Modelo de predicción para el número de especies de Coleoptera en el Departamento de Antioquia

Esteban Marentes Herrera

Proyecto Aplicado para optar al título de
Magister en Ciencia de Datos

Director
Mario Julián Mora Cardona, MSc.

FACULTAD DE INGENIERÍA Y CIENCIAS
MAESTRÍA EN CIENCIA DE DATOS
SANTIAGO DE CALI, JUNIO DE 2025

TABLA DE CONTENIDO

LISTA DE FIGURAS

LISTA DE TABLAS

AGRADECIMIENTOS

Esta tesis fue concluida gracias al apoyo de muchas personas, desde la Universidad Javeriana Cali a los docentes Gloria Alvarez, David Arango y a mi director de tesis Mario Mora, que me guiaron a través del proceso con ideas y corrigiendo el camino cuando fue necesario. También agradezco a las entomólogas Jennifer Girón y Juliana Cardona, quienes me ofrecieron su ayuda desinteresada y facilitaron la revisión de los resultados con sus opiniones expertas y sugerencias desde la perspectiva biológica y a los entomólogos expertos Diego Martínez Reveló, Larry Jiménez Ferbans, Juan Pablo Botero, John César Neita y Claudia Alejandra Uribe que revisaron mis resultados de forma constructiva y desinteresada.

Agradezco a mi jefe Ricardo Ortiz, que estuvo siempre presente y ayudándome desde antes de comenzar esta maestría. A mis amigas y amigos del Instituto Humboldt que aportaron con sus sugerencias y revisiones, Camila Plata, Stephanie Larios, Geba Jisung, Mila Parra, Nerieth Leuro, Gonzalo Cabezas, Laura Sanchez, Erika Salazar, Sebastián Correa, Claudia Rozo y Erika Suarez por ayudarme a probar el código. También a los que aportaron con consejos y apoyo moral a lo largo del proceso, Claudia María Villa, Amalia Díaz, Leidy Vallejo, Paula Salinas, Carolina Castro, David Murillo, María Cecilia Londoño y Ming.

Un especial agradecimiento a todos los involucrados en hacer públicos los datos utilizados, por creer en la ciencia abierta y los beneficios que tiene. Todos los miembros de Coleoptera de Colombia que han publicado sus listas, los publicadores del SiB Colombia, a GBIF por facilitar la infraestructura para acceder a los registros biológicos y al IDEAM por disponer de las capas meteorológicas en su portal.

Finalmente agradezco a toda mi familia, en especial a Olga, Martha, Hermencia y Pablo, cuyo apoyo fue indispensable en todo este proceso y sin ellos no hubiera sido posible.

INTRODUCCIÓN

Colombia es uno de los cinco países con mayor biodiversidad del mundo, con apenas el 0.7% de la superficie continental mundial registra cerca del 10% de la biodiversidad del planeta. Sin embargo, el número de especies registradas para nuestro país sigue estando infravalorado, en especial para grupos poco estudiados y difíciles de muestrear e identificar como los coleópteros. Esto ocasiona que los esfuerzos de conservación y uso sostenible de la biodiversidad se vean impactados negativamente, al no tener datos que permitan comparar el número de especies muestreadas en la actualidad, con respecto al número total, lo que impide tomar decisiones de conservación e inversión en recursos de muestreo e investigación. Adicionalmente es inviable calcular este número directamente con muestreos o identificaciones en campo, debido a los altos costos y el personal especializado que se requiere. Este problema es especialmente relevante a nivel departamental, teniendo en cuenta que generalmente las listas de especies que se hacen para el grupo Coleoptera se hacen a nivel de país y no se tienen estimados claros para departamentos en específico.

Por este motivo es necesario recurrir a técnicas indirectas y modelos predictivos procedentes de la ciencia de datos, que utilizando información ya disponible permitan llenar este vacío a través de predicciones y tener un valor estimado para el número de especies de coleópteros en así poder tomar decisiones informadas para la conservación. El principal objetivo de este proyecto fue predecir el número de especies de coleópteros presentes en el Departamento de Antioquia, utilizando datos de factores abióticos y hábitats disponibles en el Instituto de Hidrología, Meteorología y Estudios Ambientales (IDEAM). El grupo de interés fueron los coleópteros debido a que corresponden a más del 40% del total de especies descritas de insectos y sirven de bioindicadores [1], además se eligió el Departamento de Antioquia dado que es el que presenta una mayor cantidad de datos y una geografía variada.

Para hacer esta estimación, se realizó una búsqueda bibliográfica para identificar variables climáticas y de hábitat que estén correlacionadas con el número de especies de coleópteros, junto a registros biológicos previamente publicados en el Sistema de Información de Biodiversidad sobre Colombia, que se utilizaron para entrenar diferentes modelos tipo regresión como un perceptrón multicapa, una red neuronal profunda y un random forest. Estos fueron evaluados utilizando métricas numéricas el error cuadrático medio (MSE), el error absoluto medio (MAE) y el coeficiente de determinación (R2) y evaluaciones indirectas de las estimaciones por parte de coleopterólogos expertos. Se realizó la evaluación de diversas técnicas y modelos de predicción de especies reportadas en la literatura científica reciente, priorizando las que tuvieron alguna implementación con código abierto e incluían como grupo focal los coleópteros o insectos, junto a algún tipo de validación de los resultados, estos modelos también contaron con variables abióticas y de

hábitat. Se utilizaron los datos ya disponibles en el SiB Colombia, que comprenden desde 1900 hasta febrero de 2025 y para los datos climáticos y de hábitat, se realizó una descarga desde la página del IDEAM, complementada con los mapas de hábitat disponibles a través de imágenes satelitales.

Luego de la revisión de literatura, se escogió una lista de ocho variables relevantes para la predicción y se utilizó una red neuronal profunda para las predicciones dado que tuvo el mejor rendimiento. Esta red se utilizó para estimar el número total de especies en Colombia y compararlo con la última lista publicada para el país, dando un resultado de 6.420 especies, un 4% superior a la lista. Adicionalmente estimó un promedio de 4.210 especies de coleópteros para el departamento de Antioquia y teniendo en cuenta la alta incertidumbre, también se estimó la mediana del número de especies 3.882, un valor mínimo 2.007 y valor máximo de 9.381. Esto permitió alcanzar un modelo robusto que se puede aplicar para diversos departamentos en el futuro.

1. DEFINICIÓN DEL PROBLEMA

1.1. PLANTEAMIENTO DEL PROBLEMA

Colombia es uno de los países más megadiversos del mundo debido a su posición geográfica, procesos orogénicos de los Andes y al ser una zona de transición en las Américas. Sin embargo, el número de especies presentes en nuestro país sigue estando infravalorado en especial para grupos poco estudiados y difíciles de muestrear e identificar como algunos insectos. Esto ocasiona que los esfuerzos de conservación y uso sostenible de la biodiversidad se vean impactados negativamente al no tener datos que permitan comparar el número de especies muestreadas, con respecto al estimado total. A pesar de que existen estimaciones para el número de especies de flora y fauna vertebrada de casi 200.000 [2], no existe un estimado para un grupo megadiverso como los insectos.

Es inviable calcular directamente este número con muestreos o identificaciones en campo debido a los costos y la falta de expertos en taxonomía, por este motivo es necesario recurrir a técnicas de predicción indirectas procedentes de la ciencia de datos, que utilizando datos ya disponibles permitan llenar este vacío de información y tomar decisiones informadas de conservación. A la fecha se encontraron 1.304.629 registros biológicos de insectos en el Sistema de Información sobre Biodiversidad de Colombia y 13.985 especies de insectos con al menos una observación [3], sin embargo, al no existir un estimado del número de especies totales para ese grupo es imposible determinar si la información es representativa o no y cuántas especies hacen falta por incluir en el sistema.

Dentro de los insectos, el grupo más diverso es el de los coleópteros que a nivel mundial tienen más de 385.000 especies descritas [4], con un número potencial de especies que puede llegar a estar entre 1´700.000 y 2´100.000. Sin embargo, no existe un estimado del número total de especies para el país, ni tampoco a nivel departamental para la toma de decisiones. Teniendo en cuenta que este problema, de falta de información se presenta para todos los departamentos del país y para varios grupos biológicos, se desarrolló una metodología escalable que utiliza los datos ya disponibles para predecir el número de especies de coleópteros en un departamento en particular.

1.2. FORMULACIÓN DEL PROBLEMA

¿Cuál es el número estimado de especies de coleópteros presentes en el Departamento de Antioquia a partir de datos ya disponibles?

¿Cómo se puede modelar el número de especies de coleópteros utilizando datos previamente publicados?
¿Qué variables de hábitat y climáticas están correlacionadas con el número de especies de coleópteros?
¿Cómo evaluar el desempeño de los modelos para poder seleccionar el más adecuado?

2. OBJETIVOS DEL PROYECTO

2.1 OBJETIVO GENERAL

Predecir el número de especies de coleópteros presentes en el Departamento de Antioquia a partir de datos abiertos de biodiversidad.

2.2 OBJETIVOS ESPECÍFICOS

- Modelar el número de especies de coleópteros usando técnicas de ciencia de datos y datos publicados previamente.
- Identificar variables climáticas y de hábitat que estén correlacionadas con el número de especies de coleópteros.
- Evaluar el desempeño de los diferentes modelos para seleccionar el que tenga un mejor desempeño

3. RESULTADOS ESPERADOS

Número estimado de especies de coleópteros para el Departamento de Antioquia, que permita tomar decisiones de conservación para proteger la biodiversidad.

Documento con la evaluación de diversas técnicas de ciencia de datos empleadas para la predicción del número de especies y la cantidad de datos necesaria para que cada técnica produzca resultados adecuados. Adicionalmente el documento contendrá una lista de las variables climáticas y de hábitat que fueron significativas para predecir el número de especies.

4. ALCANCE

Este proyecto se limitará a dar un estimado general del número de especies de coleópteros para el Departamento de Antioquia, sin entrar a profundizar sobre la distribución de esas especies, ni las familias a las que pertenecen. La implementación de esta metodología

para los otros Departamentos del país o grupos de insectos diferentes, quedará para próximos trabajos.

Para el proceso de exploración de las diferentes técnicas de ciencia de datos, se va a limitar a las que hayan sido más utilizadas en los últimos cinco años y tengan alguna implementación ya realizada en un lenguaje de programación, sin explorar técnicas nuevas que no han sido implementadas aún, o que son antiguas y se elaboraban con métodos desactualizados.

Para la identificación de variables climáticas y de hábitat se realizará una revisión de literatura identificando cuáles de estas afectan a los coleópteros y se escogerán las variables que aparezcan repetidas en más ocasiones y que tengan datos abiertos disponibles de forma libre, descartando variables sin datos o que no estén libres.

Para la evaluación de desempeño de los modelos no se realizará ningún tipo de verificación del resultado en campo, ni a través de una búsqueda de literatura de las especies disponibles en el Departamento, sino con los datos disponibles, estimaciones previamente realizadas y comentarios de expertos en el grupo para la región.

5. JUSTIFICACIÓN

Este proyecto es útil para determinar la cantidad de información faltante con respecto a los datos publicados y así incentivar la publicación de información para llenar estas deficiencias, a partir de datos que han sido tomados pero son privados o de recopilación adicional a través de muestreos en campo. Además, para movilizar recursos que permitan realizar más investigaciones en campo y estudios taxonómicos que completen el inventario de los coleópteros en el Departamento.

Es viable debido a que los datos de registros biológicos que se van a utilizar están disponibles de forma abierta a través del SiB Colombia y han sido aportados por diversos actores del sistema entre los que se encuentran Universidades, colecciones biológicas e Institutos de investigación, y que son reconocidos por sus altos estándares de calidad. Junto al uso de los datos abióticos que son recolectados por estaciones del IDEAM y están disponibles públicamente.

Adicionalmente, con el avance en las nuevas técnicas y métodos de la ciencia de datos como el aprendizaje profundo y el uso de datos etiquetados como los registros biológicos, es posible descubrir nuevos patrones y hacer estimaciones más acertadas que con métodos tradicionales.

Se cuentan con aproximadamente 28.000 registros para el orden de los coleópteros en el Departamento de Antioquia, además, es posible utilizar más de 180.000 registros en el resto del país para el entrenamiento del modelo, en caso de ser necesario.

El contar con un número de especies estimado, tiene un impacto positivo a nivel de gestión

de la biodiversidad, siendo un insumo para las autoridades ambientales en temas de decisión de licencias ambientales y resaltando la cantidad de especies aún por descubrir para incentivar la inversión en ciencias básicas para llenar vacíos de información. Finalmente, ayuda a divulgar la importancia de cuidar y conservar estos insectos entre el público general, mostrando la gran diversidad que tienen con respecto a otros grupos más visibles, pero menos diversos.

6. MARCO TEÓRICO Y ANTECEDENTES

6.1 MARCO TEÓRICO

El marco teórico aborda siete aspectos relevantes para esta investigación que son: coleópteros y su importancia, estimación del número de especies biológicas, importancia de las variables ambientales en las estimaciones, datos de registros biológicos y repositorios de datos abiertos de biodiversidad, metodología de modelado CRISP-DM, modelos comunes tipo regresión y rendimiento y evaluación de los modelos.

6.1.1 Coleópteros y su importancia

Los coleópteros son un orden de insectos conocidos con algunos nombres comunes como escarabajos, cucarrones, mariquitas o gorgojos, entre sus principales características está el primer par de alas delanteras rígidas llamadas élitros, su aparato bucal tipo masticador y que son holometábolos con un estadío larval.

Este grupo corresponde aproximadamente al 25% de todos los seres vivos del mundo y al 35% del total de insectos, con alrededor de 400.000 especies descritas en el mundo [5]. Están asociados con las formaciones vegetales donde actúan como depredadores, herbívoros, polinizadores o descomponedores de materia orgánica.

Teniendo en cuenta su gran riqueza y diversidad ecológica, constituyen buenos indicadores de la biodiversidad de un territorio [6] y son importantes para los humanos en diversos sectores, desde algunas especies comestibles, biocontroladores en agricultura, útiles en investigación forense y en medicina [1].

Es de destacar que desde el 2007 se dieron a conocer las primeras especies de coleópteros con algún riesgo de extinción, estas son en su mayoría endémicas y presentan un alto grado de vulnerabilidad [7]. Por todos estos motivos es indispensable conocer el número de especies, para poder tomar mejores decisiones de conservación.

6.1.2 Estimación del número de especies

La estimación del número de especies biológicas es un problema recurrente debido a la

dificultad para hacerlo de forma directa, por este motivo se han propuesto varios métodos para hacer este proceso a través de extrapolación, algunas de las técnicas más antiguas son la extrapolación por medio de curvas de acumulación, distribuciones paramétricas de abundancia relativa o técnicas no paramétricas con base en la distribución de individuos de una especies en particular [8]. Sin embargo, en los últimos años con el avance de las herramientas procedentes de la ciencia de datos, se han realizado modelos más complejos, por ejemplo utilizando redes neuronales profundas, que pueden aprender de interacciones complejas no lineales, como las que están involucradas en el proceso de distribución de especies y el número de especies en un lugar en particular, en especial al involucrar patrones espaciales [9].

6.1.3 Variables ambientales y su efecto en el número de especies

La distribución de los coleópteros está influenciada por algunas variables abióticas y ambientales, como la temperatura, humedad relativa, cantidad sombra, la estructura de la vegetación, entre otras. Estas afectan las abundancias y distribuciones de cada especie de distinta manera e influyen en la cantidad total de especies que existen en un lugar particular, con algunas especies dentro del mismo género teniendo roles ecológicos distintos y respondiendo a diferentes variables ambientales [10].

Por este motivo, la selección de las variables que serán utilizadas para la construcción de los modelos es de vital importancia. Aunque normalmente los modelos de distribución se han limitado al uso de datos climáticos, la influencia del tipo de suelo y el tipo de cobertura también juegan un papel importante en la distribución y número de especies. En el estudio realizado por Chauvier *et al,* se determinó que los patrones espaciales de distribución de plantas estaban influenciados mayormente por las variables climáticas con un 58%, la cobertura con un 21% y el suelo con un 20% [11].

6.1.4 Datos de registros biológicos y repositorios de datos abiertos de biodiversidad

Los registros biológicos son datos primarios de biodiversidad que tienen un nivel de detalle suficiente acerca de la ubicación de un individuo en el tiempo y el espacio, es decir, ofrecen evidencia de la presencia de una especie (u otro taxón) en un lugar y fecha determinados, en ocasiones pueden tener información adicional sobre el método de muestreo, las personas o instituciones que tomaron o custodian los datos. Estos registros son utilizados en una gran variedad de escenarios, desde estudios de biogeografía, evaluación del impacto del cambio climático, planeación ambiental, entre otros [12].

Esta información de registros está disponible de forma gratis y abierta a través del Sistema Global de Información sobre Biodiversidad (GBIF por sus siglas en inglés), una iniciativa global fundada en 2001 con sede en Copenhague que facilita la publicación de datos biológicos a través de diversas herramientas como Herramienta Integrada de Publicación (IPT por sus siglas en inglés), utilizando el estándar DarwinCore para que sean interoperables en todo el mundo [13]. Este sistema funciona como una red con nodos a

nivel de país, que gestionan la publicación de datos. El nodo para el país es el Sistema de Información sobre Biodiversidad de Colombia (SiB Colombia), una iniciativa que nació con el decreto 1603 de 1994 como parte del Sistema Nacional Ambiental y tiene como principal objetivo facilitar la gestión de datos e información sobre biodiversidad para apoyar procesos de investigación, educación o toma de decisiones [14].

6.1.5 Metodología de modelado CRISP-DM

En la mayoría de los proyectos de ciencia de datos se utiliza alguna metodología para asegurar el buen resultado de este, una de estas metodologías es Cross-Industry Standard Process-Data Mining, más conocida como CRISP-DM por sus siglas en inglés [15], que es un proceso estándar utilizado para minería de datos en diversas industrias. Consiste en seis pasos principales: entendimiento del negocio, entendimiento de los datos, preparación de los datos, modelado, evaluación y despliegue como se puede ver en la figura 1.

Este es un proceso iterativo flexible, que permite enfocarse en algún aspecto en particular en caso de ser necesario siguiendo una estructura lógica. Se comienza haciendo una revisión del negocio para entender los objetivos del modelado y cuáles son los resultados esperados, seguido de una recopilación y exploración inicial del tipo de datos que se tienen disponibles, usando técnicas sencillas de visualización y estadísticas. La preparación de los datos implica realizar ajustes como fusión de datos, agregación de registros, adición de nuevos atributos, ajustes de valores faltantes, codificación de variables categóricas, división de los datos para prueba y entrenamiento y en general las modificaciones que sean necesarias para dejar los datos listos para el modelado, en general esta fase puede tomar entre 50-70% del tiempo de un proyecto. En la fase de modelado se prueban diversos modelos comenzando por los más sencillos con parámetros predeterminados y realizando ajustes de forma iterativa para mejorar los resultados. De forma paralela, se va realizando la evaluación de los modelos, para verificar que si están respondiendo la pregunta del negocio y que tienen resultados positivos y replicables. Para finalizar se realiza el despliegue, que consiste en la puesta en marcha del modelo en un ambiente real, junto con la documentación final.

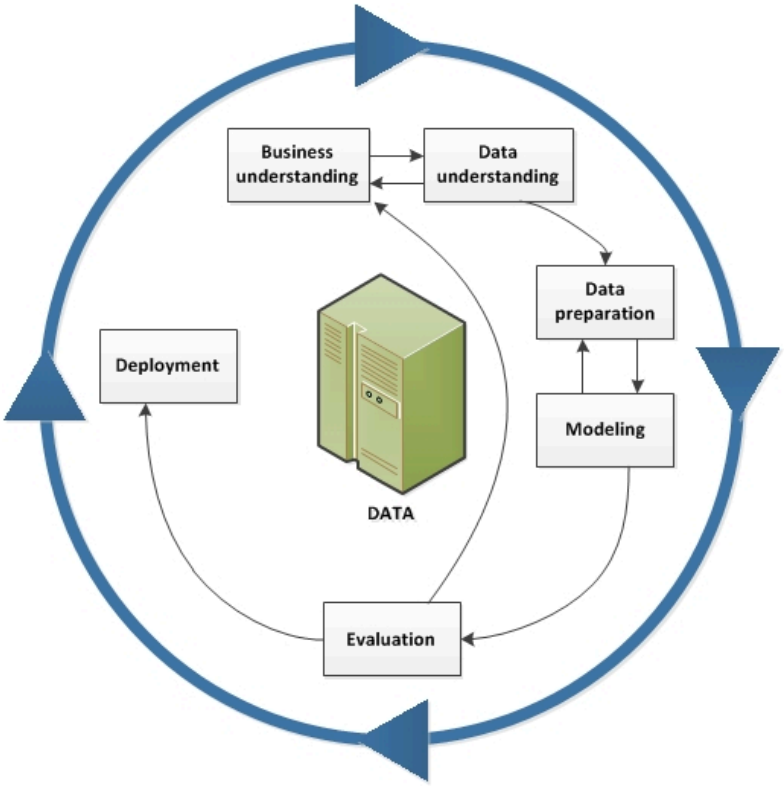

Figura 1. Ciclo de vida de minería de datos con la metodología CRISP-DM. Fuente: [16]

6.1.6 Modelos comunes en ciencia de datos para tareas de tipo regresión

Los modelos de regresión están interesados en inferir una función matemática cuyos valores correspondan al promedio de la variable dependiente o respuesta, condicionados a una o más variables independientes o de entrada. A lo largo de los años, se han creado una gran variedad de técnicas para realizar esta tarea, incluyendo métodos paramétricos, semi paramétricos y no paramétricos [17]. Algunas de las técnicas comunes para estas tareas son los modelos de regresión lineal, regresión polinomial, regresión logística, árboles de decisión, random forest, máquinas de soporte vectorial, redes neuronales, modelos bayesianos, entre otros [18].

Entre las técnicas utilizadas en este trabajo se encuentran los Random Forest, que son una combinación de predictores tipo árbol en donde cada árbol depende del valor de un vector aleatorio muestreado de forma independiente y que tiene la misma distribución para todos los árboles en el bosque [19]. A pesar de que originalmente fueron creados para tareas de clasificación, es posible utilizarlo para predecir valores numéricos continuos, promediando el resultado de múltiples árboles para reducir la varianza.

El perceptrón multicapa es un tipo de red neuronal tipo feedforward para aprendizaje supervisado, cuyas unidades tienen funciones de activación polinomiales combinando adiciones y multiplicaciones de polinomios tipo Kolmogorov-Gabor [20]. Están

compuestos de una capa de entrada, una o más capas ocultas de neuronas artificiales tipo TLU (Threshold Logic Units) y una capa final de salida, como muestra la figura 2.

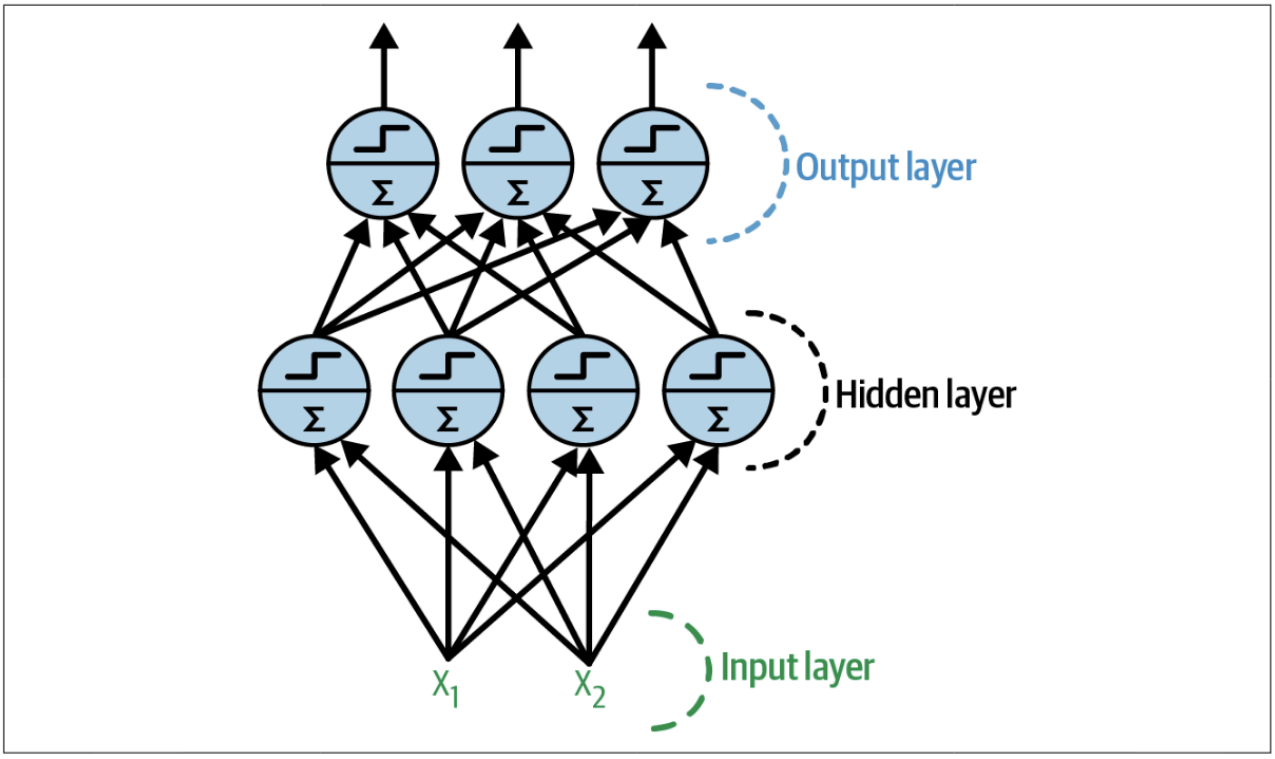

Figura 2. Arquitectura de un perceptrón multicapa. Fuente: [18]

Las redes neuronales profundas son redes con más de una capa oculta, como se puede ver en la figura 3. Son capaces de aprender características de alto nivel con más complejidad y abstracción que redes poco profundas, aunque requieren recursos adicionales de cómputo y son más difíciles de entrenar, pero dan mejores resultados si los conjuntos de datos tienen suficientes datos [21]. Para su entrenamiento se determinan valores y pesos en la red, utilizando un proceso de optimización de gradiente descendente a partir de backpropagation.

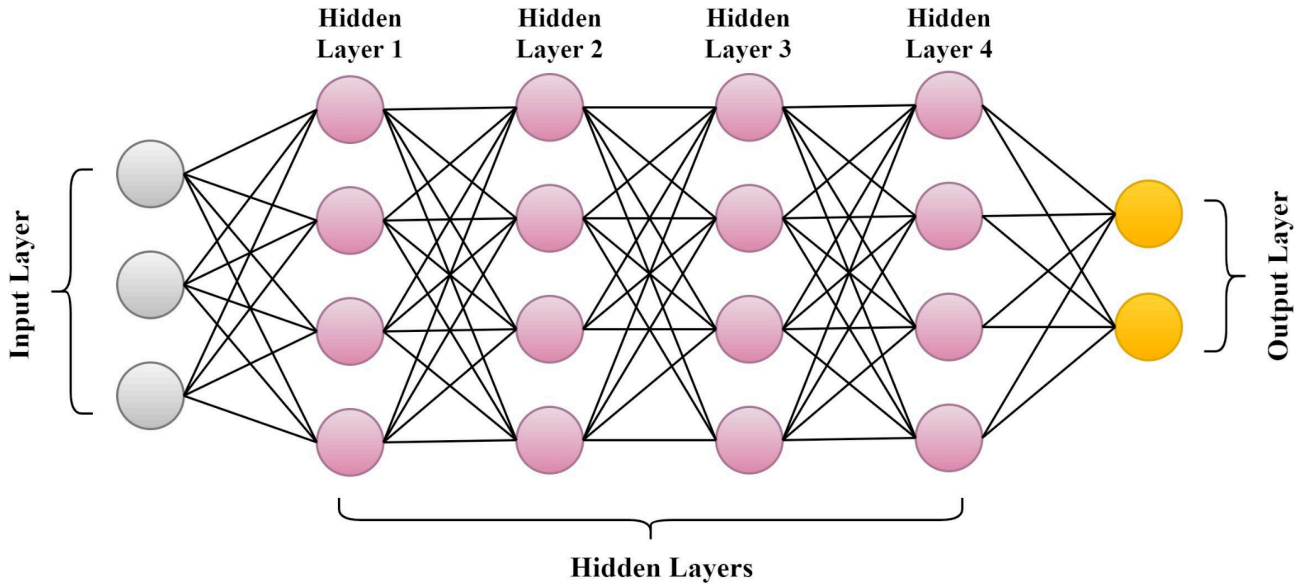

Figura 3. Arquitectura de una red neuronal profunda. Fuente: [22]

6.1.7 Evaluación del rendimiento de los modelos

En todos los procesos de modelado, es indispensable realizar una evaluación del rendimiento de los modelos resultantes, para revisar que se están comportando como deberían, mejorar los resultados y en general comprobar su validez.

En los análisis de regresión se utiliza una amplia variedad de métricas, entre las que se destacan el error cuadrático medio (MSE por sus siglas en inglés), el error absoluto medio (MAE por sus siglas en inglés) y el coeficiente de determinación (R^2), siento este último uno de los más utilizados al ser intuitivo y eficaz [23].

Las dos métricas MAE y MSE hacen parte de una familia que evalúa la calidad del ajuste en términos de distancia del valor predicho por el modelo al punto actual de entrenamiento, donde el primero calcula el promedio del valor absoluto de los errores, mientras que el segundo evalúa el cuadrado de los errores [17]. El MAE es más sencillo de interpretar al estar en las mismas unidades que la variable y no es tan sensible a los outliers, mientras que el MSE penaliza los errores grandes y es sensible a los outliers pero su interpretación no es tan simple al estar en el cuadrado de las unidades. En ambos casos se busca un valor bajo para estas métricas y depende de la escala en la que esté la variable dependiente para interpretar si un valor es elevado o no.

$$MAE = \frac{1}{m}\sum_{i=1}^{m}\left|X_i - Y_i\right|$$

$$MSE = \frac{1}{m}\sum_{i=1}^{m}\left(X_i - Y_i\right)^2$$

El coeficiente de determinación R^2 fue originalmente propuesto por Wright en 1921 y cuantifica qué tanto la variable dependiente está determinada por las variables independientes en términos de proporción de la varianza y está dado por la ecuación 1 - la relación entre la suma de los residuales al cuadrado y la suma total de los cuadrados. Este puede tomar un valor entre 0 a 1, con un valor cercano a 1 indica que el modelo explica un porcentaje elevado de la varianza [24].

$$R^2 = 1 - \frac{\sum_{i=1}^{m}\left(X_i - Y_i\right)^2}{\sum_{i=1}^{m}\left(\overline{Y}_i - Y_i\right)^2}$$

En escenarios donde hay una alta incertidumbre o no hay suficientes datos para validar los modelos de forma empírica, es posible recurrir a la opinión de expertos en el tema siguiendo un proceso de elicitación experta, que permite obtener afirmaciones de corte probabilístico basadas en las creencias de los expertos sobre alguna cantidad o parámetro desconocido [25]. Para este proceso es necesario definir las preguntas objetivo de forma cuidadosa, probar el protocolo de elicitación, entrenar a los expertos en probabilidad subjetiva, llevar a cabo entrevistas, darle retroalimentación a los expertos y analizar y documentar los resultados.

6.2 ANTECEDENTES

En el trabajo realizado por Brunn *et al* [26] se encontró el uso de redes neuronales profundas con el objetivo de predecir modelos para las comunidades de especies de plantas utilizando datos de registros biológicos derivados de ciencia ciudadana, con el objetivo de evaluar la variabilidad fenológica, predecir el potencial de especies dominantes y predecir el impacto del cambio climático. Se entrenaron distintas versiones de modelos multiespecies, usando dos funciones de costo para evaluar el resultado y varios conjuntos de datos predictores con 24 variables ambientales. A pesar de que no se intentó predecir el número total de especies y se utilizó un grupo de estudio diferente, el uso del método de redes usando registros biológicos y variables climáticas, presenta un excelente antecedente y permite la aplicación de los métodos utilizados tanto para la limpieza de datos, como para la construcción de un modelo cuyo objetivo sea predecir el número de especies.

El estudio de Alfaro y Pizarro [27], estimó un número teórico de especies de coleópteros epigeos en islas de la Reserva Nacional Pingüino de Humboldt usando curvas de acumulación y estimadores no paramétricos con muestreos realizados en diferentes años. Un modelo de dependencia lineal estimó un número teórico de especies cercano al real observado, en dos de las islas en estudio, Choros y Damas, por su parte, el modelo de Clench estimó un mayor número de especies. Todos los estimadores no paramétricos mostraron valores de riqueza estimada por encima de la riqueza observada, en todas las islas. El estudio realizó la revisión del mismo grupo biológico propuesto en este proyecto, sin embargo, está limitado a un área geográfica insular y se basa en datos con información de intensidad de muestreo y acumulación de especies, que no están disponibles en todos los casos.

El estudio de Hong *et al* [28], estimó la riqueza de especies microbianas para una muestra de sedimentos marinos, esto se hizo a través de un enfoque comparativo que empleó varias herramientas paramétricas y no paramétricas, incluidas la mayoría de métodos utilizados hasta la fecha y también dos que son nuevos para la investigación de la biodiversidad (las distribuciones de Poisson mixtas de Pareto y mixtas de exponenciales). Se analizó el rendimiento de todos los modelos y se eligió el de mejor rendimiento en función de (i) pruebas de bondad de ajuste, (ii) capacidad de producir un error estándar biológicamente significativo, y (iii) uso de la cantidad máxima de datos empíricos de frecuencia de especie. Como resultado general se identificó que la aplicación indiscriminada de un único modelo a cualquier conjunto de datos conduce a resultados erróneos. Este estudio tiene en común el objetivo de predecir el número de especies para un grupo megadiverso, aunque utilizando un enfoque distinto más enfocado en los datos moleculares.

Los antecedentes revisados durante este proyecto ofrecieron puntos de partida a nivel conceptual y metodológico, sin embargo ninguno abordó el mismo objetivo de estimar el número de coleópteros en el departamento de Antioquia, ni usaron métodos de ciencia de datos junto a listas de chequeo curadas por expertos y publicadas de forma abierta.

7. IDENTIFICACIÓN DE VARIABLES CLIMÁTICAS RELACIONADAS CON EL NÚMERO DE ESPECIES DE COLEÓPTEROS Y OBTENCIÓN DE DATOS

Para la identificación de las variables climáticas y de hábitat que tienen un impacto sobre la distribución de los coleópteros, se realizó una revisión sistemática de literatura en buscadores académicos como Google Scholar, utilizando los siguientes criterios:

- Antigüedad menor a 15 años.
- Artículos en inglés y español.
- Palabras clave de búsqueda: medidas abióticas distribución Coleoptera, hábitat y distribución Coleoptera, abiotic measurements distribution Coleoptera, habitat and

distribution Coleoptera.
- Solamente se incluyeron artículos en revistas científicas, tesis de maestría o doctorado y libros.

Posterior a la identificación de las variables, se realizó una búsqueda de datos para estas variables en Colombia que estuvieran disponibles de forma abierta en repositorios nacionales e internacionales, junto a los registros biológicos de Coleoptera.

7.1. VARIABLES CLIMÁTICAS Y DE HÁBITAT RELACIONADAS CON LOS COLEÓPTEROS

En el estudio realizado por Jiménez-Ferbans *et al* [29] para la familia Passalidae en la región Caribe, se identificaron 18 especies en cuatro departamentos, además registraron las formaciones vegetales, la elevación y la humedad relativa como los factores más influyentes en el número de especies, donde las zonas secas y en elevaciones menores tienen una menor cantidad de especies.

En el estudio realizado por Obregón-Corredor *et al* [30] se identificaron los factores climáticos para la dinámica de artrópodos en cultivos de arroz en Yopal, evaluando su importancia en la abundancia y número de especies. Se realizó un análisis de Random Forest para estimar la relevancia de cada variable climática en la abundancia de insectos, estos valores se calcularon por grupo funcional de artrópodo, en vez de por categoría taxonómica, por eso se tomaron las categorías depredadores y masticadores de follaje que presuntamente tienen la mayor cantidad de coleópteros. Para estos dos grupos, las variables temperatura mínima promedio, temperatura mínima y radiación solar tuvieron los valores más altos en la variable climática, que fue la más discriminante en el análisis.

En el estudio realizado por Rubio y Lobo [31] se evaluaron las principales condiciones microclimáticas para explicar la distribución de las 17 especies del género Eurysternus en Colombia. Para la selección de las variables se realizó un análisis de factor ecológico de nicho (ENFA por sus siglas en inglés) y se encontró que las más relevantes son la temperatura media anual seleccionada en el 94% de las especies, la precipitación anual y temperatura promedio diaria que ambas aparecen en el 41% de las especies, en conjunto con la elevación, registrando que el 60% de las especies están presentes por debajo de los 1500 metros sobre el nivel del mar.

En el estudio de Cheng *et al* [32] se realizó la identificación de las variables climáticas que más afectan a dos especies de mariquitas en agroecosistemas del norte de China. Para esto se analizaron las tendencias temporales a lo largo de 28 años utilizando tres modelos de aprendizaje de máquina: máquinas de soporte vectorial, random forest y perceptrón multicapa para relacionar los registros biológicos con diez variables medioambientales

seleccionadas. Las máquinas de soporte vectorial tuvieron el mejor rendimiento de predicción y las variables con mayor importancia relativa fueron la temperatura promedio (0.33), la humedad relativa promedio (0.32) y la velocidad promedio del viento (0.24).

En el estudio de Brasil *et al* [33] se realizó la evaluación de los principales factores para predecir la riqueza de especies en el orden Zygoptera en el Amazonas brasileño. Para esto se realizaron muestreos en 100 afluentes y se combinaron con variables abióticas en diferentes modelos lineales generalizados mixtos. Posteriormente seleccionaron el mejor modelo, utilizando como métrica de selección el pseudo R^2, que fue de 0.396 para el modelo con la productividad primaria neta, la precipitación y la temperatura. A pesar de ser un orden diferente a Coleoptera, los zigópteros tienen un estadio larval acuático y están relacionados a los cuerpos de agua, al igual que 17 familias de coleópteros acuáticos en el país [34], por lo que las mismas variables pueden llegar a ser relevantes para ambos grupos.

En el estudio de Gomes-Gonçalves [35] se realizó la evaluación de los factores meteorológicos que afectan la dinámica poblacional de los cucarrones en la familia Curculionidae en dos zonas de Brasil. Para esto se realizaron colectas en campo y toma de medidas meteorológicas radiación solar, humedad relativa, precipitación y temperatura media. Luego se realizó un análisis de correlaciones canónicas para identificar las que tenían un mayor efecto sobre las dos subfamilias. Las variables más importantes fueron radiación solar y temperatura con -0.87 y precipitación con 0.82.

En el estudio de Lobo *et al* [36] se realizó un proceso para modelar la riqueza de especies de la familia Scarabaeidae en Francia, utilizando modelos lineares generalizados para relacionar el número de especies presentes con variables climáticas, topográficas y espaciales. Para esto se dividió el territorio en 66 grillas cuadradas que habían sido previamente bien muestreadas y se probaron distintos modelos que incluyeran la relación entre las variables, la interacción de términos y la información de longitud y latitud, el modelo final se validó con un método jackknife. El modelo final predice el 86.2% de la varianza, con el 38% correspondiente al efecto combinado de los tres tipos de variables, 21% para la estructura climática espacial, 14% para las variables climáticas individuales y 11% para el componente espacial individualmente. Las únicas variables que tenían un efecto estadísticamente significativo por sí solas, fueron la temperatura mínima promedio y la temperatura máxima promedio.

De acuerdo con las recomendaciones proporcionadas por la entomóloga Juliana Cardona Duque [37], se sugiere incorporar la información relativa a las especies vegetales asociadas a los coleópteros. Esta consideración se fundamenta en el hecho de que numerosas especies de coleópteros mantienen relaciones simbióticas con una o más especies de plantas, las cuales son esenciales para su alimentación o participación en

procesos de polinización.

En el estudio de Pasek [38] se realizó una revisión general del efecto del viento en la dispersión de los insectos, identificando que aunque variaba con la especie, normalmente la frecuencia de vuelo aumenta cuando la velocidad del viento está entre 0.5 a 2 m s^{-1}. Algunos insectos pequeños o sin alas, dependen de las corrientes de aire para ser transportados a sitios nuevos y sus patrones de distribución frecuentemente reflejan los patrones del flujo del viento. Adicionalmente, los insectos voladores normalmente aterrizan en sitios con baja velocidad del viento, como la sombra de plantas o árboles. Esto se ve especialmente reflejado en la mayor presencia de insectos plaga en donde hay cortavientos de vegetación, que dependen del hábitat que afecta la composición de especies vegetales y el ángulo de incidencia del viento en los diferentes ecosistemas.

En la tabla 1 se muestra la información compilada de las variables identificadas por estudio.

Tabla 1. Variables identificadas como relevantes para el número de especies de Coleoptera y el estudio donde se identificaron.

Variables	Estudio [29]	Estudio [30]	Estudio [31]	Estudio [32]	Estudio [33]	Estudio [35]	Estudio [36]	Estudio [37]	Estudio [38]
Elevación	1		1						
Hábitat	1								1
Humedad relativa	1	1		1					
Temperatura mínima		1					1		
Temperatura media			1	1	1	1			
Temperatura máxima							1		
Radiación solar		1				1			
Precipitación			1		1	1			
Velocidad del viento				1					1
Especies de plantas								1	1
Productividad Primaria Neta					1				

Teniendo en cuenta la información recopilada, se seleccionaron las variables que aparecen de forma más frecuente y con un mayor peso en los modelos realizados, estas fueron:

- Velocidad del viento
- Humedad relativa
- Precipitación
- Radiación solar
- Temperatura
- Ecosistemas/hábitat
- Elevación sobre el nivel del mar
- Especies de plantas cercanas a los coleópteros

7.2 OBTENCIÓN DE DATOS ABIERTOS PARA LAS VARIABLES IDENTIFICADAS Y REGISTROS DE COLEÓPTEROS

Luego de tener las variables más relevantes identificadas, se realizó una búsqueda de los datos abiertos disponibles en repositorios nacionales, en especial en el portal del IDEAM (https://visualizador.ideam.gov.co/CatalogoObjetos/). Las capas de información encontradas para estas variables fueron:

1. Velocidad del viento mensual a 10 metros de altura de 2000 a 2010. Los datos son un archivo tipo ESRI shapefile con cobertura para Colombia, clasificando la velocidad en 12 rangos significativos que oscilan entre 0 hasta superior a 11 m/s [39].
2. Mapa de ecosistemas continentales y marinos-costeros de Colombia a escala 1:100.000. Los datos son un archivo tipo ESRI shapefile con cobertura para Colombia y clasifican todos los ecosistemas del país en un total de 48 categorías, adicionalmente cada uno de los polígonos tiene información adicional como tipo de clima, relieve, suelo, unidad biótica y número de anfibios, reptiles, mamíferos, aves y magnólias, presentes en el ecosistema [40].
3. Humedad relativa anual promedio multianual desde 1981 hasta 2010. Los datos son un archivo tipo ESRI shapefile con cobertura para Colombia, clasificando la humedad en 6 rangos significativos que oscilan entre 65 y 95 por ciento (%) [41].
4. Precipitación para Colombia (mm) 1976-2005. Escala 1:100,000. Los datos son un archivo raster tipo TIFF con un tamaño de celda (X,Y): 0.009, 0.009 Radianes por cada grado sexagesimal, clasificando la precipitación en 6 rangos significativos medidos en mm [42].
5. Radiación solar global promedio 2005. Los datos son un archivo tipo ESRI shapefile con cobertura para Colombia, correspondientes al valor agregado de los kWh que en

promedio inciden durante el día sobre un metro cuadrado, expresado en KWh/m^2 [43].

6. Temperatura mínima mensual promedio durante el periodo 1981-2010. Los datos son un archivo tipo ESRI shapefile con cobertura para Colombia, clasificando la temperatura en 8 rangos significativos que oscilan entre inferior a 8 y superior a 24C° [44].
7. Temperatura media mensual promedio durante el periodo 1981-2010. Los datos son un archivo tipo ESRI shapefile con cobertura para Colombia, clasificando la temperatura en 9 rangos significativos que oscilan entre inferior a 8 y superior a 28C° [45].
8. Temperatura máxima mensual promedio durante el periodo 1981-2010. Los datos son un archivo tipo ESRI shapefile con cobertura para Colombia, clasificando la temperatura en 9 rangos significativos que oscilan entre inferior a 8 y superior a 34C° [46].
9. Elevación NASA. Los datos son un archivo raster tipo TIFF con un tamaño de celda (X,Y) de 1 arco de segundo, dando un valor de elevación sobre el nivel del mar medido en metros usando el geoide EGM96 como referencia [47].
10. Adicionalmente a estas variables identificadas, se utilizó el mapa oficial de departamentos de Colombia para realizar los cruces geográficos [48].

La información de las especies de plantas asociadas a los coleópteros, se obtuvo descargando todos los registros biológicos del reino Plantae para Colombia directamente de GBIF con corte a febrero del 2025 [49]. Este conjunto de datos está en formato tabla, tiene 5’653.070 de filas y 50 columnas, que tienen información de la taxonomía, localidad, coordenadas, fecha, entidad publicadora, colector y tipo de registro.

Los datos referentes al orden Orden Coleoptera fueron descargados directamente de GBIF con corte a marzo de 2025 [50], utilizando el filtro para el orden Coleoptera en todo el mundo. Este conjunto de datos está en formato tabla, tiene 32’356.574 de filas y 50 columnas, que tienen información de la taxonomía, localidad, coordenadas, fecha, entidad publicadora, colector y tipo de registro.

Finalmente, se compiló la información del número de especies para familias de Coleoptera en Colombia, esta se obtuvo utilizando las listas de referencia de familias disponibles para el país publicadas por Coleoptera de Colombia, estas son:

Subfamilia Brachycerinae [51], subfamilia Bruchinae [52], familia Anamorphidae [53], familia Dytiscidae [54], familia Megalopodidae [55], familia Lucanidae [56], familia Gyrinidae [57], subfamilia Eumolpinae [58], familia Hydraenidae [59], familia Tenebrionidae [60], familia Elmidae [61], familia Buprestidae [62], Subfamilia Dryophthorinae [63], familia Ptilodactylidae [64], Subfamilia Conoderinae [65], familia Meloidae [66], familia

Cneoglossidae [67], familia Dryopidae [68], familia Chelonariidae [69], familia Lutrochidae [70], familia Mordellidae [71], familia Limnichidae [72], familia Georissidae [73], familia Anthribidae [74], familia Hydrochidae [75], familia Brentidae [76], Subfamilia Cyclominae [77], familia Dermestidae [78], familia Callirhipidae [79], familia Noteridae [80], familia Psephenidae [81], subfamilia Criocerinae [82], familia Anthicidae [83], Scirtidae [84], familia Melandryidae [85], familia Archeocrypticidae [86], familia Aderidae [87], subfamilia Ceutorhynchinae [88], familia Lycidae [89], familia Phengodidae [90], subfamilia Orphninae [91], familia Epimetopidae [92], familia Lampyridae [93], familia Haliplidae [94], familia Cantharidae [95], familia Hydrophilidae [96], subfamilia Cossoninae [97], familia Heteroceridae [98], familia Ochodaeidae [99], familia Leiodidae [100], familia Disteniidae [101] y la subfamilia Entiminae [102].

Para esto se realizó la descarga de las listas publicadas a través del SiB Colombia [103] y se realizó el proceso manual de agregar una nueva columna al conjunto de datos con el número de especies que había por cada familia o subfamilia, este conjunto de datos está en formato tabla, tiene 52 de filas y 18 columnas.

8. PROCESAMIENTO DE DATOS Y CONSTRUCCIÓN DE LOS MODELOS PARA PREDECIR EL NÚMERO DE ESPECIES DE COLEÓPTEROS EN ANTIOQUIA

Para el diseño del modelo, se realizó una exploración de técnicas ya utilizadas para este fin, revisando repositorios institucionales y buscadores académicos como Google Scholar, empleando los siguientes criterios:

- Antigüedad menor a 5 años.
- Artículos en inglés y español.
- Palabras clave de búsqueda: modelos de predicción, predicción especies, número especies coleópteros, model prediction, predicting species, number species coleoptera.
- Solamente se incluyeron artículos en revistas científicas, tesis de maestría o doctorado y libros.

Las técnicas encontradas se presentan en la tabla 2, que muestra los principales datos utilizados por el modelo, el software utilizado, el grupo biológico estudiado y la referencia.

Tabla 2. Modelos utilizados para la predicción del número de especies encontrados en la búsqueda de literatura

Tipo de técnica	Datos utilizados	Software	Grupo biológico	Métrica evaluación	Referencia
Regresión	Imágenes	GLIM	Coleoptera	Varianza	[104]

logística	satelitales. Datos de ausencia y presencia de Coleoptera por grillas			explicada D^2	
Redes neuronales profundas. Regresión con Random Forest.	Inventarios de plantas en campo. Imágenes satelitales.	Python	Flora	R^2	[105]
Ensamblaje de pequeños modelos.	Registros biológicos plantas. Mapas de variables climáticas, geográficas y topológicas.	R	Flora	AUC	[106]
Redes neuronales profundas.	Registros biológicos plantas. Mapas de variables climáticas.	R	Flora	Normalized Discounted Cumulative Gain(NDCG)	[26]
Redes neuronales profundas	Datos de sensores remotos	Python	General	RMSE	[107]

En la mayoría de estos artículos, los modelos más usados y con mejor resultado fueron los de redes neuronales profundas y aunque no se reportó ningún estudio donde se intente predecir el número de especies de coleópteros, se decidió utilizar estos debido a que presentaban mejores resultados en las métricas de evaluación como R^2 y MAE.

Para realizar este proceso de modelado, se siguió la metodología CRISP-DM, la correspondencia de esta con las secciones del documento se puede ver en la tabla 3. Es importante resaltar que estas fases se ejecutaron de forma iterativa a medida que avanzó el proyecto.

Tabla 3. Etapas de la metodología CRISP-DM mapeadas en las secciones del documento.

Etapa metodología CRISP-DM	Sección del documento
Entendimiento del negocio	2. Objetivos del proyecto y 3. Resultados esperados
Entendimiento de los datos	7.2 Obtención de datos abiertos
Preparación de los datos	8.1 Preparación de los datos
Modelado	8.2 Construcción de los modelos
Evaluación	9.1 Evaluación con métricas numéricas y 9.3 Evaluación por parte de expertos de los resultados del modelo
Despliegue	9.2 Predicción del número de especies de coleópteros para Antioquia y Colombia

8.1 PREPARACIÓN DE LOS DATOS

Se realizó el procesamiento de los datos utilizando el software *Python* en un ambiente local utilizando la *IDE Spyder 6.0.3*, como primer paso se realizó la carga de la información de Coleoptera en el mundo, a través de lotes de 100.000 registros debido a su tamaño. Se eliminaron algunas variables ('datasetKey', 'publishingOrgKey', 'day', 'month', 'year', 'taxonKey', 'institutionCode', 'catalogNumber', 'recordNumber', 'rightsHolder', 'typeStatus', 'establishmentMeans', 'lastInterpreted', 'mediaType') debido a que no tienen relevancia biológica y se eliminaron los registros que no tenían coordenadas.

Este primer dataframe se transformó utilizando geopandas para convertir los datos a un gdp tipo punto y poder hacer los cruces con las capas tipo shapefile, comenzando con los departamentos del país para dejar solamente la información de los coleópteros en Colombia, como se ve en la figura 4.

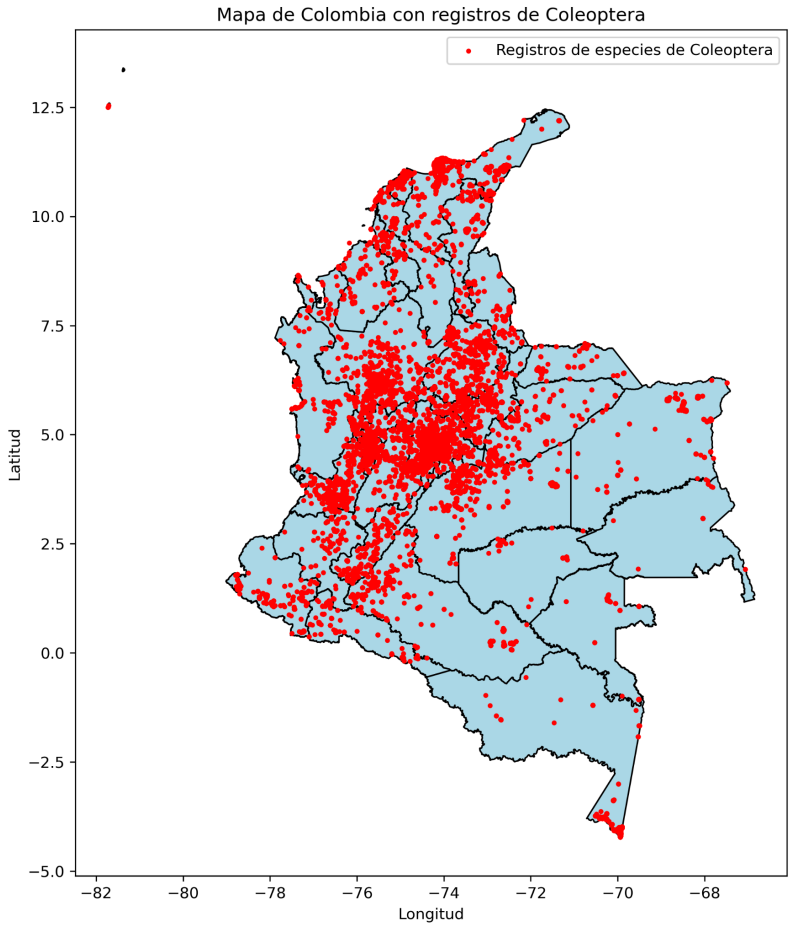

Figura 4. Mapa de Colombia con la distribución de los registros de Coleoptera del país.

Luego de la comprobación de los puntos en el mapa, se realizó el cruce espacial con las capas mencionadas en la sección 7.2 de tipo shapefile o raster, dependiendo de su origen. Esto se hizo con la función join de GeoPandas con un tipo *left*, dejando solamente la columna que tuviera información y eliminando las columnas adicionales.

Para documentar la información de las especies de plantas relacionadas, primero se realizó un preprocesamiento de los datos de plantas, eliminando los registros sin coordenadas y dejando solamente la información de plantas que estaban identificadas hasta nivel de especie.

La distancia total a la que se pueden mover los coleópteros a las plantas tanto para alimentarse, como para refugiarse depende significativamente de la especie. Algunos géneros con movimiento limitado como *Carabus* se mueven en promedio 120 m [108], otros de rango medio como *Cerambyx* se pueden mover 1080 m para hembras y 1498 para machos [109]. Unos valores muy similares se encontraron para la especie *Anoplophora glabripennis* de la familia Cerambycidae, con 1029 m para machos y 1442 m para hembras [110].

En otro estudio evaluando estrategias de hibernación de *Oreina cacaliae* se encontró que en promedio viajaban 500 m con valores máximos de 1000 m [111]. Algunos coleópteros de largo vuelo pueden encontrarse a distancias máximas de 43000 m [112]. Para los coleópteros parasitarios de las colmenas de la abeja *Aethina tumida,* la mayoría se encuentran a menos de 50 m hasta un máximo de 3200 m [113].

Debido a la amplia variedad de distancia a la que vuelan diferentes especies de coleópteros para interactuar con algunas plantas de interés, se utilizó un valor promedio de aproximadamente 1000 m para cubrir una distancia amplia sin sobrestimar el número de especies de plantas relacionadas. El cruce geoespacial se realizó para todas las especies de plantas que cayeran alrededor de todos los puntos de coleópteros, a los que se les creó un buffer de 0.01 grados y se concatenaron todas las especies de plantas que estuvieran en ese radio.

Para incorporar la variable dependiente "número de especies de coleópteros por familia" al conjunto de datos, se realizó un proceso de etiquetado mediante un cruce tipo *join-left*, utilizando como clave la columna de familia, junto con el archivo previamente construido con las listas de referencia descritas en la sección 7.2.

Este proceso permitió asignar a cada registro el número correspondiente de especies por familia, resultando en un conjunto de datos final compuesto por 373.927 registros y 60 variables que se puede observar en la tabla 4.

Tabla 4. Muestra del conjunto de datos con todas las columnas.

Columna	Ejemplo 1	Ejemplo 2	Ejemplo 3
gbifID	1891039191	5067703319	5067703487
occurrenceID	PUJ:MPUJ_ENT:0033709	PNNSFL:CEBUC:2022:ME:3:27	PNNSFL:CEBUC:2023:SACF:3:101
kingdom	Animalia	Animalia	Animalia
phylum	Arthropoda	Arthropoda	Arthropoda
class	Insecta	Insecta	Insecta
order	Coleoptera	Coleoptera	Coleoptera
family	Buprestidae	Elmidae	Elmidae
genus	Euchroma	Macrelmis	Disersus
species	Euchroma giganteum		
infraspecificEpithet			
taxonRank	SPECIES	GENUS	GENUS

Columna	Ejemplo 1	Ejemplo 2	Ejemplo 3
scientificName	Euchroma giganteum (Linnaeus, 1758)	Macrelmis Motschulsky, 1859	Disersus Sharp, 1882
verbatimScientificName	Euchroma giganteum	Macrelmis	Disersus
verbatimScientificNameAuthorship	(Linnaeus, 1758)	Motschulsky, 1859	Sharp, 1882
countryCode	CO	CO	CO
locality	Reserva El Neme, Casa de la Finca, 1 km SW de Coello	Parque Nacional Natural Selva de Florencia	Parque Nacional Natural Selva de Florencia
stateProvince	Tolima	Caldas	Caldas
occurrenceStatus	PRESENT	PRESENT	PRESENT
individualCount	1	1	4
decimalLatitude	42,801	5,512,489	5,504,469
decimalLongitude	-749,035	-75,038,481	-75,034,619
coordinateUncertaintyInMeters	1,000	4	4
coordinatePrecision			
elevation	328	1,388	1,208
elevationAccuracy	0	0	0
depth		0,47	1
depthAccuracy		0,41	0,6
eventDate	2015-03-09/2015-03-13	2022-07-28	2023-07-24
speciesKey	8150886		
basisOfRecord	PRESERVED_SPECIMEN	HUMAN_OBSERVATION	HUMAN_OBSERVATION
collectionCode	MPUJ	CEBUC	CEBUC
identifiedBy	D Moreno	Diana Rojas	Diana Rojas
dateIdentified	2015-01-01 0:00:00	2023-01-01 0:00:00	2024-01-01 0:00:00
license	CC_BY_4_0	CC_BY_4_0	CC_BY_4_0

Columna	Ejemplo 1	Ejemplo 2	Ejemplo 3
recordedBy	Tatiana Jaramillo;Karol Vera;Juanita Peñaranda	Omaira Henao;Duberney Giraldo;Orlando Marulanda;Sandra Lopez;Cesar Henao;Didier Alvarez;Jeferson Betancourt	Omaira Henao;Duberney Giraldo;Orlando Marulanda;Sandra Lopez;Cesar Henao;Didier Alvarez;Jeferson Betancourt
issue	INSTITUTION_MATCH_FUZZY	OCCURRENCE_STATUS_INFERRED_FROM_INDIVIDUAL_COUNT;COORDINATE_ROUNDED;TAXON_MATCH_SCIENTIFIC_NAME_ID_IGNORED	OCCURRENCE_STATUS_INFERRED_FROM_INDIVIDUAL_COUNT;COORDINATE_ROUNDED;TAXON_MATCH_SCIENTIFIC_NAME_ID_IGNORED
Coordinates	POINT (-74.90349999999999 4.2801)	POINT (-75.0384815 5.512489)	POINT (-75.034619000000015 5.504469)
departamento	Tolima	Caldas	Caldas
Radiacion_solar_global_promedio_multianual	4.5 - 5.0 kWh/m²	4.0 - 4.5 kWh/m²	4.0 - 4.5 kWh/m²
Humedad_Relativa_Anual_Promedio_Multianual_1981_2010	70 - 75	80 - 85	80 - 85
Temperatura_Maxima_Media_Mensual_Promedio_Multianual_1981_2010	32 – 34	20 - 24	24 - 28
Temperatura_Media_Mensual_Promedio_Multianual_1981_2010	> 28	16 - 20	16 - 20
Temperatura_Minima_Media_Mensual_Promedio_Multianual_1981_2010	22 - 24	12 - 16	12 - 16
Velocidad_viento_10_mtrs_altura_Mensual_2000_2010	2 - 3	2 - 3	2 - 3
Velocidad_viento_10_mtrs_altura_Mensual_2000_2010.1	2 - 3	2 - 3	2 - 3
elevacion_dem	304	1359	1017
ECC_Prcp_1976_2005_100K_2015	1	9	9

Columna	Ejemplo 1	Ejemplo 2	Ejemplo 3
especiePlanta	Rondeletia pubescens | Dalechampia karsteniana | Petrea volubilis | Prestonia quinquangularis | Indigofera lespedezioides | Chiococca alba | Pithecellobium dulce | Tillandsia elongata | Achatocarpus nigricans | Aphelandra glabrata | Ricinus communis | Jatropha gossypiifolia	Cordia bicolor | Solanum sycophanta | Tibouchina lepidota | Vismia baccifera | Philodendron verrucosum | Siparuna aspera	Mezobromelia capituligera | Guzmania lingulata | Vriesea elata | Guzmania glomerata | Guzmania squarrosa | Vriesea chrysostachys | Werauhia gladioliflora | Aechmea veitchii | Guzmania angustifolia | Guzmania danielii | Guzmania polycephala | Guzmania patula
ECOS_GENER	Agroecosistema de mosaico de cultivos y pastos	Vegetacion secundaria	Agroecosistema ganadero
UNIDAD_SIN	Agroecosistema de mosaico de cultivos y pastos de clima Calido Semiarido en Terrazas con suelo de Condiciones oxidantes en sedimentos jovenes y Regimen ustico	Vegetacion secundaria de clima templado superhumedo en filas y vigas con suelo de poca profundidad efectiva de los suelos y condiciones oxidantes y evolucion moderada o incipiente	Agroecosistema ganadero de clima Templado Superhumedo en Filas y vigas con suelo de Poca profundidad efectiva de los suelos y Condiciones oxidantes y evolucion moderada o incipiente
CLIMA	Calido Semiarido	Templado Superhumedo	Templado Superhumedo
RELIEVE	Terrazas	Filas y vigas	Filas y vigas

Columna	Ejemplo 1	Ejemplo 2	Ejemplo 3
SUELOS	Fluventic Haplustepts, Typic Ustifluvents	Lithic Udorthents, Typic Dystrudepts	Lithic Udorthents, Typic Dystrudepts
COBERTURA	Mosaico de cultivos y pastos	Vegetacion secundaria	Pastos
No_Anfibio	22	16	16
No_Aves	29	31	31
No_Magnoli	30	31	31
No_Mamifer	24	21	21
No_Reptile	20	17	17
NumeroEspeciesFamilia	147	60	60

En esta fase de entendimiento de los datos, se llevó a cabo un análisis exploratorio donde se identificaron un total de 16 variables numéricas, mientras que el resto corresponden a variables categóricas. Se examinó la distribución de las variables numéricas, cuyos resultados se presentan en la Figura 5. En los histogramas se observó que ninguna de estas variables sigue una distribución normal y que hay varias dicotómicas con frecuencias dispares .

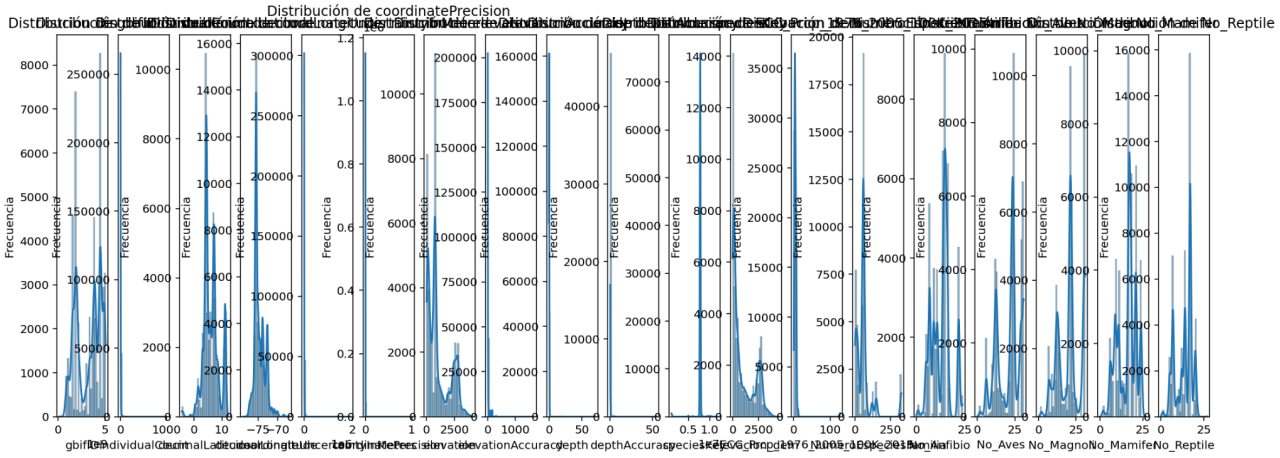

Figura 5. Distribución de las variables numéricas en el conjunto de datos.

En cuanto a las variables categóricas, se evaluó su grado de correlación que se presentan en la Figura 6. Se detectó una alta correlación en varias de estas variables, sobretodo las que tienen un componente espacial, esto podría representar redundancia informativa, sin embargo al ser parte de la información identificada como relevante para los coleopteros, no fueron eliminadas ni modificadas. Además, algunas variables presentaron un valor

constante en todos los registros, careciendo de variabilidad y, por ende, de utilidad analítica. Estas variables fueron eliminadas del conjunto de datos para optimizar el proceso de modelado y evitar posibles sesgos.

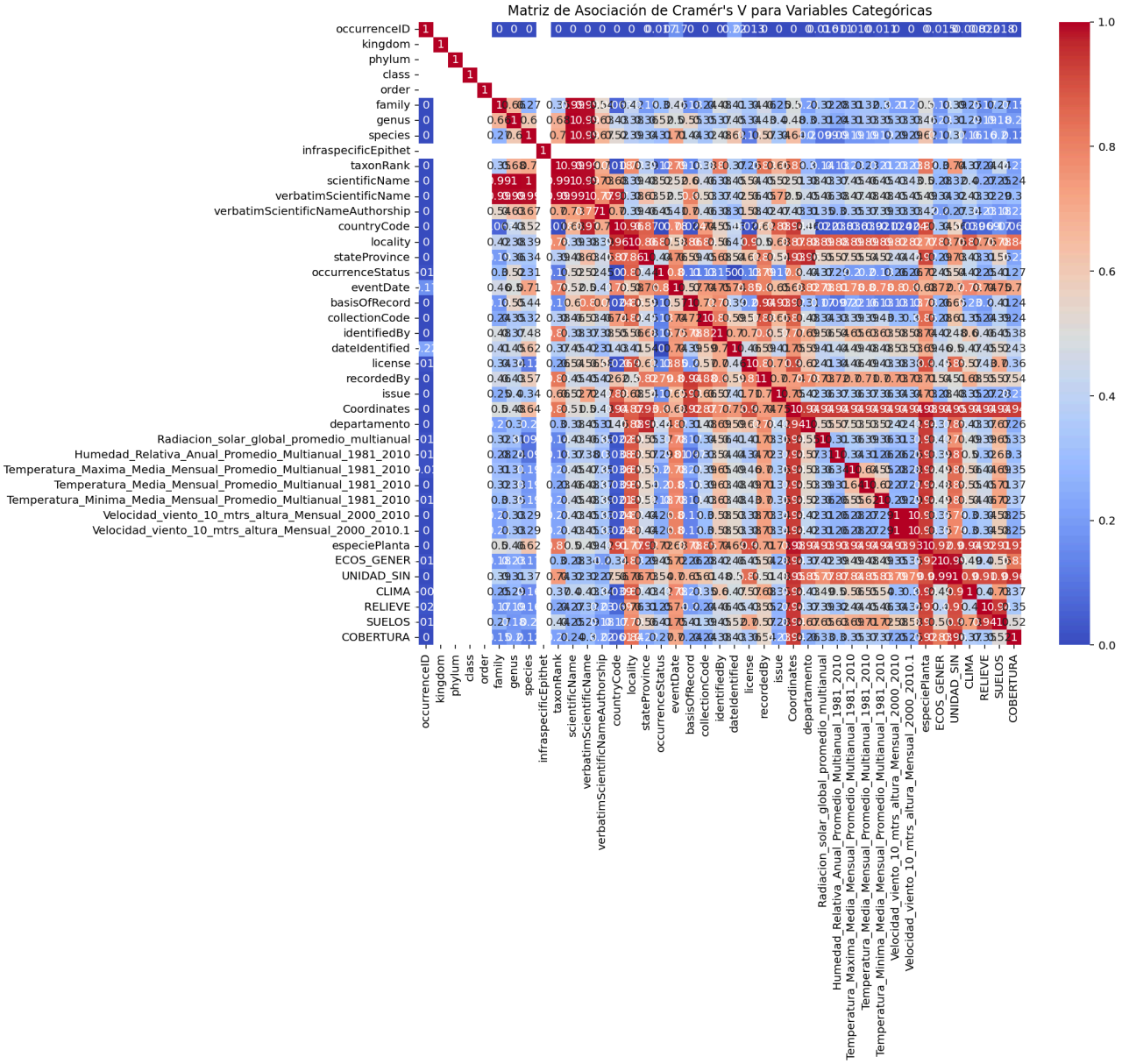

Figura 6. Correlación de las variables categóricas en el conjunto de datos.

Para abordar los datos faltantes, se realizó un proceso de imputación sencillo para datos numéricos y categóricos, utilizando la mediana y la moda respectivamente. Esto se realizó en un total de 40 columnas y un total de 3’230.088 de valores. Este proceso no fue realizado con la variable respuesta 'NumeroEspeciesFamilia', que se dejó con los datos vacíos que estaban sin etiquetar.

Los datos categóricos se transformaron a números para poder realizar un entrenamiento de modelos de aprendizaje profundo, esto se hizo utilizando un label encoder en todas las

columnas que tenían datos de tipo string, guardando la información del diccionario para poder hacer la transformación inversa al finalizar el proceso de entrenamiento.

Para finalizar el proceso de alistamiento de los datos, se quitaron algunas columnas que tenían datos iguales, estaban muy correlacionados o eran códigos únicos propios del conjunto de datos original, y se dejaron únicamente los datos que estaban etiquetados para poder hacer el entrenamiento de los modelos. El resultado final de este procesamiento fue un conjunto de datos de 46.892 filas y 50 columnas, que se puede observar en la tabla 5.

Tabla 5. Muestra del conjunto de datos con todas las columnas luego de utilizar el label encoder.

Columna	Ejemplo 1	Ejemplo 2	Ejemplo 3
family	32	35	23
genus	156	327	156
species	73	73	73
taxonRank	0	1	0
scientificName	380	506	317
verbatimScientificName	151	143	133
verbatimScientificNameAuthorship	372	372	372
countryCode	0	0	0
locality	787	787	787
stateProvince	31	31	31
occurrenceStatus	1	1	1
individualCount	1	1	1
decimalLatitude	306,417	3,067	30,675
decimalLongitude	-753,731	-75,384	-753,694
coordinateUncertaintyInMeters	505	505	505
coordinatePrecision	1	1	1
elevation	545	846	544
elevationAccuracy	0	0	0
depth	0,3	0,3	0,3
depthAccuracy	125	125	125
eventDate	1217	1217	1217
speciesKey	7745184	7745184	7745184
basisOfRecord	2	2	2
collectionCode	10	10	10

Columna	Ejemplo 1	Ejemplo 2	Ejemplo 3
identifiedBy	61	61	61
dateIdentified	1738	1738	1738
license	1	1	1
recordedBy	1213	1213	1213
issue	49	49	49
Coordinates	3650	3679	3633
departamento	17	17	17
Radiacion_solar_global_promedio_multianual	3	3	3
Humedad_Relativa_Anual_Promedio_Multianual_1981_2010	1	1	1
Temperatura_Maxima_Media_Mensual_Promedio_Multianual_1981_2010	5	5	5
Temperatura_Media_Mensual_Promedio_Multianual_1981_2010	5	4	5
Temperatura_Minima_Media_Mensual_Promedio_Multianual_1981_2010	3	3	3
Velocidad_viento_10_mtrs_altura_Mensual_2000_2010	5	5	5
Velocidad_viento_10_mtrs_altura_Mensual_2000_2010.1	5	5	5
elevacion_dem	525	793	523
ECC_Prcp_1976_2005_100K_2015	3	3	3
especiePlanta	1133	3655	537
ECOS_GENER	55	8	55
UNIDAD_SIN	1135	490	1141
CLIMA	3	3	3
RELIEVE	27	23	27
SUELOS	108	277	108
COBERTURA	1	32	17
No_Anfibio	22	22	22
No_Aves	29	29	29
No_Magnoli	30	30	30

Columna	Ejemplo 1	Ejemplo 2	Ejemplo 3
No_Mamifer	24	24	24
No_Reptile	20	20	20
NumeroEspeciesFamilia	8	25	78

Todas las variables han sido transformadas para ser numéricas y no tener ningún valor faltante, cada una de las filas del dataset final representa un registro biológico de un coleóptero en Colombia, incluyendo variables taxonómicas, espaciales y geográficas, climáticas y ambientales, ecológicas y metadatos de registro generales como código de colección, tipo de registro y persona que lo registró e identificó.

Antes de llevar a cabo el proceso de entrenamiento de los modelos, se realizó una división del conjunto de datos en un 75% para entrenamiento, un 5% para prueba (test) y un 20% para validación. La variable “NumeroEspeciesFamilia” se guardó en un vector individual por ser la dependiente y el resto de variables se usaron como variables independientes.

8.2. CONSTRUCCIÓN DE LOS MODELOS

La predicción del número de especies se realizó por medio de la construcción de cuatro tipos de modelos diferentes de regresión, regresión lineal múltiple como modelo base debido a su simplicidad y menor complejidad computacional, y tres tipos de modelos de aprendizaje automático: perceptrón multicapa, redes neuronales profundas y random forest. Estos fueron construidos con las librerías *Scikit-learn* y *Keras* en Python. En los modelos se utilizó MSE como función de pérdida, MAE como métrica y se evaluaron las predicciones utilizando R^2.

Para la regresión lineal múltiple, se ajustaron los datos de entrenamiento para predecir la variable número de especies en la familia usando la función *LinearRegression* de Sklearn.

Para las redes neuronales, se comenzó con una red sencilla tipo perceptrón multicapa con una capa de entrada con 48 neuronas, una capa oculta con 64 neuronas con función de activación *ReLU* y una capa de salida con una única neurona para predecir un valor continuo. La arquitectura resultante se puede ver en la figura 7.

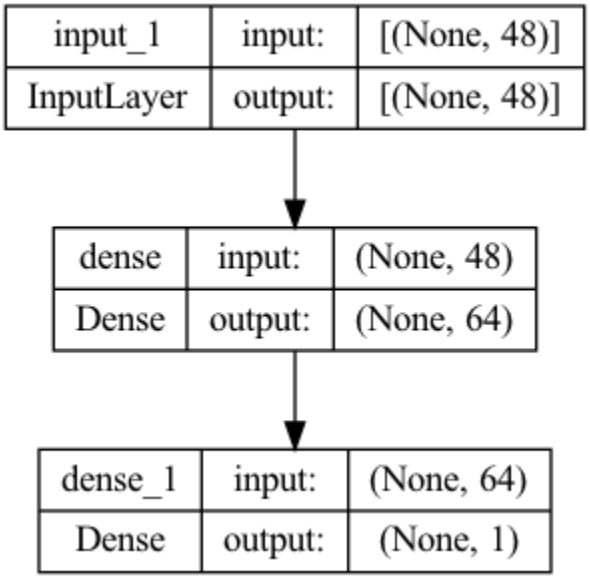

Figura 7. Arquitectura del modelo Perceptrón multicapa.

Para definir la arquitectura de las redes neuronales profundas, se realizó una búsqueda de hiperparámetros mediante el método *Hyperband*, utilizando la herramienta *Keras Tuner*, para las siguientes características:

- Número de capas ocultas de 1 a 5.
- Número de neuronas de 32 a 128, con incrementos de 32.
- Función de activación 'relu', 'tanh', 'elu', 'leaky_relu' o 'swish'.
- Learning rate entre 0.01 y 0.0001.
- Capas de dropout entre 0.2 y 0.5 con saltos de 0.1.
- Optimizador tipo ADAM.
- Función de pérdida tipo MSE.

Este proceso se realizó para dos configuraciones distintas de los datos, en la primera se utilizaron los datos en su forma original y la segunda se aplicó un proceso de escalamiento estándar en los datos para normalizar la información antes del entrenamiento. La arquitectura resultante para la red sin escalar se puede ver en la figura 8.

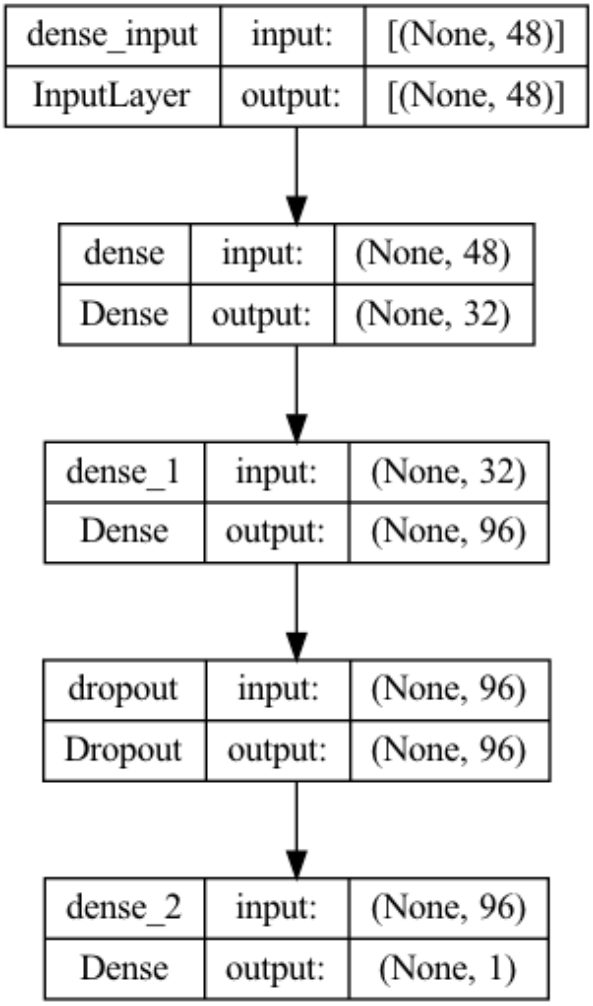

Figura 8. Arquitectura del modelo de Red Neuronal Profunda.

Para normalizar las variables de entrada y mejorar la estabilidad del proceso de entrenamiento, se aplicó un escalamiento estándar (*StandardScaler*) a todas las variables independientes del conjunto de datos. Este proceso se implementó mediante la herramienta *StandardScaler* de la biblioteca Scikit-learn, la cual transforma cada variable para que tenga una media de 0 y una desviación estándar de 1. La arquitectura de red se puede ver en la figura 9.

La variable de salida (número de especies por familia) no fue escalada, debido a que se encuentra en un rango acotado y no presenta valores extremos y que conservar la escala original de la variable facilita la interpretación directa de las predicciones generadas por el modelo.

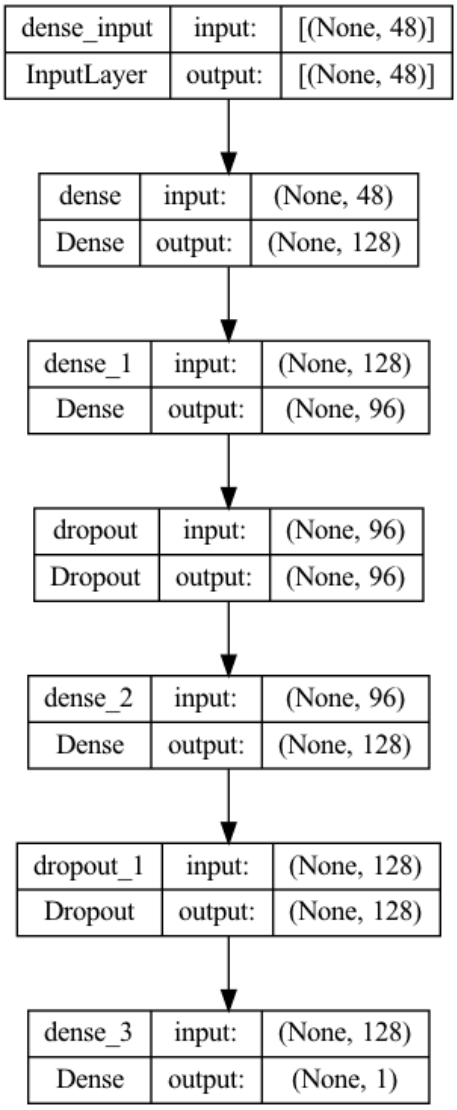

Figura 9. Arquitectura del modelo de Red Neuronal Profunda con datos escalados.

Para definir la arquitectura de del regresor tipo Random Forest, se realizó una búsqueda de hiperparámetros mediante el método grilla de búsqueda aleatoria, utilizando la herramienta *RandomizedSearchCV*, para las siguientes características:

- Número de árboles en el bosque de 20 a 200.
- Profundidad máxima de los árboles de 3 a 12.
- Mínimo número de muestras para dividir un nodo de 2 a 5.
- Mínimo número de muestras por hoja de 1 a 5.
- Parámetro de complejidad para la poda del árbol entre 0, 0.1, 0.01 y 0.001.

A partir de esta configuración se realizó la búsqueda con una validación cruzada de 10 particiones usando el Error Absoluto Medio Negativo como métrica de evaluación. El Random Forest genera diversos árboles de decisión, un ejemplo para ilustrar el resultado

de esta arquitectura se puede ver en la figura 10.

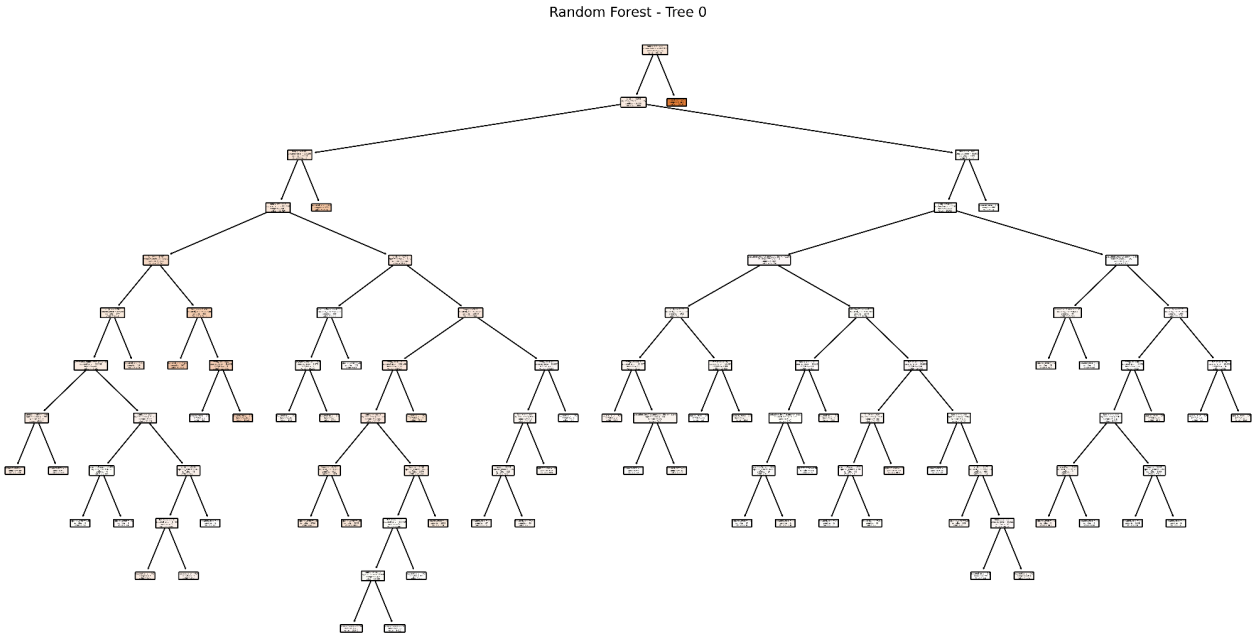

Figura 10. Ejemplo del primer árbol de decisión identificado en el RandomForest.

La configuración final de estos modelos es el resultado de múltiples ajustes y validaciones realizadas de forma iterativa, lo que implicó una carga computacional considerable y múltiples ciclos de ajuste.

Todo el código utilizado para la preparación de los datos, análisis exploratorio, entrenamiento de los modelos y predicción se encuentra disponible en el repositorio de GitHub https://github.com/EstebanMH-SiB/modelPredictColeopteraSpecies/tree/main, además están disponibles los datos anotados, los conjuntos de datos más relevantes y los modelos en formato exportable.

9. EVALUACIÓN DEL DESEMPEÑO DE LOS MODELOS Y ESTIMACIÓN DEL NÚMERO DE ESPECIES

El proceso de evaluación del desempeño de los modelos se realizó de dos formas, la primera utilizando métricas matemáticas comunes para evaluar el rendimiento de modelos de regresión, como el error cuadrático medio (MSE), el error absoluto medio (MAE) y el coeficiente de determinación (R^2). La segunda fue a través de métodos indirectos, con una elicitación experta por parte de entomólogos del país y una comparación con el número total de especies reportada para el país hasta el momento.

9.1 EVALUACIÓN CON MÉTRICAS NUMÉRICAS

Las métricas matemáticas fueron calculadas para todos los modelos y a partir de estas se seleccionó el modelo que tuviera un mejor desempeño para realizar la predicción que se compartió con los expertos para su validación.

La regresión lineal obtuvo valores de 0.20 para el R^2, 41.32 para MAE y 3716.32 para el MSE. En la figura 11 se puede ver los valores predichos del modelo difieren de forma significativa de la línea punteada ideal, mostrando que no es un buen ajuste. También el modelo predice valor menores a los reales, sin llegar a predecir nunca valores por encima de 250.

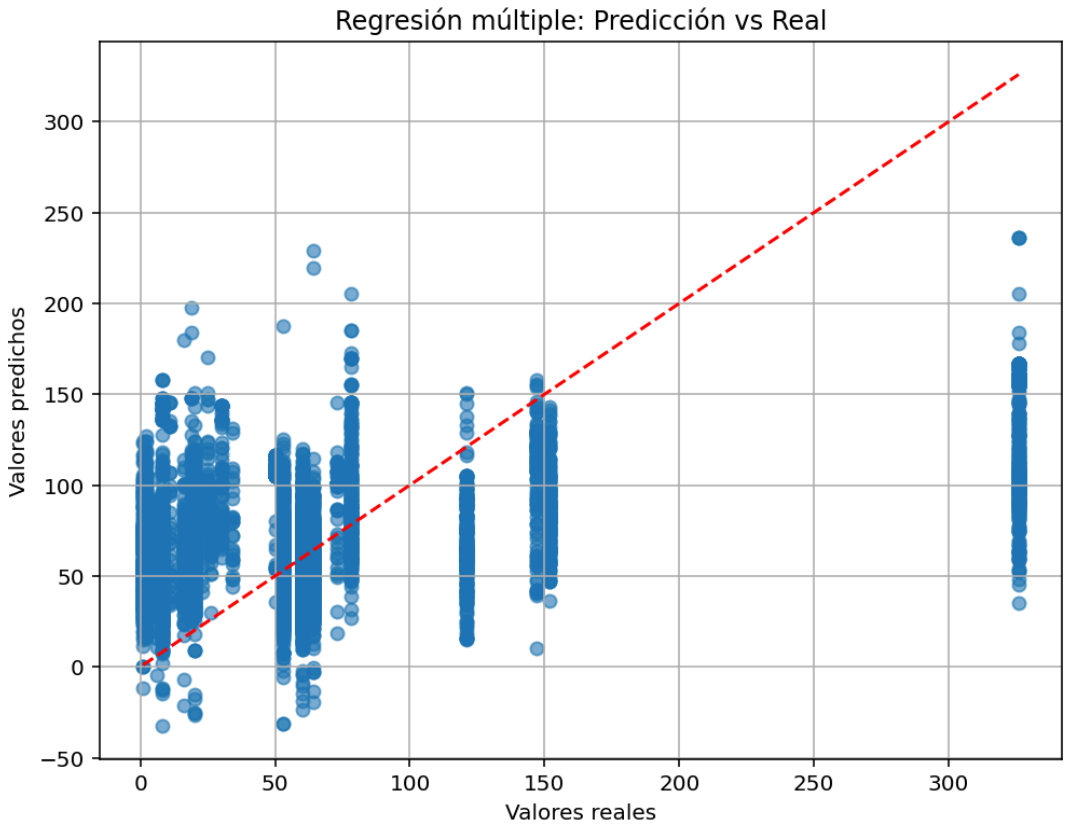

Figura 11. Valores predichos versus reales para el modelo de regresión lineal.

Para todas las redes neuronales, se realizó la evaluación del MAE y MSE tanto para los datos de prueba como para los de entrenamiento y se calculó el R^2 con la predicción.

El perceptrón multicapa obtuvo valores de -0.074 para el R^2, 46.53 para MAE y 5048.65 para el MSE.
En la figura 12 presenta la evolución de las métricas durante 50 épocas de entrenamiento, en esta se puede ver el MSE comienza con un valor extremadamente algo que disminuye en las primeras épocas y luego se mantiene estable hasta el final donde aumenta un poco

para el set de prueba.

Para el MAE se presenta una mayor estocasticidad, sobre todo en los datos de prueba y no termina de converger, con un aumento importante en las últimas épocas. Este modelo se podría beneficiar con un early stopping para usar el modelo antes de que empeoren las métricas.

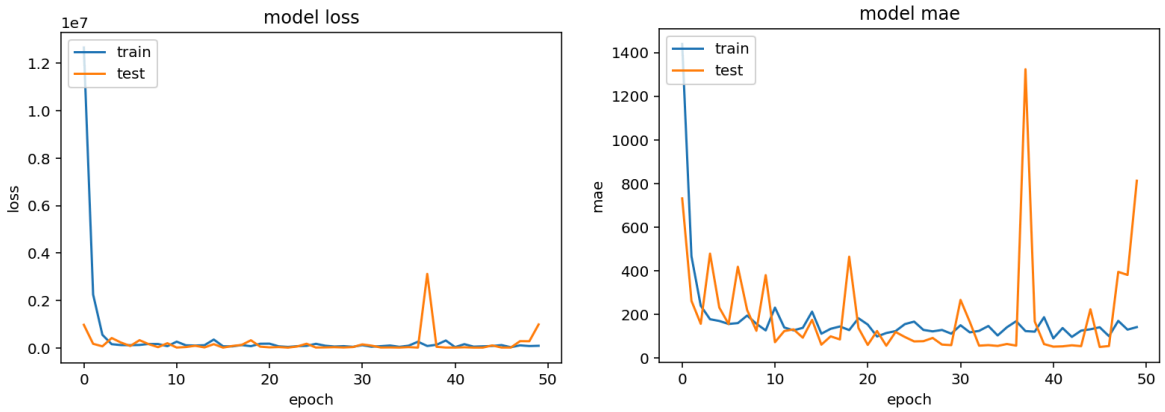

Figura 12. Valores de MAE y MSE para el perceptrón multicapa en 50 épocas.

La red neuronal profunda sin escalar obtuvo valores de 0.35 para el R2, 34.24 para MAE y 3041.94 para el MSE.

En la figura 13 presenta la evolución de las métricas durante 50 épocas de entrenamiento, en esta se puede ver el MSE se estabiliza luego de las primeras épocas en un valor relativamente bajo, con los datos de prueba siempre debajo de la línea de entrenamiento, lo que demuestra un buena estabilidad.

Para el MAE se presenta una mayor estocasticidad, sobre todo en los datos de prueba y no termina de converger, los datos de entrenamiento se mantienen relativamente estables luego de la época 15, pero las primeras épocas tienen valores menores y se podrían beneficiar de un early stopping.

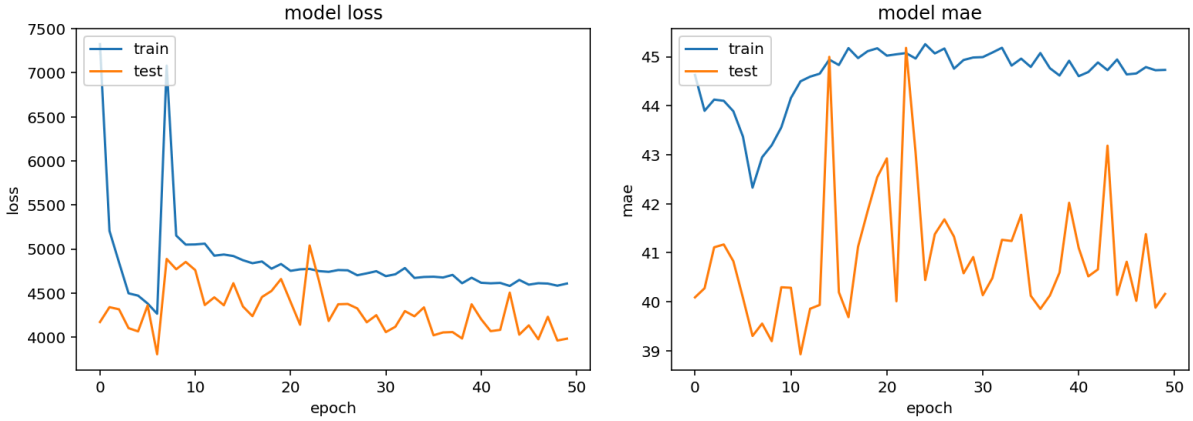

Figura 13. Métricas numéricas modelo red neuronal no escalada en 50 épocas.

La red neuronal profunda escalada obtuvo valores de 0.98 para el R2, 4.07 para MAE y 92.6 para el MSE.

En la figura 14 presenta la evolución de las métricas durante 50 épocas de entrenamiento, en esta se puede ver el MSE para el entrenamiento tiene una gran varianza y solamente se estabiliza luego de 40 épocas en un valor relativamente bajo, mientras que los datos de prueba no tienen tanta variación y se mantienen bajos en todo momento.

Para el MAE tanto los datos de entrenamiento como de prueba se estabilizan en las primera 10 épocas reduciendo su valor, y luego disminuye lentamente hasta la última época.

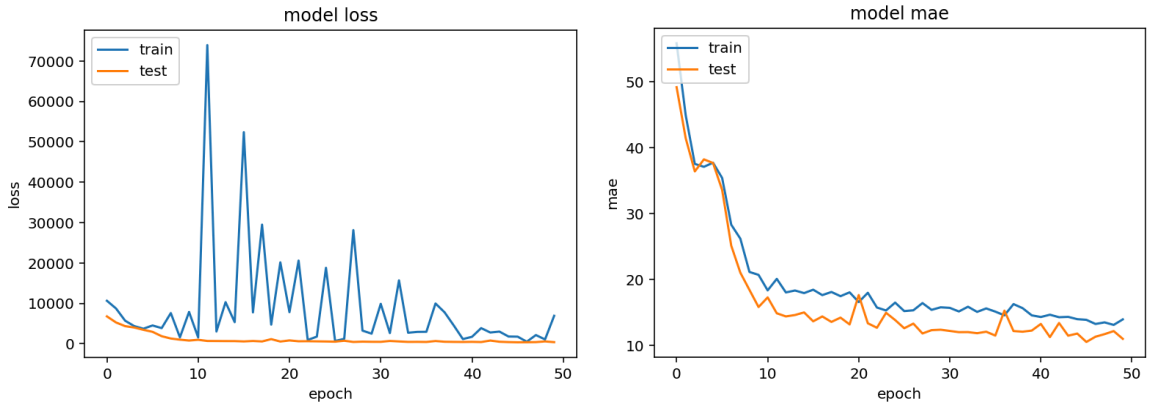

Figura 14. Métricas numéricas modelo red neuronal escalada en 50 épocas.

El Random Forest obtuvo valores de 0.96 para el R2, 4.42 para MAE y 166.98 para el MSE.

En la figura 15 se puede ver los valores predichos del modelo se comportan relativamente bien hasta 150, pero luego de este valor no logran hacer predicciones tan buenas y en muchos casos son cercanas a 0 cuando debería ser el valor máximo.

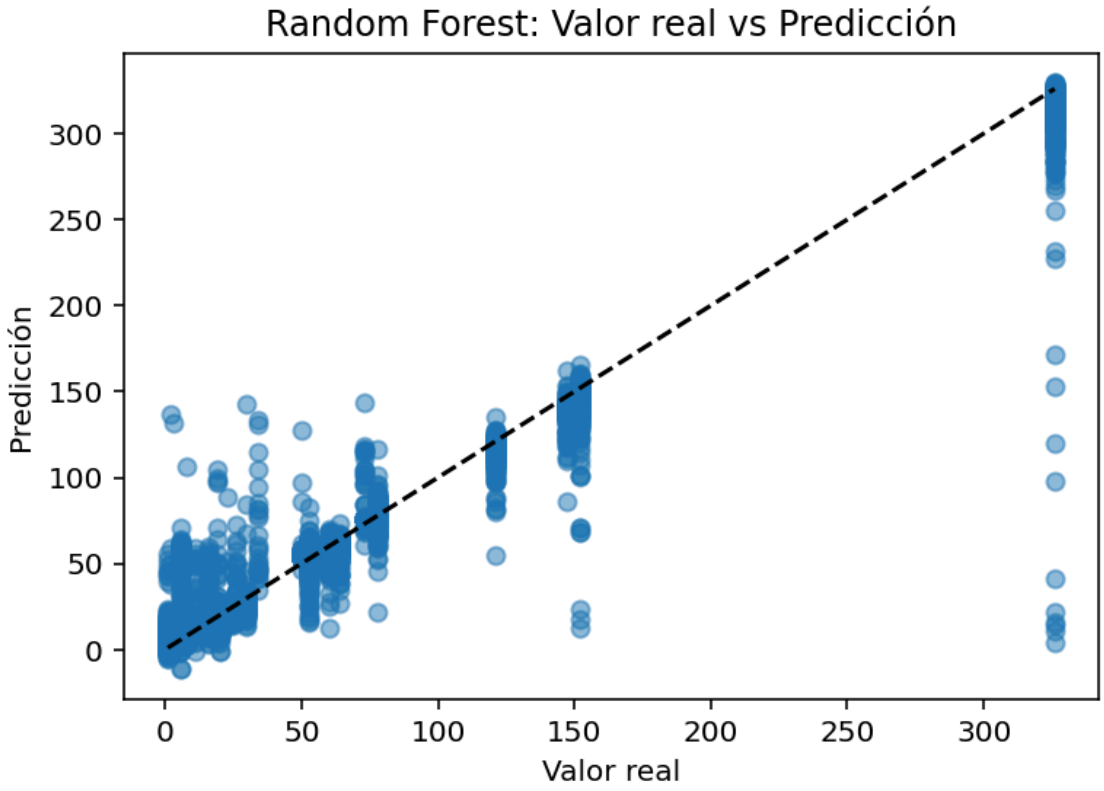

Figura 15. Métricas numéricas modelo random forest

El resultado compilado de las métricas para todos los modelos se presenta en la tabla 6, donde se muestra que la Red Neuronal Profunda con escalamiento tuvo el mejor rendimiento para en las tres métricas, con el valor más alto de R^2 muy cercano a 1 que indica un ajuste casi perfecto y los valores más bajos de MAE y MSE que demuestran el bajo error obtenido en las predicciones. Por estos motivos este fue el modelo seleccionado para realizar las predicciones y algunas validaciones adicionales.

Tabla 6. Métricas de evaluación para los modelos realizados.

Modelo	R^2	MAE	MSE
Regresión lineal múltiple	0.20	41.32	3716.32
Perceptrón multicapa simple	-0.074	46.53	5048.65
Red Neuronal Profunda sin escalar	0.35	34.24	3041.94

Red Neuronal Profunda con escalamiento	**0.98**	**4.07**	**92.6**
Random Forest	0.96	4.42	166.98

Para evaluar la estabilidad y capacidad de generalización del modelo escogido, se realizó una validación cruzada de 10 particiones, lo que permite estimar el rendimiento de una manera más robusta evitando sesgos presentes en la partición original de los datos. El promedio y la desviación estándar para las métricas fueron las siguientes:

R^2: 0.9849 ± 0.0065
MSE promedio: 69.2735 ± 32.7074
MAE promedio: 3.6190 ± 0.2924

Estos resultados se muestran en la figura 16 para cada una de las particiones.

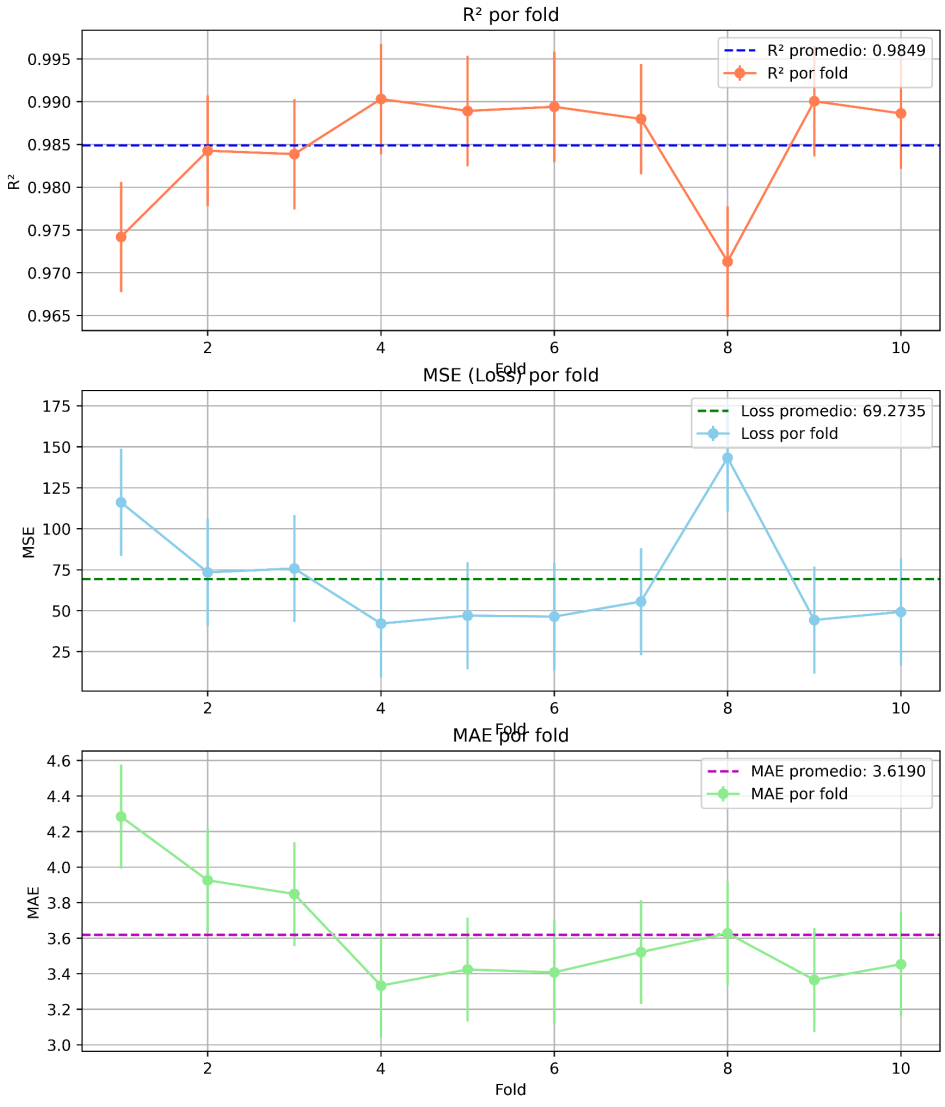

Figura 16. Métricas numéricas obtenidas para cada partición en la validación cruzada de la red neuronal profunda escalada.

Los valores reflejan un comportamiento consistente y estable del modelo ante distintos subconjuntos de datos, con bajas desviaciones estándar para todas las métricas y un desempeño fiable en todos los casos.

9.2 PREDICCIÓN DEL NÚMERO DE ESPECIES DE COLEÓPTEROS PARA ANTIOQUIA Y COLOMBIA

Para realizar la estimación del número de especies, se utilizó la red neuronal profunda con datos escalados, aplicando la función *predict* sobre el conjunto de datos sin etiquetar para Colombia luego del proceso de codificación y escalamiento, este conjunto de datos contiene 373.927 registros y 48 columnas.

Teniendo en cuenta que los datos están etiquetados a nivel de familia, pero los modelos de regresión generan un valor numérico para cada registro individual, fue necesario agrupar las predicciones a nivel de familia. Para ello, se utilizó una tabla dinámica (*pivot table*) que agrupó las predicciones por la columna familia y se calcularon cuatro métricas: promedio, mediana, valor mínimo y valor máximo.

Después de obtener la tabla de resultados, se realizó un proceso de limpieza para dejar únicamente las familias presentes en la taxonomía de Lawrence & Newton [114], eliminando 31 subfamilias para Colombia y 17 para Antioquia, con el fin de tener un mejor estimado del número de especies presentes. Finalmente se sumaron todos los estimados por familia para obtener los siguientes resultados a nivel país:

- Promedio: 6.420 especies
- Mediana: 6.181 especies
- Mínimo: 3.226 especies
- Máximo: 14.991 especies

Para la estimación del número de especies de Coleoptera en el departamento de Antioquia, se tomó el conjunto de datos sin etiquetar para Colombia luego del proceso de codificación y escalamiento y se realizó un filtro para el departamento, obteniendo un nuevo conjunto con 79.127 registros y 48 columnas. Sobre este se realizó el mismo proceso de agrupamiento en la tabla dinámica y se sumaron todos los estimados por familia luego de la limpieza para obtener los siguientes resultados en Antioquia:

- Promedio: 4.210 especies
- Mediana: 3.882 especies

- Mínimo: 2.007 especies
- Máximo: 9.381 especies

El valor estimado de especies de coleópteros por familia para Antioquia se puede ver en la tabla 7.

Tabla 7. Número estimado de especies de coleópteros por familia para el Departamento de Antioquia.

Familia	Promedio	Mediana	Mínimo	Máximo
Aderidae	51	51	51	51
Anobiidae	94	60	28	228
Anthicidae	15	19	0	21
Anthribidae	16	15	3	37
Attelabidae	92	92	92	92
Bostrichidae	124	103	19	270
Brentidae	83	86	6	146
Buprestidae	146	174	20	221
Cantharidae	67	59	8	257
Carabidae	82	54	1	304
Cerambycidae	98	80	6	278
Ceratocanthidae	40	40	40	40
Chrysomelidae	86	67	-42	257
Cleridae	64	64	58	71
Coccinellidae	114	117	21	447
Curculionidae	74	62	14	214
Dascillidae	50	50	50	50
Dryopidae	41	31	15	110
Dytiscidae	65	59	22	182
Elateridae	104	103	39	171
Elmidae	51	47	-110	298
Endomychidae	83	59	43	169
Erotylidae	115	106	54	207
Geotrupidae	74	74	72	76
Gyrinidae	73	68	30	164
Haliplidae	116	116	116	116
Heteroceridae	57	55	55	59
Histeridae	110	112	45	163
Hybosoridae	66	69	47	99
Hydraenidae	87	74	43	132
Hydrophilidae	70	68	44	158
Hydroscaphidae	48	48	43	53

Lampyridae	85	83	42	124
Leiodidae	64	59	55	112
Limnichidae	61	63	44	76
Lucanidae	99	102	59	120
Lutrochidae	62	62	52	73
Lycidae	88	90	48	116
Lymexylidae	73	60	51	110
Meloidae	92	92	68	157
Melyridae	87	66	59	144
Mordellidae	52	50	44	62
Mycetophagidae	77	74	73	84
Nitidulidae	70	65	39	108
Noteridae	70	69	15	144
Oedemeridae	35	35	29	42
Passalidae	86	60	33	647
Phengodidae	54	51	31	72
Psephenidae	65	64	24	134
Ptiliidae	71	71	69	75
Ptilodactylidae	63	70	19	117
Scarabaeidae	37	24	-14	920
Scirtidae	38	21	12	130
Scydmaenidae	27	26	23	34
Silphidae	19	21	13	24
Silvanidae	23	23	23	23
Staphylinidae	37	23	-1	216
Tenebrionidae	63	19	11	218
Trogossitidae	156	157	153	158
Total	4210	3882	2007	9381

La figura 17 muestra la distribución del número estimado de especies promedio por familia en Antioquia, donde se observa una alta variabilidad, la mayoría de las familias tienen menos de 100 especies estimadas y en promedio hay 71 por familia. *Trogossitidae* tiene el mayor número estimado con 156 y *Anthicidae* el menor con 15.

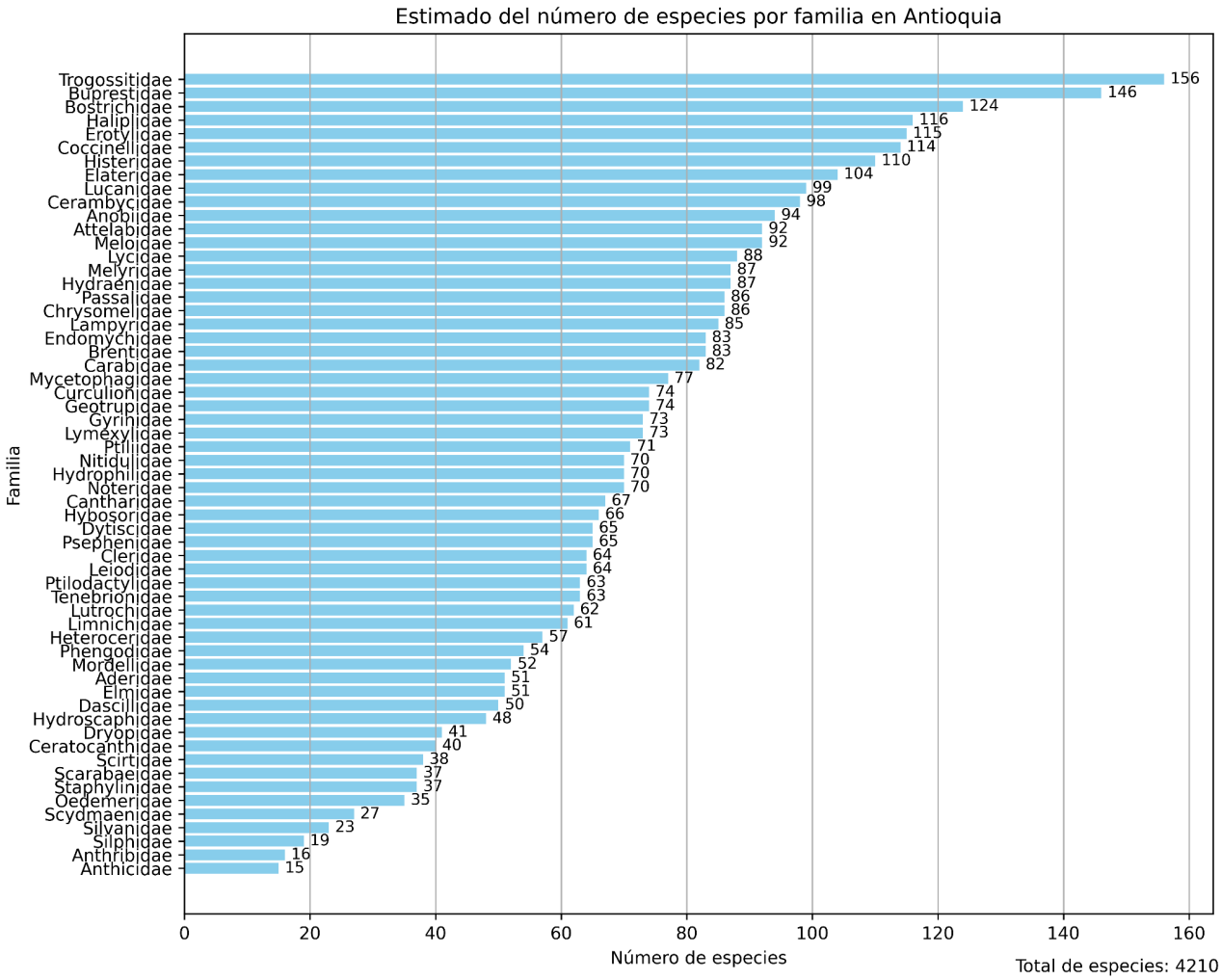

Figura 17. Número de especies promedio por familia predicho para el Departamento de Antioquia.

Estas estimaciones no corresponden a una ventana temporal específica, por el contrario pretenden representar el número real de especies de Coleoptera total en el departamento. En los próximos años se espera que se realicen más estudios de taxonomía reportando nuevas especies o presencia de especies en el departamento y el número actual de especies debería aumentar y luego ubicarse en alguna parte del rango estimado.

Sin embargo, no es posible estimar con certeza en que fecha particular se logre llegar a este número, pero podría ser entre 10 a 20 años si sigue la tendencia de especies con al menos algún registro biológico publicado a través del SiB Colombia, teniendo en cuenta que en 2018 habían 320 especies para el país [115] y en promedio han aumentado 500 especies por año hasta las 3.249 especies en 2024.

A pesar de que el modelo presenta un buen desempeño general en las métricas de ajuste, se identificaron algunas limitaciones importantes, entre las que se encuentra la posibilidad de predecir valores negativos, lo que es imposible en un contexto biológico y es una

oportunidad de mejorar en trabajos futuros, implementando restricciones para garantizar que el valor siempre sea ≥ 0. También se presentan incongruencias entre las escalas departamental y nacional, donde en algunos casos el número estimado de especies en Antioquia supera el estimado nacional para la misma familia, lo que es inconsistente desde un punto de vista biogeográfico.

9.3 EVALUACIÓN POR PARTE DE EXPERTOS DE LOS RESULTADOS DEL MODELO

Se realizó un proceso de elicitación experta para validar el resultado de la predicción del número de especies para Colombia y para Antioquia. Para esto se realizó una consulta con algunos coleopterólogos para identificar expertos en el departamento y una búsqueda en el directorio de especialistas del Registro Único Nacional de Colecciones Biológicas (RNC) para identificar curadores que trabajaran en Coleoptera en el departamento de Antioquia [116]. Posteriormente se envió un correo electrónico con una descripción del proyecto, la estimación del número de especies a nivel de país y departamento y la solicitud de evaluar esta estimación con base en la experiencia personal. En total se enviaron 13 solicitudes y se obtuvieron 5 respuestas.

Según la entomóloga Juliana Cardona Duque [117], el valor promedio estimado del número de especies para Colombia en este estudio (6.420 especies) podría estar subestimado, considerando que para el año 2021 habían registrados 5.000 especies únicamente para 20 familias.

En cuanto a la estimación para el departamento de Antioquia, su evaluación resulta compleja porque presenta una alta heterogeneidad geográfica, con la posible presencia de hasta siete cordilleras, además de una marcada influencia del Chocó Biogeográfico, el Caribe y Centroamérica.

De acuerdo con el entomólogo Diego Martínez Revelo [118], el valor promedio estimado del número de especies para Colombia de la subfamilia Scarabaeinae se encuentra subestimado, teniendo en cuenta que existen aproximadamente 400 especies registradas para el país, mientras que el presente modelo solo predice 41.

Respecto a la estimación para el departamento de Antioquia, está también parece estar subestimada, considerando que la colección del Museo Francisco Luis Gallego, se han identificado al menos 67 especies pertenecientes a esta subfamilia.

De acuerdo con el entomólogo Larry Jiménez Ferbans [119], la estimación promedio del número de especies de la familia Passalidae para Colombia, se encuentra subestimada, ya que se han registrado más de 130 especies en el país, mientras que el modelo utilizado en este estudio predice únicamente 56.

En contraste, la estimación para el departamento de Antioquia parece estar sobreestimada, teniendo en cuenta que se estima la presencia de aproximadamente 40 especies, aunque el modelo predice 86. Esta discrepancia sugiere que los datos disponibles reportan una diversidad menor a la proyectada por el modelo.

Según el entomólogo Juan Pablo Botero [120], la estimación promedio del número de especies Coleóptera para Colombia se encuentra subestimada. Esta apreciación se fundamenta en la comparación con Brasil, donde se han reportado 36.043 especies de Coleoptera. Si bien Brasil posee una mayor extensión territorial, la diferencia en riqueza específica no debería ser tan marcada, dado que Colombia también presenta una alta diversidad de ecosistemas.
En este sentido, se estima que el número real de especies de coleópteros en Colombia podría superar las 20.000.

De manera específica, la estimación para la familia Cerambycidae también presenta un sesgo hacia la baja. Aunque actualmente se han documentado aproximadamente 1.023 especies en el país, el modelo predictivo empleado en este estudio reporta únicamente 97 en Colombia. Según Botero, el número real para esta familia podría alcanzar, como mínimo, las 2.000 especies.

En cuanto a la estimación del número de especies para la familia Cerambycidae para el departamento de Antioquia, también parece estar subestimada. Aunque actualmente se han registrado menos 87 especies en el departamento y el modelo predice en promedio 98, se estima que el número real podría situarse entre 150 y 200 especies. Esta proyección se basa en comparaciones con otros departamentos del país que presentan registros más elevados, como Bolívar, con 164 especies, y Magdalena, con 145.

Según la entomóloga Jennifer Girón Duque [121], la estimación promedio del número de especies para el departamento de Antioquia presenta una posible sobrestimación. Esta observación se sustenta en la comparación con el estado de solamente Paraíba, cuya extensión territorial es comparable a la de Antioquia y donde se han registrado únicamente 236 especies. En contraste, el modelo empleado en este estudio estima un total de 4.210 especies para Antioquia, lo que sugiere una discrepancia considerable entre los datos existentes y los valores proyectados.

Según la entomóloga Claudia Medina Uribe [122], la estimación del valor máximo para la familia Scarabaeidae en Colombia es alta, aunque no descarta que sea posible. Adicionalmente considera que las predicciones están circunscritas a las fuente de datos y al no existir estudios sistemáticos para ningún departamento es complicado dar una opinión más acertada.

Finalmente, se comparó la estimación del número total de especies de coleóptera en Colombia obtenida en este estudio con la última lista taxonómica completa publicada para el país, elaborada por Blackwelder en 1957, la cual registra 6.170 especies [123]. El valor promedio estimado por el modelo empleado en este trabajo asciende a 6.420 especies, lo que representa un incremento del 4% con respecto al listado histórico, No obstante, dado que la referencia de Blackwelder tiene más de cinco décadas de antigüedad, es razonable suponer que dicho valor se encuentra subestimado, especialmente considerando que desde entonces se han descrito numerosas especies adicionales de coleópteros en el país.

Un resumen general de las opiniones de los expertos se puede ver en la tabla 8, donde hay siete opiniones para subestimación, tres para sobrestimación y una ajustada. Es relevante señalar que, entre los expertos consultados, existe un consenso general respecto a la alta incertidumbre asociada a las estimaciones del número de especies, tanto a nivel nacional como departamental. Esta incertidumbre se atribuye a las limitaciones en los inventarios biológicos disponibles, la subrepresentación de ciertas regiones y la continua descripción de nuevas especies en el país.

Tabla 8. Resultado cualitativo de la elicitación experta de seis entomólogos a nivel nacional o regional.

Experto	Taxón y región evaluada	Opinión de la estimación
Juliana Cardona Duque	General, Colombia	Subestimado
Diego Martínez Reveló	Scarabaeinae, Colombia	Subestimado
Diego Martínez Reveló	Scarabaeinae, Antioquia	Subestimado
Larry Jiménez Ferbans	Passalidae, Colombia	Subestimado
Larry Jiménez Ferbans	Passalidae, Antioquia	Sobrestimado
Juan Pablo Botero	General, Colombia	Subestimado
Juan Pablo Botero	Cerambycidae, Colombia	Subestimado
Juan Pablo Botero	Cerambycidae, Antioquia	Subestimado
Jennifer Girón Duque	General, Colombia	Sobrestimado
Jennifer Girón Duque	General, Antioquia	Sobrestimado
Claudia Medina Uribe	Scarabaeidae, Colombia	Sobrestimado
Blackwelder 1945	General, Colombia	Ajustado

A partir del análisis integral de las opiniones expertas, se propone adoptar el valor máximo estimado por el modelo como una cota superior para representar el número potencial de especies de coleópteros a nivel nacional y departamental. Esta elección se fundamenta en

que dicho valor se aproxima con más consistencia a las estimaciones reportadas por los especialistas y proporciona un marco conservador ante el alto grado de incertidumbre existente, tanto en términos taxonómicos como en relación con la cantidad y calidad de datos disponibles. Se presume que el número real de especies se encuentra dentro del intervalo definido por las estimaciones del modelo, con una mayor probabilidad de ubicarse hacia el extremo superior de dicho rango.

10. CONCLUSIONES Y TRABAJOS FUTUROS

10.1 Conclusiones

El modelo desarrollado cumplió con su objetivo principal de predecir el número de especies de coleópteros en el Departamento de Antioquia, utilizando exclusivamente información previamente publicada y de acceso abierto, tanto de registros biológicos como de variables climáticas. Debido a la alta incertidumbre asociada a este tipo de estimaciones, se propuso un rango para la predicción y se consideró que el valor real probablemente se situa en el extremo superior del intervalo. Adicionalmente, se realizó una estimación del número total de especies en Colombia, lo cual permitió establecer comparaciones con los listados taxonómicos existentes y valorar la representatividad del departamento de Antioquia a nivel nacional.

El análisis permitió identificar las principales variables climáticas y de hábitat que influyen en la distribución de los Coleópteros y por ende en el número de especies presentes en una región. Entre estas variables se destacan : velocidad del viento, humedad relativa, precipitación, radiación solar, temperatura, ecosistemas/hábitat, elevación sobre el nivel del mar y especies de plantas cercanas a los coleópteros.

En el proceso de modelado se llevó a cabo de forma iterativa siguiendo la metodología CRISP-DM. En la fase de entendimiento de los datos se recolectó información biológica a través del Sistema de Información sobre Biodiversidad de Colombia (SiB Colombia) y variables ambientales del IDEAM, complementadas con fuentes internacionales. La etapa de preparación de datos se enfocó en la depuración de las variables sin relevancia biológica, integración de bases de datos, tratamiento de valores faltantes y codificación de las variables categóricas, con el fin de preparar el conjunto de datos entrenamiento de modelos de regresión.

Para el modelado, se emplearon técnicas comúnmente utilizadas en tareas de regresión, incluyendo regresión lineal, perceptrón multicapa, redes neuronales profundas y Random Forest, estas fueron implementadas en Python.

El desempeño de los modelos fue evaluado mediante las métricas numéricas R2, MAE y MSE. Con base en estos indicadores, se seleccionó la red neuronal profunda con escalamiento como el modelo con mejor rendimiento que fue sometida a una validación

cruzada de 10, lo cual permitió verificar su robustez.

Posteriormente, se realizaron predicciones tanto a nivel nacional como departamental, las cuales fueron sometidas a un proceso de elicitación experta. En este ejercicio, coleopterólogos evaluaron cualitativamente las predicciones del modelo y, en la mayoría de los casos, consideraron que los valores obtenidos representaban una subestimación del número real de especies.

10.2 Trabajos futuros

A partir del modelo desarrollado, se prevé la posibilidad de replicar esta metodología para otros departamentos del país, con el fin de generar estimaciones comparables del número de especies de coleóptera. Estos resultados podrían ser integrados a la plataforma Biodiversidad en Cifras: https://cifras.biodiversidad.co/, fortaleciendo así el acceso a datos cuantitativos sobre la diversidad biológica a nivel nacional. Para ello será necesario realizar ajustes iterativos al modelo, tanto en la selección de variables como en la incorporación de nuevos datos o mapas que se generen en los próximos años, con el objetivo de mejorar su precisión y aplicabilidad

Asimismo, se contempla la extensión de este enfoque de modelado a otros grupos de insectos para los cuales no existen actualmente estimaciones del número de especies a escala departamental. Entre estos se incluyen órdenes relevantes de la clase Insecta como Hymenoptera, Hemiptera y Diptera. Esta aplicación permitirá avanzar en la caracterización de la biodiversidad insectil del país mediante herramientas analíticas consistentes y reproducibles.

En el mediano y largo plazo, se proyecta adaptar esta metodología a otros grupos biológicos con alta diversidad y escasa información como hongos y bacterias. Sin embargo, para estos organismos será indispensable rediseñar el conjunto de variables predictoras, priorizando aquellas de naturaleza abiótica, dada la diferencia sustancial en sus patrones de distribución y requerimientos ecológicos. Esta expansión metodológica representa un paso estratégico hacia una comprensión más integral de la biodiversidad colombiana y su distribución espacial.

11. BIBLIOGRAFÍA

[1] J. C. Girón *et al.*, “Consideraciones sobre el estado del conocimiento de la diversidad de Coleoptera (Arthropoda: Insecta) en Colombia,” *Revista Colombiana de Entomología*, vol. 47, no. 2, Art. no. 2, Jul. 2021, doi: 10.25100/socolen.v47i2.10717.

[2] E. Arbeláez Cortés[1], “DESCRIBIENDO ESPECIES: UN PANORAMA DE LA BIODIVERSIDAD COLOMBIANA EN EL ÁMBITO MUNDIAL,” *Acta Biológica Colombiana*, vol. 18, no. 1, pp. 165–178, Apr. 2013, Accessed: Jun. 01, 2025. [Online]. Available: http://www.scielo.org.co/scielo.php?script=sci_abstract&pid=S0120-548X2013000100012&lng

=en&nrm=iso&tlng=es
[3] SiB Colombia, "Biodiversidad en Cifras 2024." [Online]. Available: https://biodiversidad.co/cifras
[4] S. A. Slipinski, R. a. B. Leschen, and J. F. Lawrence, "Order Coleoptera Linnaeus, 1758. In: Zhang, Z.-Q. (Ed.) Animal biodiversity: An outline of higher-level classification and survey of taxonomic richness," *Zootaxa*, vol. 3148, no. 1, Art. no. 1, Dec. 2011, doi: 10.11646/zootaxa.3148.1.39.
[5] P. Eggleton, "The State of the World's Insects," *Annual Review of Environment and Resources*, vol. 45, no. Volume 45, 2020, pp. 61–82, Oct. 2020, doi: 10.1146/annurev-environ-012420-050035.
[6] H. B. Salazar, S. M. B. Burgos, J. R. C. Chinome, J. R. Piñeros, M. P. Mendieta, and G. L. G. Gómez, "Diversidad de Coleópteros en un bosque alto andino del municipio de Santa Rosa de Viterbo (Boyacá)," *Revista Mutis*, vol. 6, no. 2, Art. no. 2, Oct. 2016, doi: 10.21789/22561498.1149.
[7] Germán. Amat García, M. G. Andrade-C., and E. Amat-García, *Libro rojo de los invertebrados terrestres de Colombia*, 1. ed. Bogotá: Conservación Internacional Colombia : Universidad Nacional de Colombia, Sede Bogotá, Facultad de Ciencias Naturales, Instituto de Ciencias Naturales, 2007.
[8] R. K. Colwell and J. A. Coddington, "Estimating terrestrial biodiversity through extrapolation," *Phil. Trans. R. Soc. Lond. B*, vol. 345, no. 1311, pp. 101–118, Jul. 1994, doi: 10.1098/rstb.1994.0091.
[9] C. Botella, A. Joly, P. Bonnet, P. Monestiez, and F. Munoz, "A Deep Learning Approach to Species Distribution Modelling," in *Multimedia Tools and Applications for Environmental & Biodiversity Informatics*, A. Joly, S. Vrochidis, K. Karatzas, A. Karppinen, and P. Bonnet, Eds., Cham: Springer International Publishing, 2018, pp. 169–199. doi: 10.1007/978-3-319-76445-0_10.
[10] M. Koivula, "Useful model organisms, indicators, or both? Ground beetles (Coleoptera, Carabidae) reflecting environmental conditions," *ZooKeys*, vol. 100, pp. 287–317, May 2011, doi: 10.3897/zookeys.100.1533.
[11] Y. Chauvier *et al.*, "Influence of climate, soil, and land cover on plant species distribution in the European Alps," *Ecological Monographs*, vol. 91, no. 2, p. e01433, 2021, doi: 10.1002/ecm.1433.
[12] A. D. Chapman, *Uses of Primary Species-Occurrence Data*. Copenhagen: Report for the Global Biodiversity Information Facility, 2005.
[13] T. Robertson *et al.*, "The GBIF Integrated Publishing Toolkit: Facilitating the Efficient Publishing of Biodiversity Data on the Internet," *PLoS ONE*, vol. 9, no. 8, p. e102623, Aug. 2014, doi: 10.1371/journal.pone.0102623.
[14] SiB Colombia, "¿Qué es el SiB Colombia? - SiB Colombia." Accessed: May 12, 2025. [Online]. Available: https://biodiversidad.co/acercade/sib-colombia/
[15] P. Chapman, R. Kerber, J. Clinton, T. Khabaza, T. Reinartz, and R. Wirth, "The CRISP-DM Process Model," *Discussion Paper*, Mar. 1999.
[16] IBM, "IBM SPSS Modeler CRISP-DM Guide," Jun. 2021, doi: https://www.ibm.com/docs/it/SS3RA7_18.3.0/pdf/ModelerCRISPDM.pdf.
[17] C. Sammut and G. I. Webb, *Encyclopedia of Machine Learning and Data Mining*, 2nd ed. Springer, 2017. [Online]. Available: 10.1007/978-1-4899-7687-1
[18] A. Géron, *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow*, 3rd ed. United States of America: O'Reilly Media, Inc., 2023. Accessed: May 14, 2025. [Online]. Available: https://www.oreilly.com/library/view/hands-on-machine-learning/9781492032632/
[19] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, Oct. 2001, doi: 10.1023/A:1010933404324.
[20] A. Ivakhnenko and V. Lapa, *Cybernetic Predicting Devices*. New York: CCM Information Corporation, 1965.
[21] V. Sze, Y.-H. Chen, T.-J. Yang, and J. S. Emer, "Efficient Processing of Deep Neural Networks: A

Tutorial and Survey," *Proceedings of the IEEE*, vol. 105, no. 12, pp. 2295–2329, Dec. 2017, doi: 10.1109/JPROC.2017.2761740.

[22] G. Alam, I. Ihsanullah, Mu. Naushad, and M. Sillanpää, "Applications of artificial intelligence in water treatment for optimization and automation of adsorption processes: Recent advances and prospects," *Chemical Engineering Journal*, vol. 427, p. 130011, Jan. 2022, doi: 10.1016/j.cej.2021.130011.

[23] D. Chicco, M. J. Warrens, and G. Jurman, "The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation," *PeerJ Comput. Sci.*, vol. 7, p. e623, Jul. 2021, doi: 10.7717/peerj-cs.623.

[24] S. Wright, "Correlation and causation," *Journal of agricultural research*, vol. 20, no. 7, p. 557, 1921, Accessed: May 13, 2025. [Online]. Available: https://cir.nii.ac.jp/crid/1370567187556110595

[25] A. R. Colson and R. M. Cooke, "Expert Elicitation: Using the Classical Model to Validate Experts' Judgments," *Review of Environmental Economics and Policy*, vol. 12, no. 1, pp. 113–132, Jan. 2018, doi: 10.1093/reep/rex022.

[26] P. Brun *et al.*, "Multispecies deep learning using citizen science data produces more informative plant community models," *Nat Commun*, vol. 15, no. 1, p. 4421, May 2024, doi: 10.1038/s41467-024-48559-9.

[27] F. M. Alfaro, J. Pizarro-Araya, F. M. Alfaro, and J. Pizarro-Araya, "Estimation of the richness of epigean coleopterans of the Pingüino of Humboldt National Reserve (Atacama And Coquimbo Regions, Chile)," *Gayana (Concepción)*, vol. 81, no. 2, pp. 39–51, Dec. 2017, doi: 10.4067/S0717-65382017000200039.

[28] S.-H. Hong, J. Bunge, S.-O. Jeon, and S. S. Epstein, "Predicting microbial species richness," *Proceedings of the National Academy of Sciences*, vol. 103, no. 1, pp. 117–122, Jan. 2006, doi: 10.1073/pnas.0507245102.

[29] L. Jiménez-Ferbans, G. Amat-García, and P. Reyes-Castillo, "Diversity and distribution patterns of Passalidae (Coleoptera Scarabaeoidea) in the Caribbean Region of Colombia," *Tropical Zoology*, vol. 23, pp. 147–164, 2010.

[30] D. OBREGÓN-CORREDOR, F. J. HERNÁNDEZ-GUZMÁN, and D. K. RIOS-MOYANO, "Efecto de los factores climáticos, variedades y densidades de siembra en la dinámica de artrópodos en cultivos de arroz en Yopal-Casanare, Colombia," *Revista Colombiana de Entomología*, vol. 47, no. 1, p. e9364, 2021, Accessed: Apr. 04, 2025. [Online]. Available: https://doi.org/10.25100/socolen.v47i1.9364

[31] E. C. Rubio and J. M. Lobo, "DISTRIBUCIÓN CONOCIDA Y POTENCIAL DE LAS ESPECIES DEL GÉNERO EURYSTERNUS DALMAN, 1824 (COLEOPTERA: SCARABAEIDAE) DE COLOMBIA," *Boletín de la Sociedad Entomológica Aragonesa*, vol. 47, pp. 254–264, 2010.

[32] J. Cheng, P. Li, Y. Zhang, Y. Zhan, and Y. Liu, "Quantitative assessment of the contribution of environmental factors to divergent population trends in two lady beetles," *Biological Control*, vol. 145, p. 104259, Jun. 2020, doi: 10.1016/j.biocontrol.2020.104259.

[33] L. S. Brasil *et al.*, "Net primary productivity and seasonality of temperature and precipitation are predictors of the species richness of the Damselflies in the Amazon," *Basic and Applied Ecology*, vol. 35, pp. 45–53, Mar. 2019, doi: 10.1016/j.baae.2019.01.001.

[34] D. M. Hincapié-Montoya, L. M. González-Rodríguez, and M. González-Córdoba, "Estado actual del conocimiento de los coleópteros acuáticos en Colombia," *Caldasia*, vol. 45, no. 3, Art. no. 3, Jul. 2023, doi: 10.15446/caldasia.v45n3.104455.

[35] M. P. G. Gonçalves, "Beetles and Meteorological Conditions: A Case Study," in *Agrometeorology*, IntechOpen, 2020. doi: 10.5772/intechopen.94517.

[36] Jorge M. Lobo, J. Lumaret, and P. Jay-Robert, "Modelling the species richness distribution of French dung beetles (Coleoptera, Scarabaeidae) and delimiting the predictive capacity of different groups of explanatory variables," *Global Ecology and Biogeography*, vol. 11, no. 4, pp. 265–277, Jul. 2002, doi: 10.1046/j.1466-822X.2002.00291.x.

[37] J. Cardona-Duque, "Variables importantes para la distribución de Coleoptera," 2024.

[38] J. E. Pasek, "Influence of wind and windbreaks on local dispersal of insects," *Agriculture, Ecosystems & Environment*, vol. 22–23, pp. 539–554, Aug. 1988, doi: 10.1016/0167-8809(88)90044-8.
[39] R. R. Angela Maria, "Velocidad del Viento más Probable a 10 metros de altura Mensual durante el periodo 2000-2010. República de Colombia. Año 2015." Accessed: Apr. 03, 2025. [Online]. Available: https://visualizador.ideam.gov.co/geonetwork/srv/spa/catalog.search#/metadata/6b916fb1-67cb-4c42-a7c6-6080a92e592b
[40] IDEAM *et al.*, "Ecosistemas Continentales, Costeros y Marinos de Colombia. Escala 1:100.000. versión 2.1.Año 2017." Accessed: Apr. 03, 2025. [Online]. Available: https://visualizador.ideam.gov.co/geonetwork/srv/spa/catalog.search#/metadata/0684d637-5b6a-40e8-80f4-bdf915b3e3da
[41] A. M. Ruiz Rotta, "Humedad Relativa Anual promedio Multianual durante el periodo 1981-2010. República de Colombia. Año 2014." Accessed: Apr. 10, 2025. [Online]. Available: https://visualizador.ideam.gov.co/geonetwork/srv/spa/catalog.search#/metadata/048126d7-0ba7-4bd6-bd1c-3e475385e377
[42] IDEAM, "Precipitación para Colombia (mm) 1976-2005. Escala 1:100,000. Año 2015." Accessed: Apr. 10, 2025. [Online]. Available: https://visualizador.ideam.gov.co/geonetwork/srv/spa/catalog.search#/metadata/b2771c78-47eb-4b25-ad64-20a6f992e410
[43] J. Albarracin, "Radiación global media recibida en una superficie horizontal durante el día promedio anual multianual (KWH/M2). República de Colombia." Accessed: Apr. 10, 2025. [Online]. Available: https://visualizador.ideam.gov.co/geonetwork/srv/spa/catalog.search#/metadata/c2d36ff5-41de-47ff-a866-8a60932b8a31
[44] A. M. Ruiz Rotta, "Temperatura Mínima Mensual Promedio Multianual durante el periodo 1981-2010. República de Colombia. Año 2014." Accessed: Apr. 10, 2025. [Online]. Available: https://visualizador.ideam.gov.co/geonetwork/srv/spa/catalog.search#/metadata/7123a846-33bf-4831-b51c-2d11f629905f
[45] R. R. Angela Maria, "Temperatura Media Mensual Promedio Multianual durante el periodo 1981-2010. República de Colombia. Año 2014." Accessed: Apr. 10, 2025. [Online]. Available: https://visualizador.ideam.gov.co/geonetwork/srv/spa/catalog.search#/metadata/5f504388-2ad8-404c-a36f-e2b023357589
[46] R. R. Angela Maria, "Temperatura Máxima Mensual Promedio Multianual durante el periodo 1981-2010." Accessed: Apr. 10, 2025. [Online]. Available: https://visualizador.ideam.gov.co/geonetwork/srv/spa/catalog.search#/metadata/378ab856-167f-4d01-813a-f8b9e74934a5
[47] NASA/METI/AIST/Japan Spacesystems And U.S./Japan ASTER Science Team, "ASTER Global Digital Elevation Model V003." NASA Land Processes Distributed Active Archive Center, 2019. doi: 10.5067/ASTER/ASTGTM.003.
[48] DANE, "Nivel geográfico de Departamentos del Marco Geoestadístico Nacional (MGN) versión 2020." Accessed: Apr. 04, 2025. [Online]. Available: https://www.arcgis.com/home/webmap/viewer.html?url=https%3A%2F%2Fportalgis.dane.gov.co%2Fmparcgis%2Frest%2Fservices%2FMGN2020%2FServ_CapaDepartamentos_2020%2FMapServer&source=sd
[49] GBIF.org, "GBIF Occurrence Download, Plantae Colombia," https://www.gbif.org/. Accessed: Apr. 03, 2025. [Online]. Available: https://doi.org/10.15468/dl.j6xumw
[50] GBIF.org, "GBIF Occurrence Download." Accessed: Apr. 03, 2025. [Online]. Available: https://doi.org/10.15468/dl.retzxm
[51] J. C. Girón Duque, "Listado de las especies de Brachycerinae (Coleoptera: Curculionidae) de Colombia." Accessed: Apr. 03, 2025. [Online]. Available: https://doi.org/10.15472/suwoq6
[52] J. Utria Suarez and A. Rendón Ramírez, "Listado de las especies de Bruchinae (Coleoptera:

Chrysomelidae) de Colombia." Accessed: Apr. 03, 2025. [Online]. Available: https://doi.org/10.15472/tvuxwk
[53] E. Arriaga-Varela and W. D. Rodríguez, "Listado de las especies de Anamorphidae (Coleoptera: Coccinelloidea) de Colombia." Accessed: Apr. 03, 2025. [Online]. Available: https://doi.org/10.15472/hxxkei
[54] D. M. Hincapié-Montoya, "Listado de las especies de Dytiscidae (Coleoptera: Adephaga) de Colombia." Accessed: Apr. 03, 2025. [Online]. Available: https://doi.org/10.15472/m1c5nc
[55] J. P. Botero Rodríguez and G. M. Rodríguez-Mirón, "Listado de las especies de Megalopodidae (Coleoptera: Chrysomeloidea) de Colombia." Accessed: Apr. 03, 2025. [Online]. Available: https://doi.org/10.15472/niicpm
[56] J. S. Dueñas Cáceres, J. Clavijo Bustos, and I. C. Ríos Málaver, "Listado de las especies de Lucanidae (Coleoptera: Scarabaeoidea) de Colombia." Accessed: Apr. 03, 2025. [Online]. Available: https://doi.org/10.15472/p9fc2m
[57] D. M. Hincapié-Montoya and J. C. Girón Duque, "Listado de las especies de Gyrinidae (Coleoptera: Adephaga) de Colombia." Accessed: Apr. 03, 2025. [Online]. Available: https://doi.org/10.15472/aeddqk
[58] J. S. Palacios Rodríguez, "Listado de las especies de Eumolpinae (Coleoptera: Chrysomelidae) de Colombia." Accessed: Apr. 09, 2025. [Online]. Available: https://doi.org/10.15472/off6un
[59] D. Hincapié-Montoya, "Listado de las especies de Hydraenidae (Coleoptera: Staphylinoidea) de Colombia." Accessed: Apr. 09, 2025. [Online]. Available: https://doi.org/10.15472/33qho2
[60] O. Ascuntar-Osnas, P. López-Bedoya, M. A. Johnston, and J. C. Girón Duque, "Listado de las especies de Tenebrionidae (Coleoptera: Tenebrionoidea) de Colombia." Accessed: Apr. 09, 2025. [Online]. Available: https://doi.org/10.15472/3bphqb
[61] D. M. Hincapié-Montoya and M. González-Córdoba, "Listado de las especies de Elmidae (Coleoptera: Byrrhoidea) de Colombia." Accessed: Apr. 09, 2025. [Online]. Available: https://doi.org/10.15472/8lqsij
[62] K. García, D. Uchima Taborda, and L. J. Migliore, "Listado de las especies de Buprestidae (Coleoptera: Buprestoidea) de Colombia." Accessed: Apr. 09, 2025. [Online]. Available: https://doi.org/10.15472/g5u6tg
[63] M. F. Bermúdez-Higinio and J. C. Girón Duque, "Listado de las especies de Dryophthorinae (Coleoptera: Curculionidae) de Colombia." Accessed: Apr. 09, 2025. [Online]. Available: https://doi.org/10.15472/wtd0tk
[64] D. M. Hincapié-Montoya and M. González-Córdoba, "Listado de las especies de Ptilodactylidae (Coleoptera: Byrrhoidea) de Colombia." Accessed: Apr. 09, 2025. [Online]. Available: https://doi.org/10.15472/5fywbh
[65] J. C. Girón Duque and M. Barros-Barrios, "Listado de las especies de Conoderinae (Coleoptera: Curculionidae) de Colombia." Accessed: Apr. 09, 2025. [Online]. Available: https://doi.org/10.15472/yifvet
[66] D. E. Martínez-Revelo, "Listado de las especies de Meloidae (Coleoptera: Tenebrionoidea) de Colombia." Accessed: Apr. 09, 2025. [Online]. Available: https://doi.org/10.15472/9ixuud
[67] D. M. Hincapié-Montoya and M. González-Córdoba, "Listado de las especies de Cneoglossidae (Coleoptera: Byrrhoidea) de Colombia." Accessed: Apr. 09, 2025. [Online]. Available: https://doi.org/10.15472/rsgvlo
[68] D. M. Hincapié-Montoya and M. González-Córdoba, "Listado de las especies de Dryopidae (Coleoptera: Byrrhoidea) de Colombia." Accessed: Apr. 09, 2025. [Online]. Available: https://ipt.biodiversidad.co/sib/resource?r=dryopidae_colombia
[69] D. M. Hincapié-Montoya and M. González-Córdoba, "Listado de las especies de Chelonariidae (Coleoptera: Byrrhoidea) de Colombia." Accessed: Apr. 09, 2025. [Online]. Available: https://doi.org/10.15472/kxvmlk
[70] D. M. Hincapié-Montoya and M. González-Córdoba, "Listado de las especies de Lutrochidae (Coleoptera: Byrrhoidea) de Colombia." Accessed: Apr. 09, 2025. [Online]. Available: https://doi.org/10.15472/5w8sxt

[71] D. E. Martínez-Revelo, “Listado de las especies de Mordellidae (Coleoptera: Tenebrionoidea) de Colombia.” Accessed: Apr. 09, 2025. [Online]. Available: https://doi.org/10.15472/abvxw3
[72] D. M. Hincapié-Montoya and M. González-Córdoba, “Listado de las especies de Limnichidae (Coleoptera: Byrrhoidea) de Colombia.” Accessed: Apr. 09, 2025. [Online]. Available: https://doi.org/10.15472/fmeri8
[73] L. M. González-Rodríguez, “Listado de las especies de Georissidae (Coleoptera: Hydrophiloidea) de Colombia.” Accessed: Apr. 09, 2025. [Online]. Available: https://doi.org/10.15472/bdcxbe
[74] J. C. Girón Duque, “Listado de las especies de Anthribidae (Coleoptera: Curculionoidea) de Colombia.” Accessed: Apr. 09, 2025. [Online]. Available: https://doi.org/10.15472/tolans
[75] L. M. González-Rodríguez, “Listado de las especies de Hydrochidae (Coleoptera: Hydrophiloidea) de Colombia.” Accessed: Apr. 09, 2025. [Online]. Available: https://doi.org/10.15472/3uxhlf
[76] J. C. Girón Duque, “Listado de las especies de Brentidae (Coleoptera: Curculionoidea) de Colombia.” Accessed: Apr. 09, 2025. [Online]. Available: https://doi.org/10.15472/iyrpyc
[77] J. C. Girón Duque, “Listado de las especies de Cyclominae (Coleoptera: Curculionidae) de Colombia.” Accessed: Apr. 10, 2025. [Online]. Available: https://doi.org/10.15472/mv3ql4
[78] J. C. Girón Duque, “Listado de las especies de Dermestidae (Coleoptera: Bostrichoidea) de Colombia.” Accessed: Apr. 10, 2025. [Online]. Available: https://doi.org/10.15472/pisuan
[79] D. Hincapié-Montoya and M. González-Córdoba, “Listado de las especies de Callirhipidae (Coleoptera: Byrrhoidea) de Colombia.” Accessed: Apr. 10, 2025. [Online]. Available: https://doi.org/10.15472/elddwn
[80] J. C. Girón Duque, I. T. Cardenas Espitia, and D. Hincapié-Montoya, “Listado de las especies de Noteridae (Coleoptera: Adephaga) de Colombia.” Accessed: Apr. 10, 2025. [Online]. Available: https://doi.org/10.15472/rwfj6h
[81] D. M. Hincapié-Montoya and M. González-Córdoba, “Listado de las especies de Psephenidae (Coleoptera: Byrrhoidea) de Colombia.” Accessed: Apr. 10, 2025. [Online]. Available: https://doi.org/10.15472/6mzbwp
[82] J. C. Girón Duque and M. Geiser, “Listado de las especies de Criocerinae (Coleoptera: Chrysomelidae) de Colombia.” Accessed: Apr. 10, 2025. [Online]. Available: https://doi.org/10.15472/cnxkgm
[83] Y. P. Sissa-Dueñas, “Listado de las especies de Anthicidae (Coleoptera: Tenebrionoidea) de Colombia.” Accessed: Apr. 10, 2025. [Online]. Available: https://ipt.biodiversidad.co/sib/resource?r=anth
[84] C. A. Taboada-Verona, “Listado de las especies de Scirtidae (Coleoptera: Scirtoidea) de Colombia.” Accessed: Apr. 10, 2025. [Online]. Available: https://doi.org/10.15472/bw5o1u
[85] Y. P. Sissa-Dueñas, “Listado de las especies de Melandryidae (Coleoptera: Tenebrionoidea) de Colombia.” Accessed: Apr. 10, 2025. [Online]. Available: https://doi.org/10.15472/3eiijl
[86] Y. P. Sissa-Dueñas, “Listado de las especies de Archeocrypticidae (Coleoptera: Tenebrionoidea) de Colombia.” Accessed: Apr. 10, 2025. [Online]. Available: https://doi.org/10.15472/vfcq3h
[87] Y. P. Sissa-Dueñas, “Listado de las especies de Aderidae (Coleoptera: Tenebrionoidea) de Colombia.” Accessed: Apr. 10, 2025. [Online]. Available: https://doi.org/10.15472/bo97af
[88] J. C. Girón Duque, “Listado de las especies de Ceutorhynchinae (Coleoptera: Curculionidae) de Colombia.” Accessed: Apr. 10, 2025. [Online]. Available: https://ipt.biodiversidad.co/sib/resource?r=ceut
[89] E. A. Nascimento, “Listado de las especies de Lycidae (Coleoptera: Elateroidea) de Colombia.” Accessed: Apr. 10, 2025. [Online]. Available: https://doi.org/10.15472/iftego
[90] U. T. Diego Alejandro, “Listado de las especies de Phengodidae (Coleoptera: Elateroidea) de Colombia.” Accessed: Apr. 10, 2025. [Online]. Available: https://doi.org/10.15472/1bxorj
[91] J. Clavijo-Bustos, H. J. Gasca Álvarez, and J. C. Neita Moreno, “Listado de las especies de Orphninae (Coleoptera: Scarabaeidae) de Colombia.” Accessed: Apr. 10, 2025. [Online]. Available: https://doi.org/10.15472/0msljd

[92] J. Clavijo Bustos and L. M. González-Rodríguez, "Listado de las especies de Epimetopidae (Coleoptera: Hydrophiloidea) de Colombia." Accessed: Apr. 10, 2025. [Online]. Available: https://doi.org/10.15472/wqyqbs
[93] A. Ladino-Peñuela, "Listado de las especies de Lampyridae (Coleoptera: Elateroidea) de Colombia." Accessed: Apr. 10, 2025. [Online]. Available: https://doi.org/10.15472/ii4qjk
[94] I. T. Cardenas Espitia, "Listado de las especies de Haliplidae (Coleoptera: Adephaga) de Colombia." Accessed: Apr. 10, 2025. [Online]. Available: https://doi.org/10.15472/pgwexm
[95] J. P. Botero and G. Biffi, "Listado de las especies de Cantharidae (Coleoptera: Elateroidea) de Colombia." Accessed: Apr. 10, 2025. [Online]. Available: https://doi.org/10.15472/l1vnio
[96] L. M. González-Rodríguez, "Listado de las especies de Hydrophilidae (Coleoptera: Hydrophiloidea) de Colombia." Accessed: Apr. 10, 2025. [Online]. Available: https://doi.org/10.15472/l4lymo
[97] J. C. Girón Duque and J. Cardona-Duque, "Listado de las especies de Cossoninae (Coleoptera: Curculionidae) de Colombia." Accessed: Apr. 10, 2025. [Online]. Available: https://doi.org/10.15472/9mqxev
[98] J. Clavijo Bustos, "Listado de las especies de Heteroceridae (Coleoptera: Byrrhoidea) de Colombia." Accessed: Apr. 10, 2025. [Online]. Available: https://doi.org/10.15472/4ymj2e
[99] J. Clavijo Bustos, "Listado de las especies de Ochodaeidae (Coleoptera: Scarabaeoidea) de Colombia." Accessed: Apr. 10, 2025. [Online]. Available: https://doi.org/10.15472/mew1hc
[100] M. I. Salinas Cano, J. C. Girón Duque, and P. Gnaspini, "Listado de las especies de Leiodidae (Coleoptera: Staphylinoidea) de Colombia." Accessed: Apr. 10, 2025. [Online]. Available: https://doi.org/10.15472/6xhtdd
[101] J. P. Botero, "Listado de las especies de Disteniidae (Coleoptera: Chrysomeloidea) de Colombia." Accessed: Apr. 10, 2025. [Online]. Available: https://doi.org/10.15472/r77dxh
[102] J. C. Girón Duque, "Listado de las especies de Entiminae (Coleoptera: Curculionidae) de Colombia." Accessed: Apr. 10, 2025. [Online]. Available: https://doi.org/10.15472/jdwfao
[103] Coleoptera de Colombia, "Conjuntos de datos Coleoptera de Colombia." Accessed: May 04, 2025. [Online]. Available: https://www.gbif.org/dataset/search?publishing_org=2c39be5c-c11e-46d0-bcb4-552f2072d19f
[104] M. D. Eyre, S. P. Rushton, M. L. Luff, and M. G. Telfer, "Predicting the distribution of ground beetle species (Coleoptera, Carabidae) in Britain using land cover variables," *Journal of Environmental Management*, vol. 72, no. 3, pp. 163–174, Sep. 2004, doi: 10.1016/j.jenvman.2004.04.007.
[105] J. Muro *et al.*, "Predicting plant biomass and species richness in temperate grasslands across regions, time, and land management with remote sensing and deep learning," *Remote Sensing of Environment*, vol. 282, p. 113262, Dec. 2022, doi: 10.1016/j.rse.2022.113262.
[106] F. T. Breiner, A. Guisan, A. Bergamini, and M. P. Nobis, "Overcoming limitations of modelling rare species by using ensembles of small models," *Methods in Ecology and Evolution*, vol. 6, no. 10, pp. 1210–1218, 2015, doi: 10.1111/2041-210X.12403.
[107] G. J. Chang, "Biodiversity estimation by environment drivers using machine/deep learning for ecological management," *Ecological Informatics*, vol. 78, p. 102319, Dec. 2023, doi: 10.1016/j.ecoinf.2023.102319.
[108] J. Růžičková and M. Veselý, "Using radio telemetry to track ground beetles: Movement of Carabus ullrichii," *Biologia*, vol. 71, no. 8, pp. 924–930, Aug. 2016, doi: 10.1515/biolog-2016-0108.
[109] L. Drag and L. Cizek, "Radio-Tracking Suggests High Dispersal Ability of the Great Capricorn Beetle (Cerambyx cerdo)," *J Insect Behav*, vol. 31, no. 2, pp. 138–143, Mar. 2018, doi: 10.1007/s10905-018-9669-x.
[110] M. T. Smith, J. Bancroft, G. H. Li, R. T. Gao, and S. Teale, "Dispersal of Anoplophora glabripennis (Cerambycidae)," *Environmental entomology*, vol. 30, no. 6, pp. 1036–1040, 2001, Accessed: Apr. 16, 2025. [Online]. Available: https://doi.org/10.1603/0046-225X-30.6.1036

[111] N. M. Kalberer, T. C. J. Turlings, and M. Rahier, “An alternative hibernation strategy involving sun- exposed ‘hotspots’, dispersal by flight, and host plant finding by olfaction in an alpine leaf beetle,” *Entomologia Experimentalis et Applicata*, vol. 114, pp. 189–196.

[112] K. D. Chase, D. Kelly, A. M. Liebhold, M. K.-F. Bader, and E. G. Brockerhoff, “Long-distance dispersal of non-native pine bark beetles from host resources,” *Ecological Entomology*, vol. 42, no. 2, pp. 173–183, 2017, doi: 10.1111/een.12371.

[113] B. Cornelissen *et al.*, “The small hive beetle’s capacity to disperse over long distances by flight,” *Sci Rep*, vol. 14, no. 1, p. 14859, Jun. 2024, doi: 10.1038/s41598-024-65434-1.

[114] J. F. Lawrence and A. F. Newton, “Families and subfamilies of Coleoptera (with selected genera, notes, references and data on family-group names),” in *Biology, Phylogeny, and Classification of Coleoptera: Papers Celebrating the 80th Birthday of Roy A. Crowson*, Warszawa, 1995, pp. 779–1006.

[115] SiB Colombia, “Biodiversidad en cifras 2018.” Accessed: Jun. 08, 2025. [Online]. Available: https://web.archive.org/web/20180117231846/https://cifras.biodiversidad.co/

[116] RNC, “Registro Único Nacional de Colecciones Biológicas - Especialista Curador.” Accessed: May 13, 2025. [Online]. Available: http://rnc.humboldt.org.co/admin/index.php/Curador/Especialista

[117] J. Cardona Duque, “Opinión sobre el número de especies estimadas registradas en Colombia. Comunicación privada,” 2025.

[118] D. E. Martínez Revelo, “Opinión sobre el número de especies estimadas de la familia Scarabaeinae. Comunicación privada,” 2025.

[119] L. Jiménez Ferbans, “Opinión sobre el número de especies estimadas de la familia Passalidae. Comunicación privada,” 2025.

[120] J. P. Botero R, “Opinión sobre el número de especies estimadas de la familia Cerambycidae. Comunicación privada,” 2025.

[121] J. C. Girón Duque, “Opinión sobre el número de especies estimadas registradas en Antioquia. Comunicación privada,” 2025.

[122] C. A. Medina Uribe, “Opinión sobre el número de especies estimadas de la familia Scarabaeidae. Comunicación privada,” 2025.

[123] R. E. Blackwelder, “Checklist of the coleopterous insects of Mexico, Central America, the West Indies, and South America, pt. 3,” *Bulletin of the United States National Museum*, vol. 185, no. i–iv, pp. 343–550, 1945, doi: https://doi.org/10.5479/si.03629236.185.3.