Pseudocolaptes boissonneautii</i>	0.000	0
<i>Psarocolius oseryi</i>	0.000	0
<i>Pionopsitta barrabandi</i>	0.000	0
<i>Porphyrospiza alaudina</i>	0.000	0
<i>Porphyrio martinica</i>	0.000	0
<i>Pitangus lictor</i>	0.000	0
<i>Pipra chloromeros</i>	0.000	0
<i>Pteroglossus beauharnaesii</i>	0.000	0
<i>Phalacrocorax brasilianus</i>	0.000	0
<i>Phylloscartes ophthalmicus</i>	0.000	0
<i>Phylloscartes orbitalis</i>	0.000	0
<i>Xenops rutilans</i>	0.000	0
<i>Troglodytes musculus</i>	0.000	0
<i>Upucerthia harterti</i>	0.000	0
<i>Upucerthia andaecola</i>	0.000	0
<i>Tyto furcata</i>	0.000	0
<i>Thraupis palmarum</i>	0.000	0
<i>Thraupis sayaca</i>	0.000	0
<i>Sturnella supercilialis</i>	0.000	0
<i>Terenura sharpei</i>	0.000	0
<i>Stilpnia nigrocincta</i>	0.000	0
<i>Strix virgata</i>	0.000	0

Mamíferos (37 especies)

Tabla 16: Predicciones completas de amenaza — Mamíferos.

Espece	Prob.	Pred.
<i>Ovis aries</i> †	0.387	1
<i>Dasypterus ega</i>	0.283	1
<i>Equus caballus</i> †	0.267	1

Continúa

Especie	Prob.	Pred.
<i>Pteronura brasiliensis</i>	0.267	1
<i>Bos taurus</i> †	0.237	1
<i>Ateles chamek</i>	0.220	1
<i>Inia geoffrensis</i>	0.187	1
<i>Capra hircus</i> †	0.170	1
<i>Bubalus bubalis</i> †	0.143	1
<i>Catagonus wagneri</i>	0.117	1
<i>Plecturocebus modestus</i>	0.113	1
<i>Cyclopes catellus</i>	0.110	1
<i>Equus asinus</i> †	0.110	1
<i>Sylvilagus brasiliensis</i>	0.097	1
<i>Hydrochoeris hydrochaeris</i>	0.097	1
<i>Puma yagouaroundi</i>	0.073	1
<i>Leopardus braccatus</i>	0.073	1
<i>Lama glama</i> †	0.070	1
<i>Cavia porcellus</i>	0.057	0
<i>Natalus macrourus</i>	0.057	0
<i>Ateles belzebuth</i>	0.050	0
<i>Felis catus</i>	0.050	0
<i>Coendou longicaudatus</i>	0.040	0
<i>Carollia brevicaudum</i>	0.027	0
<i>Abrothrix andinus</i>	0.027	0
<i>Lonchophylla dekeyseri</i>	0.023	0
<i>Marmosops dorothea</i>	0.017	0
<i>Oryctolagus cuniculus</i>	0.013	0
<i>Vicugna pacos</i>	0.013	0
<i>Artibeus glaucus</i>	0.007	0
<i>Cyclopes rufus</i>	0.003	0
<i>Micoureus demerarae</i>	0.003	0
<i>Artibeus gnomus</i>	0.000	0
<i>Hsunnycteris thomasi</i>	0.000	0
<i>Eligmodontia hirtipes</i>	0.000	0
<i>Myotis midastactus</i>	0.000	0
<i>Lophostoma silvicola</i>	0.000	0