

# Modelo de predicción de nicho de especies utilizando técnicas de machine learning

Daniel Rojas Díaz, John Esteban Castro Ramírez, Luis Miguel Caicedo Jimenez

Estadística Multivariada, Departamento de Ciencias Matemáticas, Escuela de Ciencias Aplicadas e Ingeniería  
Medellín, Colombia

*drojasd@eafit.edu.co, jecastror@eafit.edu.co, lmcaicedoj@eafit.edu.co*

## 1. Pregunta de investigación

Los modelos de distribución de especies (SDMs) son herramientas matemáticas que permiten inferir los patrones de ocurrencia de cierta especie a nivel espacial a partir de información sobre las variables que describen el hábitat donde ésta ha sido observada (Beery et al., 2021). Por tal razón, los SDMs han ganado relevancia durante los últimos años en el campo de la ecología, especialmente en las áreas relacionadas con la conservación, el estudio de especies invasoras y el estudio de posibles efectos del cambio climático (Walther et al., 2009; Thuiller et al., 2005). Actualmente, existe una preponderancia de modelos estadísticos paramétricos en el área de los SDMs Fitzpatrick, Gotelli y Ellison, 2013, sin embargo, se ha visto un creciente interés por el uso de técnicas provenientes del machine learning y el deep learning debido a la creciente información disponible y la necesidad de extraer características cada vez más complejas a partir de los datos de presencia de cierta especie o especies (Rew, Cho y Hwang, 2021; Estopinan et al., 2022; Pichler y Hartig, 2023).

En general, puede decirse que el desempeño de los algoritmos para SMDs pueden medirse de acuerdo a su capacidad para reproducir correctamente los patrones de ocurrencia de las especies de las distintas especies (nicho) a partir de algunas observaciones (Austin et al., 2006), sin embargo, la información del nicho real de la especie es, en general, desconocida y por tanto se debe recurrir a la generación de nichos artificiales y especies virtuales (Meynard, Leroy y Kaplan, 2019). En el presente trabajo se plantea valorar el potencial de distintos algoritmos de aprendizaje automático para inferir correctamente el nicho de una especie virtual compleja con respecto al desempeño de una técnica tradicional de SDMs.

## 2. Objetivos

### 2.1. Objetivo general

Evaluar el desempeño de distintos algoritmos de machine learning, con respecto al algoritmo del estado del arte Biomat-Niche Modeling (BNM), para determinar el nicho real de una especie a partir de limitadas observaciones de campo contribuyendo así a una gestión de los recursos naturales y a una planificación de la conservación de la biodiversidad más efectiva.

### 2.2. Objetivos específicos

- Identificar diferentes algoritmos de aprendizaje automático e inteligencia artificial cuya estructura y enfoque sean adecuados para predecir el nicho de una especie dada la naturaleza del problema.
- Evaluar la precisión de los algoritmos para predecir el nicho de una especie utilizando un conjunto de datos de entrenamiento y validación generados con ayuda de un generador de nichos virtual.
- Identificar las variables ambientales más importantes para predecir el nicho de una especie y determinar cómo afectan a la distribución de la misma.

### 3. Revisión de literatura

Los modelos de distribución de especies (SDMs) han ganado considerable atención en el campo de la ecología y la biología de la conservación durante los últimos 20 años (Botella et al., 2018; Bauer et al., 2019). Los SDMs constituyen valiosas herramientas que permiten predecir la distribución potencial de una especie a través de diferentes terrenos representados mediante variables ambientales y espaciales, permitiendo además una ganancia en la comprensión de las condiciones espaciales y ambientales que influyen en la ocurrencia de dicha especie (Seo et al., 2021). De manera concreta, el objetivo de los SDMs es inferir la distribución espacial de cierta especie con base en un conjunto de limitadas observaciones geolocalizadas de presencias de la especie en cuestión (Botella et al., 2018; Beery et al., 2021). De forma tal, los mayores retos que afrontan los SDMs son el limitado número de observaciones disponibles de la especie (que suele ser muy difícil de aumentar), la naturaleza de los datos respecto a la presencia pero no a la ausencia de la especie, y finalmente, el sesgo en el esfuerzo de muestreo de la especie en comparación con la verdadera distribución subyacente, es decir, que las bases de datos de observaciones presentan correlación con los hábitos y preferencias de los observadores, lo cual, en principio, no guarda ninguna relación con los hábitos y preferencias de la especie (Botella et al., 2018).

De forma tal, el SDM para cierta especie suele obtenerse a través de técnicas de modelado de nicho ambiental (Beery et al., 2021). Un nicho ambiental para una especie, en el sentido de Hutchinson (1957), es equivalente a la hipótesis de que dicha especie tiende a habitar un lugar particular en el espacio de variables ambientales (nicho fundamental) que puede ser representado mediante una función de densidad unimodal (Botella et al., 2018). Sin embargo, en realidad, las preferencias de la especie no solo dependen de características abióticas del medio (como es el caso de las variables ambientales), sino también de la dinámica interna de la especie, las interacciones con otras especies y los cambios que sufre el entorno a través del tiempo (Seo et al., 2021; Estopinan et al., 2022; Lee et al., 2022). Dicha configuración dinámica de todas las variables que afectan la presencia de la especie terminan determinando el nicho real o *realizado* de la especie (Botella et al., 2018; Seo et al., 2021).

Debido a que en la práctica se hace casi imposible adquirir todas las variables que determinan el nicho realizado de la especie, las técnicas de modelado de distribución de especies basadas en nichos tienen como objetivo determinar  $P(Y = 1|Z)$ , siendo  $Y = 1$  la presencia de la especie y  $Z$  un vector aleatorio que representa a las variables ambientales (Merow y Silander, 2014). Resolver este problema, por el teorema de Bayes, es equivalente a resolver

$$\frac{f(Z|Y = 1)P(Y = 1)}{f(Z)}. \quad (1)$$

Sin embargo, suele ser imposible determinar  $P(Y = 1)$  y por tanto el problema se centra en determinar  $\frac{f_1(Z)}{f(Z)}$ , siendo  $f_1(Z) = f(Z|Y = 1)$  (Elith et al., 2010). De tal forma, estrictamente hablando, los SDMs no estiman como tal la probabilidad de presencia de la especie sino que determinan un índice relacionado con la intensidad de probabilidad de presencia de la especie que es conocido como la *idoneidad ambiental* (Beery et al., 2021). Los algoritmos modernos más comúnmente utilizados en la literatura para construir los SDMs son básicamente aproximaciones estadísticas que abordan la estimación de  $\frac{f_1(Z)}{f(Z)}$  mediante un enfoque frecuentista y paramétrico, empleando además técnicas de optimización, como es el caso de Maxent (Phillips, Anderson y Schapire, 2006) y Maxlike (Merow y Silander, 2014). Otros enfoques algorítmicos utilizados en el campo incluyen modelos lineales generalizados (Beery et al., 2021) y algunas técnicas de aprendizaje automático como el random forest (Lee et al., 2022).

Si bien podría pensarse que los SDMs pueden tratarse fácilmente como problemas clásicos en el aprendizaje automático, la realidad es bastante distinta debido a la naturaleza de los datos, pocas observaciones fácilmente sesgadas, que reflejan solamente la presencia de la especie pero no su ausencia (Rew, Cho y Hwang, 2021). Adicionalmente, aunque la meta de los algoritmos de predicción de nicho es inferir el nicho a partir de los datos de presencia de la

especie, es decir, caracterizar cierto mapa de acuerdo a la idoneidad ambiental para la proliferación de la especie con base en las condiciones ambientales donde la especie ha sido observada, el nicho como tal permanece desconocido para los investigadores (Merow y Silander, 2014). Lo anterior sugiere que el desempeño de un modelo de SDM no puede ser medido directamente por su habilidad para predecir el nicho como tal puesto que el nicho es desconocido, lo cual obliga a los investigadores a usar métricas que privilegian la asignación de idoneidad ambiental elevada a aquellas locaciones donde la especie ha sido observada (Elith et al., 2010; Mouton, Baets y Goethals, 2010).

Aún así, en años recientes, ha venido aumentando el número de publicaciones donde se emplean técnicas de aprendizaje profundo para construir SMDs a partir del trabajo seminal de Deneu, Servajean et al. (2021) donde se pone en evidencia el potencial de las redes neuronales convolucionales (CNNs) para incluir información espacial en la estructura del nicho y no depender únicamente de las variables ambientales. En este tipo de trabajos se suele modificar ligeramente el enfoque del modelamiento para llevar el problema a un terreno conocido por las técnicas de aprendizaje automático, por ejemplo Deneu, Joly et al. (2022) opta por trabajar con un dataset que tiene registros de miles de especies al mismo tiempo y por ende puede trabajar el problema en términos de clasificación dado un conjunto de variables ambientales. Otro ejemplo sería el de Rew, Cho y Hwang (2021), donde se propone una metodología para generar, de manera artificial, información sobre la ausencia de la especie, llevando nuevamente el problema hacia el terreno de la clasificación; o el trabajo de Bourhis et al. (2023), donde se lleva aún más lejos la exploración del modelamiento conjunto de especies para extraer características de cómo se relaciona cada especie con el ambiente en una forma distinta a la otra especie. Entre los algoritmos que se utilizan para comparar el rendimiento de dichas CNNs encontramos usualmente a los árboles de decisión, que facilitan la interpretabilidad de los resultados y Maxent, por su amplio uso dentro del campo, aunque los resultados de ambas técnicas no sean exactamente comparables dadas las variaciones en el enfoque.

Las dificultades para acceder a datos sobre especies de manera consistente también se ha traducido en dificultades a la hora de validar el desempeño de los algoritmos de SDMs, especialmente aquellos cuyo su enfoque consiste en determinar el nicho potencial de la especie (Grimmett, Whitsed y Horta, 2021). Como respuesta a dicha situación se han desarrollado recientemente algunos métodos de generación de especies virtuales que permiten modelar la forma en la que la especie responde a las variables ambientales, es decir, son técnicas que permiten diseñar funciones de idoneidad ambiental en el espacio de las variables ambientales (Leroy et al., 2016). Posteriormente, tras emplear algún método, bien sea dinámico, estocástico o determinista, para muestrear el nicho, se procede a aplicar un SDM con base en los muestreos obtenidos (Garzon-Lopez et al., 2016). Las especies virtuales constituyen entonces herramientas poderosas que permiten generar conjuntos de datos para poner a prueba a los algoritmos de SDMs al mismo tiempo que permiten tener acceso a información del nicho como tal (Leroy et al., 2016). Por tal razón, pueden emplearse para evaluar el potencial de los distintos algoritmos para reconstruir el nicho potencial de la especie a partir de los datos de solo presencia.

## 4. Metodología de investigación

El presente trabajo se centra en estimar  $f_1(Z)$  en primer lugar y, posteriormente, estimar  $\frac{f_1(Z)}{f(Z)}$ , en ambos casos mediante datos generados a partir de un nicho artificial. Los nichos virtuales consisten en una función  $\Omega(Z_i)$  que asigna a cada pixel de cierto mapa una idoneidad ambiental  $f_1(Z)$  de acuerdo a la combinación de variables ambientales para dicho pixel ( $Z_i$ ). Una vez la función asigna la idoneidad ambiental al mapa, este se muestrea considerando que la probabilidad de muestrear al pixel  $i$ -ésimo es  $P(i = 1) = \frac{\Omega(Z_i)}{\sum_{i=1}^m \Omega(Z_i)}$ . Para este proyecto, se cuenta con un mapa de Sudamérica de  $1637 \times 1117$  pixeles en 19 variables ambientales y se implementa un algoritmo de generación de nichos virtuales de alta variabilidad con el cual se construye un dataset para una especie virtual con aproximadamente 5000 registros.

Dado que es posible acceder a la información del nicho de la especie como tal al emplear paquetes de generación de especies virtuales, como es el caso de *Virtual species* propuesto por Leroy et al. (2015), se propone abordar la estimación de  $f_1(Z)$  como un problema de aprendizaje supervisado. Aún así, los datos reales en esta área de trabajo no traen dicha información. Por tanto, se plantea continuar el trabajo buscando la estimación de  $f_1(Z) = f(Z|Y = 1)$  o la optimización de alguna métrica de rendimiento clásica como las reportadas por Mouton, Baets y Goethals (2010) sin hacer uso directo de la información del nicho de la especie. Por tal razón, durante las distintas fases del proyecto se hará uso de diferentes algoritmos de aprendizaje automático e inteligencia artificial como, por ejemplo, la regresión lineal, los árboles de decisión, el Random Forest, y la Red neuronal y se evaluarán sus desempeños mediante la estimación del error de generalización con el fin de identificar cual de estos algoritmos es el más idóneo para nuestro problema particular en su versión más general. En cuanto a las métricas, utilizaremos, en primera instancia, algunas comúnmente utilizadas para problemas de regresión, como es el caso del *mean absolute error* (MAE), el *mean squared error* (MSE), y el score  $R^2$ . También deben tenerse en cuenta alternativas previamente propuestas para este tipo de problemas como las mencionadas por Mouton, Baets y Goethals (ibíd.).

El método del estado del arte BNM es un toolbox de modelado de nicho y generación de especies virtuales desarrollado por el grupo de modelado matemático de la Universidad EAFIT e implementado en el ambiente de MatLab que puede encontrarse en el siguiente repositorio de GitHub [\[4\]](#). El algoritmo BNM explota el concepto de independencia entre las variables explicativas para determinar las zonas con mayor densidad de probabilidad mediante un análisis marginal ponderado con una media geométrica. De forma tal, en el presente trabajo utilizaremos MatLab para la generación del nicho y el muestreo de la especie virtual que estudiaremos como también para la aplicación de la técnica BNM de predicción de nicho y la comparación final con los nichos inducidos por las técnicas de aprendizaje automático que implementaremos. Es necesario mencionar que ningún algoritmo del estado del arte, según podemos concluir a partir de la revisión de literatura, utiliza actualmente las etiquetas del nicho durante el proceso de predicción del nicho. Lo anterior también aplica para el método BNM y por tanto lo pone en una clara desventaja con relación a los métodos que implementaremos. Como parte de nuestro trabajo futuro planteamos proponer un nuevo marco de trabajo que permita incluir técnicas de aprendizaje automático sin recurrir a las etiquetas del nicho en cuestión durante el proceso de entrenamiento.

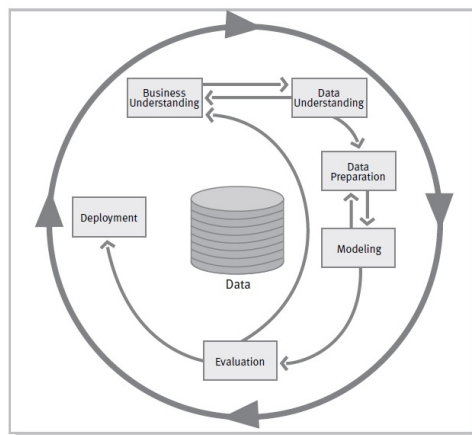


Figura 1: Representación de las fases de la metodología Cross Industry Standard Process for data Mining (CRISP-DM), tomada del artículo titulado *Metodología para Machine Learning (I)* de Gavilán (2021).

La implementación de las técnicas de aprendizaje automático será llevada a cabo en el lenguaje python, incluyendo los datos generados por el toolbox BNM en el formato especificado en la siguiente sección y siguiendo los pasos

propuestos por la metodología CRISP-DM para ciencia de datos y aprendizaje automático con excepción de la fase de despliegue 1. En el presente trabajo nos proponemos evaluar el rendimiento de 9 algoritmos de aprendizaje automático apropiados para problemas de regresión, empezando desde los más tradicionales de la literatura, como es el caso de la regresión lineal y la regresión lineal con regularización, y aumentando la complejidad de la técnica hasta llegar al bagging y boosting de árboles de decisión y a las redes neuronales. Debido a que no incluiremos información espacial en el presente estudio, como se discute en la siguiente sección, emplearemos la red neuronal como un regresor sin necesidad de implementar capas convolucionales. Una vez se corrieron los distintos algoritmos con una configuración estándar, se procedió a seleccionar aquellos con mejor desempeño para realizar un proceso de identificación de hiperparámetros mediante validación cruzada con k-folds.

## 5. Análisis de los datos

La fuente que se utilizó para extraer la información de las variables ambientales fue WorldClim (Fick e Hijmans, 2017). WorldClim es una base de datos de variables ambientales a nivel global que proporciona información sobre todos los lugares de la Tierra en alta resolución espacial. Dicha información se deriva de observaciones climáticas a largo plazo de estaciones meteorológicas terrestres y de satélite que se interpolan para proporcionar información completa sobre todo el mundo. Además, estos datos suelen utilizarse en la modelización de la distribución de especies y en técnicas de modelización ecológica afines, que es precisamente nuestro problema de interés.

La información ambiental del mapa de Sudamérica se le proporciona como entrada al algoritmo de generación de especies virtuales, el cual procede a generar el nicho y a muestrearlo con aproximadamente 5 mil observaciones. El formato de estas observaciones generadas con la máquina virtual puede ser como se desee, sin embargo en este caso por facilidad estarán almacenados en formato *.csv*. El proceso de generación de la especie virtual será desconocido para los algoritmos de inteligencia artificial, por tal razón, se realizará un análisis exploratorio para cada conjunto de datos empezando por un análisis de correlación y de detección de datos atípicos. La naturaleza de los datos para cada especie virtual se ilustra en la Figura 2.

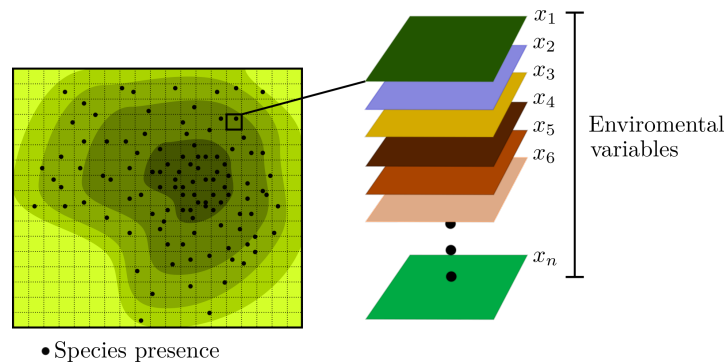


Figura 2: Representación de los datos generados a partir de un muestreo de una especie virtual. En el mapa de la izquierda tenemos, en un gradiente de color, la idoneidad ambiental dada por la función  $\Omega(Z_i)$ . Dicha idoneidad se utiliza para producir un muestreo de la especie (puntos negros) que tienen una representación en 19 variables ambientales ( $n = 19$ ).

De este modo, el conjunto de datos consta de 23 variables, 19 de ellas son variables bioclimáticas que representan tendencias anuales, estacionalidad y factores ambientales extremos o limitantes, 3 de ellas indican el nombre de la especie y la ubicación en donde se ha visto en términos de latitud y longitud mientras que la restante corresponde a la idoneidad ambiental, es decir que tan apto o no es el lugar para que la especie prospere allí. Así, del proceso de

generación de muestreos a partir de especies virtuales se obtiene un conjunto de datos para una especie que consistirá en un arreglo cuyo número de filas dependerá del tamaño del muestreo, en este caso 4982 y con 23 columnas. Es decir, nuestros datos serán la representación en el espacio de varias variables de las observaciones de la especie virtual junto con su ubicación espacial. En la siguiente tabla se resume la información acerca del conjunto de datos.

Variable	Descripción	Tipo de variable
Name	Nombre de la especie	Variable tipo palabra
LONG	Longitud en donde se ha visto la especie	Variable continua
LAT	Latitud en donde se ha visto la especie	Variable continua
BIO1	Temperatura media anual	Variable continua
BIO2	Rango diurno medio (media mensual (temperatura máxima-temperatura mínima))	Variable continua
BIO3	Isotermia (Esta se calcula como $\frac{BIO2}{BIO7} \times 100$ )	Variable continua
BIO4	Estacionalidad de la temperatura (Desviación estandar $\times 100$ )	Variable continua
BIO5	Temperatura máxima del mes más cálido	Variable continua
BIO6	Temperatura mínima del mes más frío	Variable continua
BIO7	Rango anual de temperatura (BIO5-BIO6)	Variable continua
BIO8	Temperatura media del trimestre más húmedo	Variable continua
BIO9	Temperatura media del trimestre más seco	Variable continua
BIO10	Temperatura media del trimestre más calido	Variable continua
BIO11	Temperatura media del trimestre más frío	Variable continua
BIO12	Precipitación anual	Variable continua
BIO13	Precipitaciones del mes más húmedo	Variable continua
BIO14	Precipitaciones del mes más seco	Variable continua
BIO15	Estacionalidad de las precipitaciones (Coeficiente de variación)	Variable continua
BIO16	Precipitaciones del trimestre más húmedo	Variable continua
BIO17	Precipitaciones del trimestre más seco	Variable continua
BIO18	Precipitaciones del trimestre más cálido	Variable continua
BIO19	Precipitaciones del trimestre más frío	Variable continua
BIO20	Idoneidad ambiental del lugar para la especie	Variable continua

Tabla 1: Información resumida de las variables presentes en el conjunto de datos.

Para analizar la naturaleza de las variables ambientales y su relación entre ellas computamos, en primer lugar, la matriz de correlaciones de pearson para todas las variables incluyendo la variable objetivo tal y como puede verse en la Figura 3. A partir de este análisis es posible concluir que existen algunos clústeres de variables altamente correlacionadas entre sí, como es el caso de las variables *BIO* con índices entre 1 y 3 y entre 15 y 19. Algunas



de dichas correlaciones son perfectamente comprensibles debido a la naturaleza de las variables: básicamente se dividen en dos categorías, la primera relacionada con la temperatura y la segunda relacionada con la precipitación. Sin embargo, algunas otras correlaciones son bastante interesantes porque revelan interacciones entre variables de distinta categoría. Vale la pena notar que nuestra variable objetivo (Prob) no presenta correlaciones fuertes con ninguna de las variables explicativas, lo cual parece indicar que el nicho está determinado por interacciones no lineales entre las variables ambientales. Se aprecia, sin embargo, una correlación que quizá deba tenerse en cuenta entre la variable objetivo y la variable de temperatura media del trimestre más cálido (*BIO10*). Dicha correlación nos lleva a hipotetizar que *BIO10* puede desempeñar un papel relevante en el proceso de predicción del nicho. Aún así, se debe tener en cuenta que una correlación de 1 no implica causalidad, es decir, esto no significa que una de las variables cause la otra.

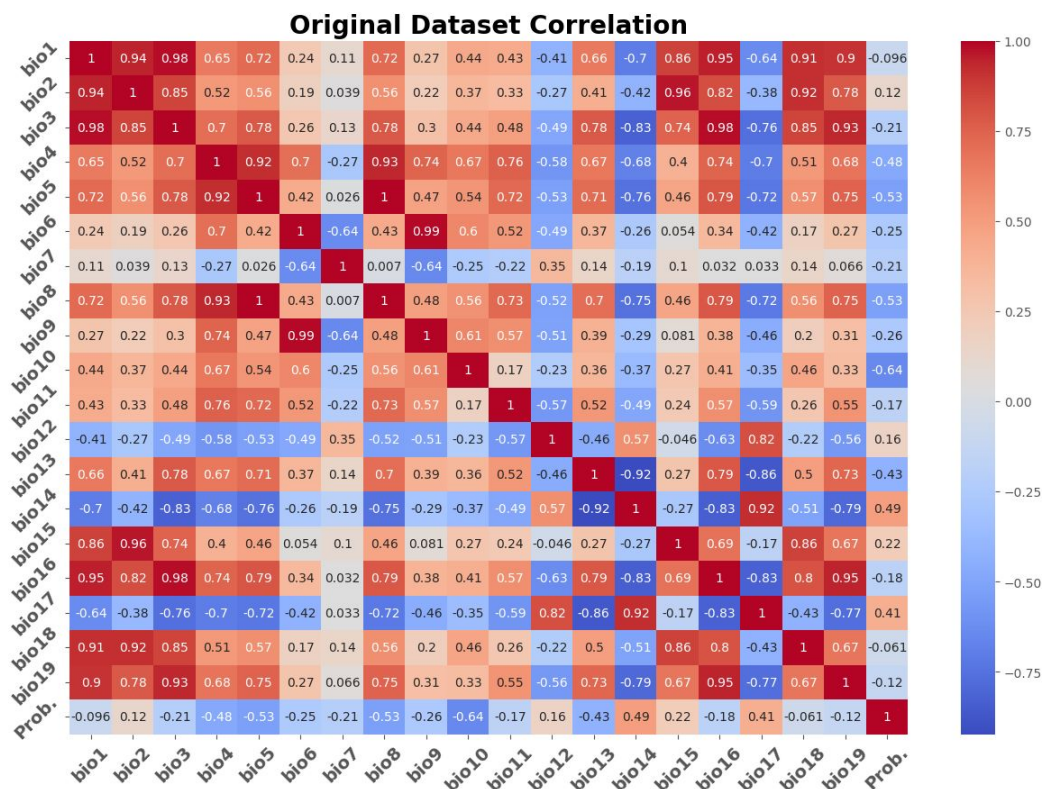


Figura 3: Matriz de correlación de las variables que representan a los datos bio1-bio19, incluyendo a la variable objetivo de idoneidad ambiental (prob).

Otra técnica que utilizamos de manera preliminar para explorar los datos fue la normalización y posterior reducción de dimensión por proyecciones de los datos sobre direcciones ortogonales de máxima varianza, también conocido como análisis de componentes principales (PCA), tal y como el lector puede verlo en la Figura 4. Para graficar los resultados de este proceso elegimos las dos componentes ortogonales con máxima varianza. También procedimos a graficar sobre dicho plano la contribución de cada variable a cada componente principal mediante vectores (en azul). La Figura 4 nos permite visualizar las fuertes correlaciones presentes entre algunas de las variables cuyos vectores de contribución se superponen y también nos permite identificar la presencia de variables que posiblemente no tendrán gran influencia sobre el comportamiento del modelo: aquellas cuyos vectores tienen poca

contribución al primer componente (eje x) de la PCA. Otra ventaja de esta técnica es que nos permite visualizar datos que se encuentran embebidos en espacios de mayor dimensión para buscar posibles patrones. De forma tal, es posible visualizar que los datos se concentran notoriamente en ciertas regiones del plano de PCA, lo que sugiere que efectivamente hay un proceso generador de los datos con una distribución que podría ser aprendida por los algoritmos a implementar.

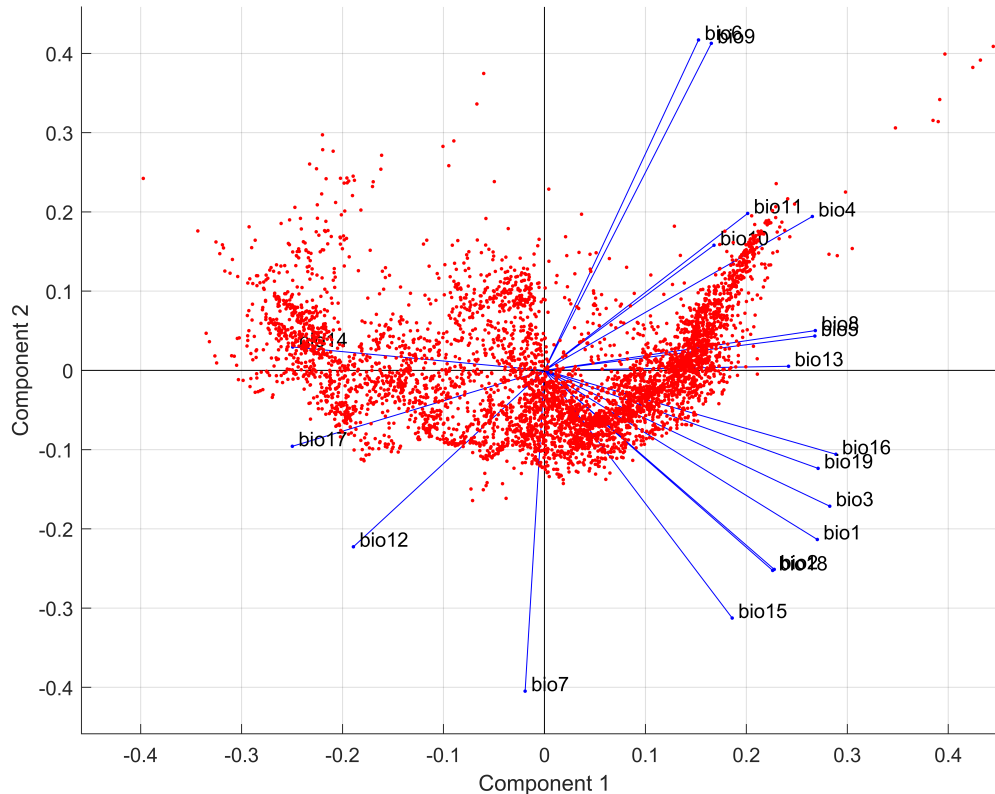


Figura 4: PCA 2 dimensional de los datos (en rojo). También se grafica la contribución de cada variable a los respectivos componentes principales (en azul).

Finalmente decidimos explorar las variables ambientales de forma marginal a fin de obtener información relacionada con su naturaleza. Para tal fin computamos algunos indicadores estadísticos para cada una de las variables predictoras tales como la media, desviación estándar, cuartiles, asimetría y curtosis. En la Tabla 2 se muestra el resultado para algunas de las variables de forma que se intuyan las notables diferencias entre ellas. Por ejemplo, a partir de los datos de la tabla es posible evidenciar que la variable *BIO2* presenta una alta curtosis y asimetría negativa, lo que indica que hay una mayor concentración alrededor de la media y los datos atípicos son poco frecuentes, sin embargo los valores más bajos están más dispersos y hay una mayor concentración en los valores más altos.

La Tabla 2 también nos permite evidenciar diferencias notables en términos órdenes de magnitud para las variables ambientales. Lo anterior sugiere que los datos deben ser normalizados como parte de la rutina de preprocesamiento antes de proceder a aplicar los distintos algoritmos de aprendizaje automático. Vale la pena aclarar que la curtosis



en la Tabla 2 está calculada alrededor de 0, por lo que un valor por debajo de 0 indica que la distribución tiene colas más ligeras y presenta menor apuntamiento alrededor de la moda en comparación con la distribución normal, por otro lado una curtosis positiva indica una presencia de colas más pesadas y mayor apuntamiento alrededor de la moda en comparación con la distribución normal.

Variable	Media	Desviación	Mínimo	25 %	50 %	75 %	Máximo	Asimetría	Curtosis
<b>LONG</b>	-60.53	10.02	-81.1	-68.43	-62.06	-53.72	-35.02	0.44	-0.57
<b>LAT</b>	-15.86	15.25	-55.39	-27.02	-13.16	-4.10	11.93	-0.42	-0.61
<b>BIO1</b>	21.33	6.10	-7.87	18.06	24.02	25.94	29.35	-1.23	0.64
<b>BIO2</b>	23.83	4.81	-3.35	22.78	25.70	26.81	30.69	-1.81	2.91
<b>BIO3</b>	18.60	7.90	-12.53	13.19	21.74	25.23	28.16	-0.90	-0.41
<b>BIO4</b>	1426.14	854.58	0	755	1397	2046	6592	0.49	0.67
<b>BIO18</b>	22.09	6.59	-11.38	21.69	24.92	26.02	29.27	-1.82	2.32
<b>BIO19</b>	20.09	6.84	-3.35	15.42	22.49	25.85	29.26	-0.86	-0.29
<b>Prob</b>	0.65	0.12	0.04	0.59	0.67	0.73	0.98	-0.86	1.71

Tabla 2: Estadística descriptiva para algunas de las variables del conjunto de datos.

## 6. Algoritmos

### 6.1. Preprocesamiento de los datos

Para el preprocesamiento de los datos, se procede a eliminar del conjunto de datos a las variables *Name*, *LONG* y *LAT* ya que en primer lugar la variable *Name* al trabajar con una única especie no es relevante para los algoritmos y por otro lado, las variables *LONG* y *LAT* introducen un sesgo espacial que no se puede tener en cuenta dentro del marco de trabajo propuesto. Las variables predictoras restantes corresponden a *BIO1* hasta *BIO19*, que representan variables ambientales, mientras que la variable objetivo es *Prob* que representa la idoneidad ambiental del lugar para la especie. Por otro lado, notamos que el proceso de generación de las observaciones de la especie virtual produce registros en los que no se cuenta con información para algunas variables. Para lidiar con esta situación decidimos eliminar directamente toda la observación de la especie, ya que la técnica de imputación de valores perdidos con la media no es adecuada debido a la naturaleza de los datos y al contexto multivariado del problema. Finalmente, se escalan los datos a través de la normalización de las variables, ya que de esta manera todas las variables estarán en una escala común y esto evita que una variable con valores más grandes o pequeños domine el modelo incorrectamente.

### 6.2. Entrenamiento de modelos

Teniendo en cuenta que la variable objetivo (*Prob*) es de tipo continuo, se entrenaron diferentes modelos empleando múltiples algoritmos útiles para problemas de regresión. En primer lugar implementamos el algoritmo más básico, que es la regresión lineal, que sirve además como punto de partida para comparar el rendimiento de otros algoritmos. Luego procedimos con la implementación de la regresión lineal tipo Lasso, Ridge y elastic net, que aplican técnicas de regularización para penalizar los coeficientes del modelo y limitar su magnitud, lo cual ayuda a reducir la complejidad

del modelo y evitar el sobreajuste. Luego, pasamos a implementar algoritmos más complejos y efectivos como la regresión por máquina de soporte vectorial, decision trees, XGBoost, random forest y multilayer perceptron. Se reservó el 20 % del conjunto de datos para la validación de los modelos resultantes, de forma que nunca se entrena ni se compara con ellos durante el proceso de entrenamiento de los modelos. El dataset de validación nos ayuda a evaluar la capacidad de generalización de los modelos y da una idea de cómo se comportará el modelo cuando se utilice en situaciones reales con nuevos datos. El 80 % restante de los datos se dividen en 70 % para entrenamiento y el 30 % para testeo. En la siguiente tabla se muestra diferentes métricas y medidas de interés obtenidas para cada uno de los modelos desarrollados con ayuda de la librería *scikit learn*.

Algoritmo	MAE	MSE	R2	Variables principales	Media	Var.	Asimetría	Curtosis
<b>XGBoost</b>	0.01122	0.0004	0.97527	BIO10, BIO14, BIO15	0.0001	0.0004	-2.3972	21.291
<b>Random Forest</b>	0.01199	0.00054	0.96718	BIO10, BIO14, BIO5	-0.0015	0.0005	-3.4512	24.0623
<b>MLP (NN)</b>	0.01836	0.00066	0.95948	BIO11, BIO5, BIO12	0.0003	0.0007	-0.0753	4.681
<b>Decision Tree</b>	0.01973	0.00125	0.92382	BIO10, BIO15, BIO5	-0.0023	0.0012	-0.8696	10.6396
<b>Regresión Ridge</b>	0.02096	0.00091	0.94461	BIO5, BIO10, BIO8	0.0004	0.0009	-0.8512	6.9324
<b>Regresión lineal</b>	0.02097	0.00091	0.94467	BIO16, BIO17, BIO15	0.0004	0.0009	-0.8353	6.8769
<b>ElasticNet</b>	0.02112	0.00093	0.94336	BIO5, BIO10, BIO2	0.0006	0.0009	-0.9743	6.8689
<b>SVM</b>	0.03503	0.00189	0.88434	BIO10, BIO6, BIO9	0.0105	0.0018	-0.1316	0.3856
<b>Regresión Lasso</b>	0.09456	0.01641	-0.00305	None	-0.0071	0.0164	-0.8723	1.8581
<b>BNM</b>	0.1542	0.0028	-1.402	NA	-0.1241	0.0226	-0.1964	-0.0196

Tabla 3: Diferentes métricas e indicadores para evaluar el desempeño de cada uno de los modelos en los datos de testeo. Los algoritmos se implementaron con la configuración dada por defecto.

En la tabla 3, las columnas etiquetadas con Media, Var, Asimetría y Curtosis se computan estos indicadores estadísticos para los residuales de cada uno de los modelos a fin de obtener un proxy que evidencie ciertos atributos de algunas de las técnicas, como es el caso de reducción de varianza del Random Forest o la reducción de Bias del boosting. Revisando con detenimiento los residuales obtenidos de los diferentes modelos aplicados, encontramos que la asimetría en todos los casos es negativa (hay mas datos atípicos a la izquierda de la media muestral), osea que los atípicos por lo general fueron casos donde la predicción de la variable objetivo tomo un valor menor comparado con su medición original.

En el caso de la curtosis vemos que cuando la métrica MAE se reduce la curtosis aumenta en la mayoría de casos. Los valores positivos  $> 3$  de la curtosis, indican que todos los residuales están cobijados por la misma distribución (diferente para cada modelo), pero las colas de dicha distribución están abultadas de datos atípicos. Estos resultados sugieren que para mejorar la predicción hay que analizar esos datos que crecen al mejorar la predicción. Sin embargo, hay que rescatar que en el caso de XGBoost porque redujo el valor de la curtosis y a su vez disminuyo el valor de MAE. Esto puede ser porque XGboost en comparacion con modelos como Random forest tiene principalmente (i) mejor rendimiento (construye un modelo de regresión aditiva a partir de árboles de decisión débiles, ajustando secuencialmente los árboles para corregir los errores cometidos por modelos anteriores), y (ii) regulariza con múltiples hiperparámetros para controlar el sobreajuste (como la tasa de aprendizaje, la profundidad máxima del árbol, la regularización L1 y L2, entre otros).

Retomando la tabla 3, la métrica que seleccionamos como aquella de mayor interés, fue *Mean Absolute Error* debido a que la variable objetivo se encuentra entre 0 y 1 y el *Mean Squared Error* no se desempeña tan bien para este tipo

de datos. Debido a los rendimientos preliminares de los algoritmos que obtuvimos en la Tabla 3 decidimos centrar nuestra atención en analizar más a fondo el los algoritmos de regresión tipo Ridge, Random Forest y XGBoost. Elegimos la regresión Ridge debido a que sigue teniendo una estructura clara de regresión lineal y tuvo el mejor desempeño de este tipo de algoritmos. Por otro lado, escogimos el Random Forest y el XGBoost porque son técnicas con mucho potencial para este tipo de problemas de regresión que cuentan con múltiples hiperparámetros y, si bien parten desde la misma estructura base del árbol de decisión, utilizan estrategias diferentes entre sí para lograr un mayor desempeño.

### 6.2.1. Regresión lineal tipo Ridge

En la regresión lineal básica, se minimiza la suma de los errores cuadrados para encontrar los coeficientes óptimos. Sin embargo, en presencia de multicolinealidad, los coeficientes pueden tener una alta varianza y volverse inestables (Hastie, Tibshirani y Friedman, 2009). La regresión lineal tipo Ridge aborda este problema introduciendo una penalización  $L^2$  en la función de costo (ibíd.). Esta penalización reduce la magnitud de los coeficientes de regresión al agregar un término proporcional a la suma de los cuadrados de los coeficientes (ibíd.). Como resultado, los coeficientes se reducen, pero no se eliminan por completo. De esta manera, la regularización de Ridge ayuda a controlar la multicolinealidad al limitar la magnitud de los coeficientes, lo que puede mejorar la estabilidad y la precisión de las estimaciones (ibíd.). Además, este tipo de regularización contribuye a que todas las variables predictoras contribuyen al modelo, aunque algunas pueden tener un efecto más pequeño. A continuación se muestran algunos resultados importantes al realizar la regresión lineal tipo Ridge con ayuda de *scikit learn* y un factor de regularización igual a 0.5.

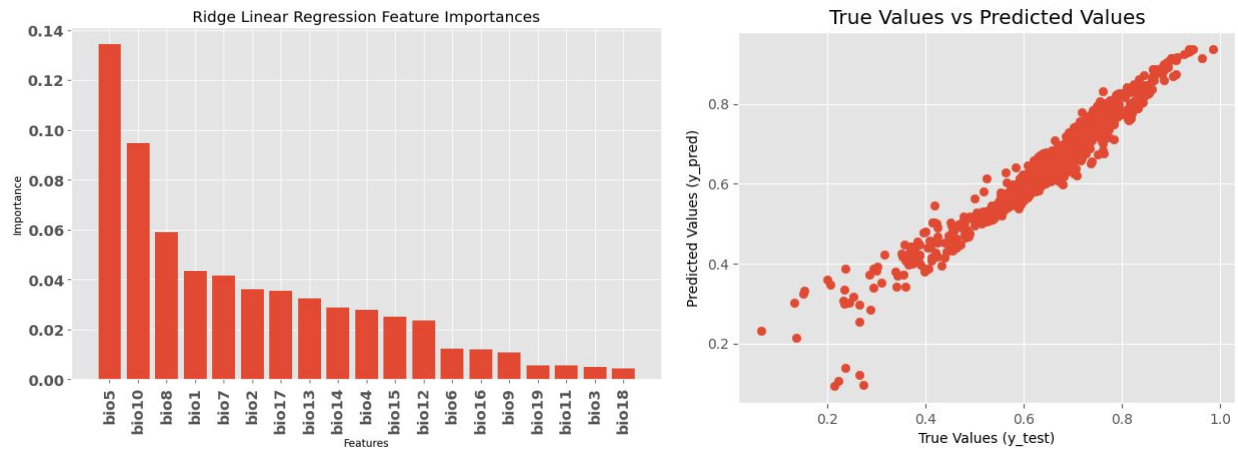


Figura 5: Resultados importantes del modelo de regresión lineal tipo Ridge

Así, según las gráficas anteriores, en la gráfica de la izquierda se puede ver lo dicho anteriormente, es decir que al hacer los coeficientes más pequeños por medio de la regularización de Ridge todas las variables predictoras contribuyen al modelo, sin embargo, en este caso las variables que más contribuyen son BIO5: Temperatura máxima del mes más cálido, BIO10: Temperatura media del trimestre más cálido y BIO8: Temperatura media del trimestre más húmedo. Por otro lado, en la gráfica de la derecha correspondiente al diagrama de dispersión de valores reales contra valores predichos por el modelo tiene una forma bastante lineal, lo que indica que el modelo está capturando correctamente la relación entre las características de entrada y la variable objetivo y está proporcionando estimaciones precisas.

### Cross validation y selección de hiperparámetros para la regresión lineal tipo Ridge

Con el objetivo de encontrar los hiperparámetros del algoritmo que producen los mejores resultados, se hace uso de

la función *RandomizedSearchCV* de la biblioteca de *Scikit-learn*, ya que ésta proporciona una forma eficiente de buscar una combinación óptima de hiperparámetros al muestrear aleatoriamente diferentes combinaciones de valores de hiperparámetros de un espacio de búsqueda definido y comprueba su idoneidad mediante validación cruzada con 5 folds. Esta función evalúa el rendimiento del modelo con algunas de las posibles combinaciones de hiperparámetros en lugar de probarlas todas, es decir, realiza una búsqueda aleatoria seleccionando una muestra específica de combinaciones de hiperparámetros del espacio de búsqueda e infiere el rendimiento de las otras combinaciones con base en el muestreo.

Para este caso, definimos el espacio de búsqueda de hiperparámetros sobre el factor de regularización ( $\alpha$ ) y el solver, encontrando que los hiperparámetros que producen los mejores resultados son  $\alpha = 0,92725$  y *solver*= *sparse\_cg*. Los otros hiperparámetros quedan establecidos por defecto. El *Mean Absolute Error* de este modelo con hiperparámetros optimizados es 0.020964, por lo tanto no se tiene una mejoría significativa respecto al modelo original. La Figura 6 muestra cómo el factor de regularización afecta las diferentes métricas de desempeño del modelo en entrenamiento y testeo. Estos resultados evidencian que la diferencia entre las métricas de train y test no suele ser muy significativas, todas las diferencias rondan alrededor de 0.008. Adicionalmente, según el MAE, que es nuestra métrica de mayor interés, se puede apreciar que el mejor valor para el factor de regularización que corresponde a la menor diferencia entre la métrica en train y test, para evitar el sobreajuste esta muy cercana a uno, lo que coincide con el resultado obtenido previamente ( $\alpha = 0,92725$ ).

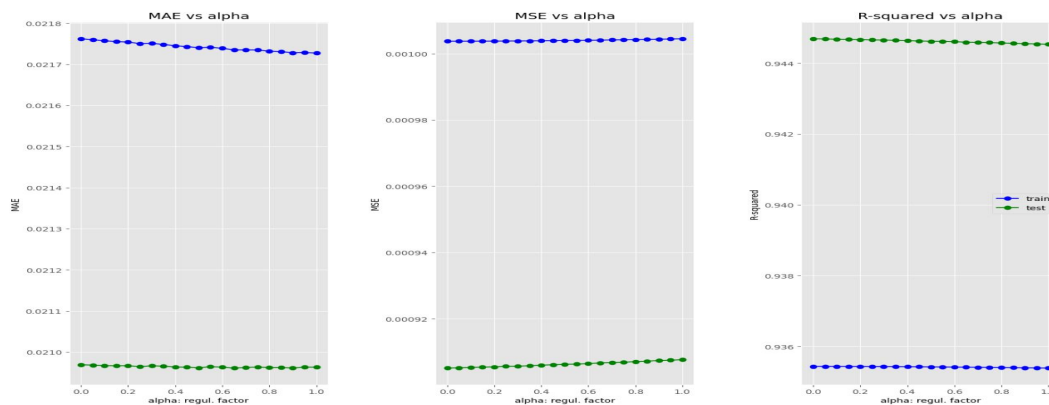


Figura 6: Métricas en train y test con diferentes factores de regularización.

### 6.2.2. Random forest

De manera similar, a continuación se muestran algunos resultados importantes obtenidos al implementar el algoritmo de random forest con ayuda de *Scikit-learn* con hiperparametrización por defecto. En la Figura 7 izquierda se puede ver que las variables que más contribuyen son BIO10: Temperatura media del trimestre más cálido, BIO14: Precipitaciones del mes más seco y BIO5: Temperatura máxima del mes más cálido; los cuales son resultados muy similares a los de la regresión lineal tipo Ridge en donde las variables que más contribuían al modelo son BIO10, BIO5 y BIO8. Por otro lado, en la Figura 7 derecha, correspondiente al diagrama de dispersión de valores reales contra valores predichos por el modelo puede apreciarse una clara tendencia lineal, lo que indica que el modelo está capturando correctamente la relación entre las características de entrada y la variable objetivo y está proporcionando estimaciones satisfactorias.

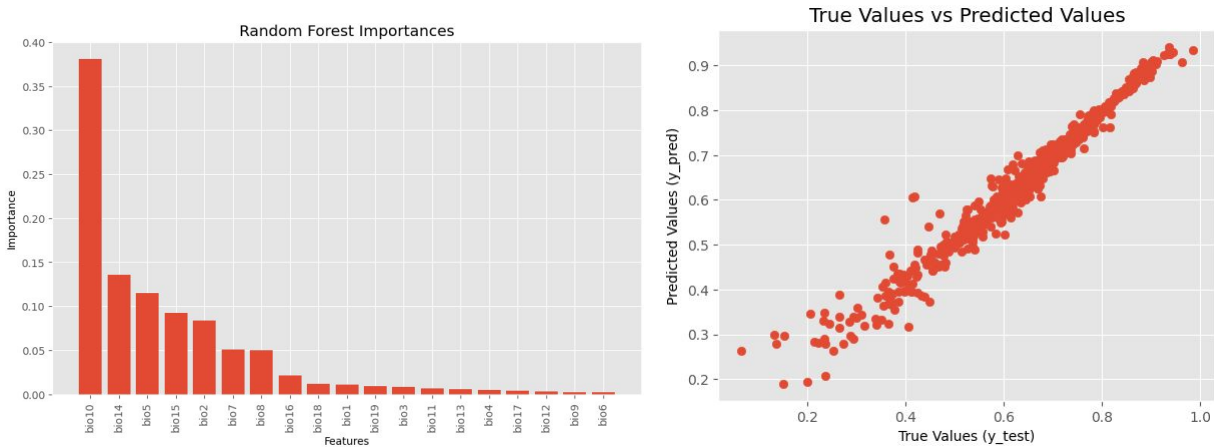


Figura 7: Resultados importantes del modelo de random forest

### Cross validation y selección de hiperparámetros para la random forest

Al igual que en el caso anterior, se utiliza la función *randomizedSearchCV* para buscar la combinación de hiperparámetros del algoritmo asociados al menor error de generalización. El espacio de búsqueda de hiperparámetros es definido sobre el número de árboles en el bosque (*num\_estimators*), el número mínimo de muestras necesarias para dividir un nodo interno (*min\_samples\_split*), el número mínimo de muestras necesarias para estar en un nodo hoja (*min\_samples\_leaf*), la cantidad de características que hay que tener en cuenta a la hora de buscar el mejor split (*max\_features*) y la profundidad máxima del árbol (*max\_depth*). De este modo, encontramos que los hiperparámetros que producen los mejores resultados son *n\_estimators* = 166, *min\_samples\_split* = 2, *min\_samples\_leaf* = 1, *max\_features* = *auto* y *max\_depth* = 50, los demás hiperparámetros quedan establecidos por defecto. De forma tal, el *Mean Absolute Error* de este modelo mejorado es 0.01179, que comparado con el desempeño del modelo original presenta una mejora de aproximadamente en un 0.05 % en cuanto a su exactitud, lo cual, al igual que para el caso anterior, no representa una mejora significativa a primera vista. La Figura 8 muestra como la cantidad de árboles en el bosque afecta las diferentes métricas de desempeño del modelo en entrenamiento y testeo. Es posible evidenciar que la diferencia entre las métricas de train y test no suelen ser muy significativas, la mayoría de las diferencias rondan alrededor de 0.0075. Sin embargo, cuando la cantidad de árboles es muy baja se puede apreciar que el desempeño del algoritmo en train y test disminuye considerablemente, por lo tanto, para superar el bias, se debe tener un número considerable de árboles.

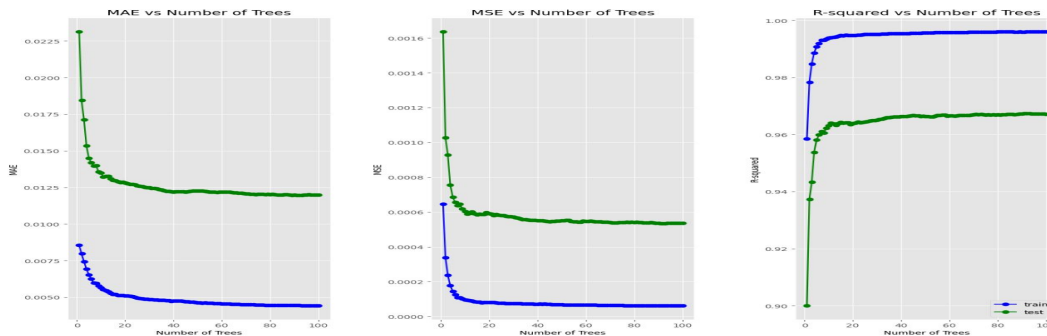


Figura 8: Métricas en train y test con diferente cantidad de árboles en el bosque

### 6.2.3. XGBoost

La Figura 9 resume algunos resultados importantes al realizar el modelo XGBoost con ayuda de la función *xgb.XGBRegressor* y sus valores por defecto. En la Figura 9 izquierda se puede ver que las variables que más contribuyen en el modelo son BIO10: Temperatura media del trimestre más cálido, BIO14: Precipitaciones del mes más seco y BIO15: Estacionalidad de las precipitaciones, los cuales son resultados muy similares a los obtenidos por medio de la regresión lineal tipo Ridge y Random Forest. Por otro lado, en la Figura 9 derecha, correspondiente al diagrama de dispersión de los valores reales contra los valores predichos por el modelo, se puede apreciar que éste tiene una forma bastante lineal, lo que indica que el modelo está capturando correctamente la relación entre las características de entrada y la variable objetivo y está proporcionando estimaciones satisfactorias.

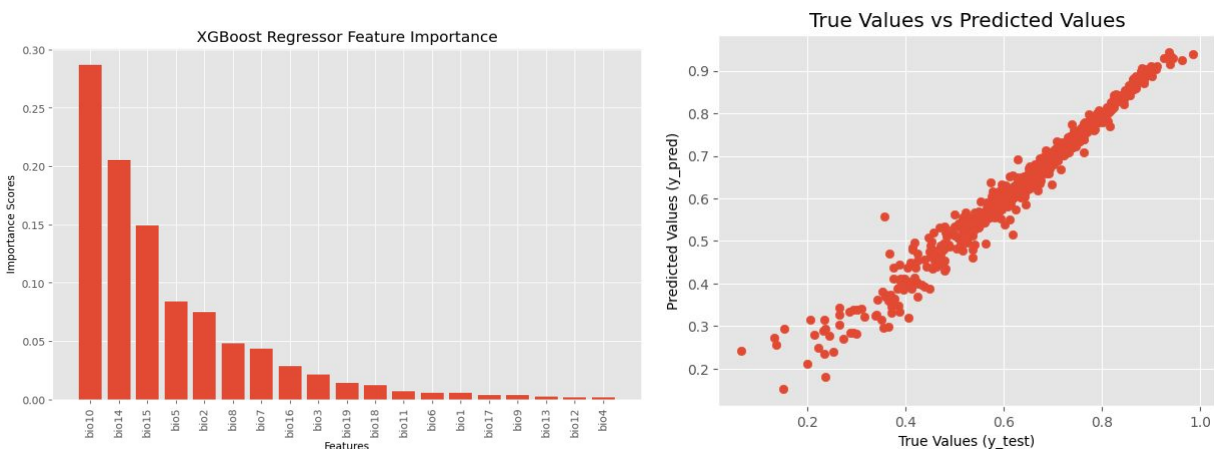


Figura 9: Resultados importantes del modelo XGBoost

### Cross validation y selección de hiperparámetros para XGBoost

Al igual que en los casos anteriores, se utiliza la función *randomizedSearchCV* en busca de la combinación de hiperparámetros que causen la máxima reducción del error de generalización del para el algoritmo. De esta manera, el espacio de búsqueda de hiperparámetros es definido sobre la tasa de aprendizaje, también conocido como el término de regularización  $L^1$  (Lasso) para controlar la complejidad del modelo ( $\alpha$ ), el potenciador (*booster*) y el número de árboles en el modelo (*n\_estimators*). De este modo, los hiperparámetros que producen los mejores resultados son  $\alpha = 0,121$ , *booster* : *dart* y *n\_estimators* : 539, los demás hiperparámetros quedan establecidos por defecto. El *Mean Absolute Error* de este modelo mejorado es 0.010108 que comparado con el modelo original que tenía un MAE de 0.01122, el modelo mejoró aproximadamente en un 0.17% en cuanto a su exactitud, siendo esta la mayor tasa de mejora obtenida hasta el momento. La Figura 10 muestra como la tasa de aprendizaje afecta las diferentes métricas de desempeño del modelo en entrenamiento y testeo. Los resultados en la Figura 10 permiten evidenciar que la diferencia entre las métricas de train y test no suelen ser muy significativas, la mayoría de diferencias están por debajo de 0.025. Sin embargo, teniendo en cuenta que la métrica de mayor interés es el MAE se puede apreciar que el mejor valor para la tasa de aprendizaje que corresponde a la menor diferencia entre la métrica en train y test para evitar el sobreajuste, está entre 0 y 0.2, lo que coincide con el resultado encontrado previamente ( $\alpha = 0,1210$ ).



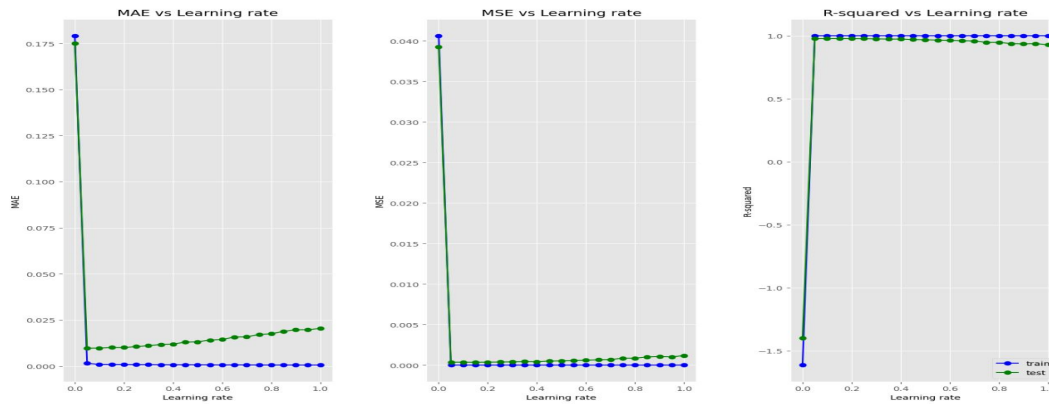


Figura 10: Métricas en train y test con diferentes tasas de aprendizaje para XGBoost

### 6.3. Comparación final entre algoritmos

Para finalizar, en la Tabla 4 se muestra la comparación entre los diferentes modelos optimizados para cada algoritmo, similar a la Tabla 3, sólo que en este caso se agregan los modelos mejorados mediante cross validation para XGBoost, Random Forest y la regresión Ridge. Adicionalmente, debido a que cada modelo entrenado puede asignar una probabilidad para cada pixel del mapa de Sudamérica original, en la Figura 11 decidimos presentar los diferentes nichos obtenidos por los algoritmos implementados en contraste con el nicho real para la especie virtual. Claramente, los algoritmos con el peor desempeño fueron la regresión lasso y el algoritmo BNM, sin embargo, vale la pena resaltar que dicho éste último algoritmo no pertenece al contexto del aprendizaje supervisado y por ende está en clara desventaja con respecto a los otros algoritmos implementados.

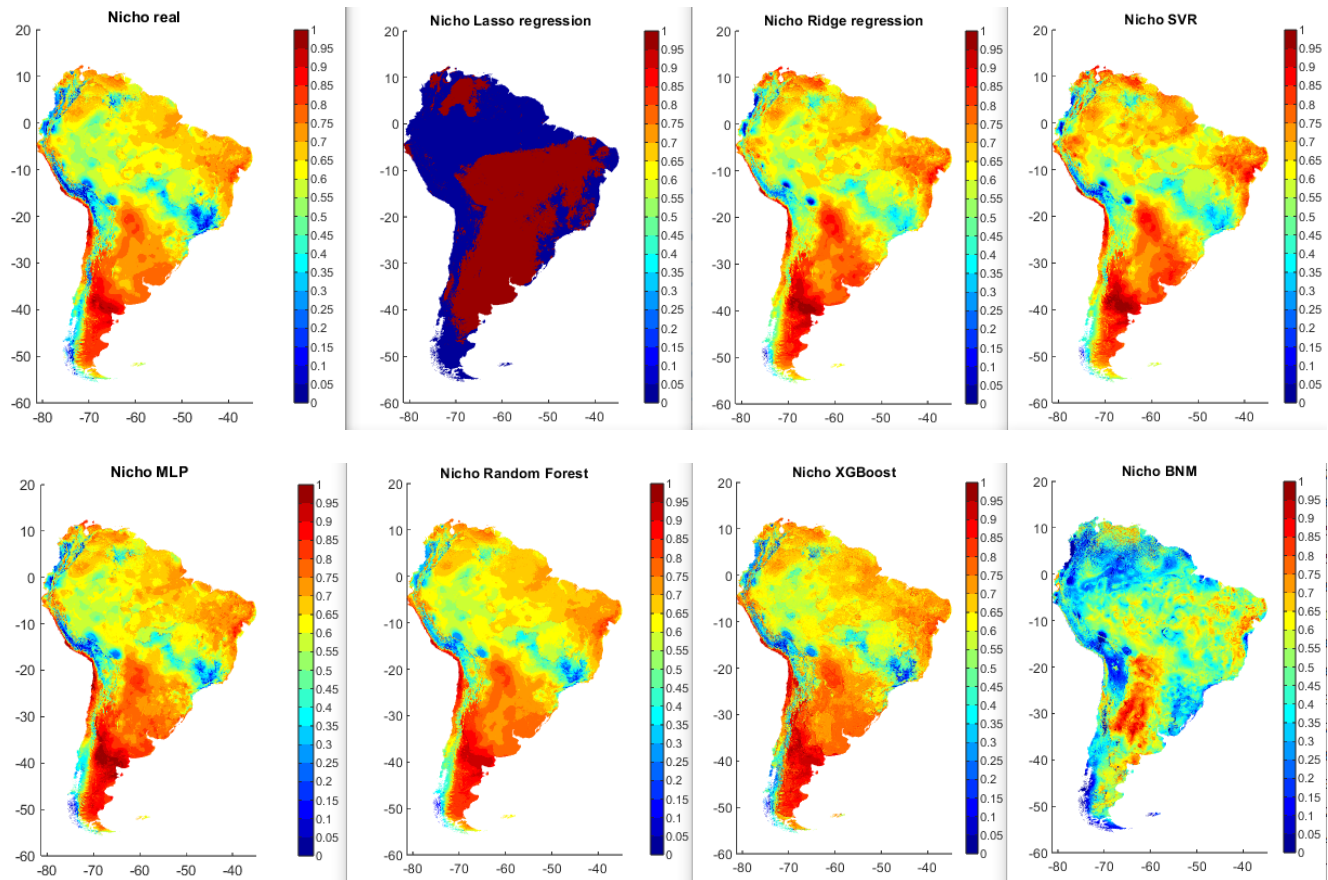


Figura 11: Nicho real generado virtualmente contrastado contra los distintos nichos inducidos por algunos de los algoritmos implementados. Puede apreciarse que la dinámica más difícil de capturar fue la de las zonas con menor idoneidad ambiental para la especie (azul oscuro en el nicho real). Los algoritmos de mayor complejidad como el MLP, el Random Forest y el XGBoost fueron los que mejor capturaron dicha dinámica.

Algoritmo	MAE	MSE	R2	Variables principales	Media	Var.	Asimet.	Curt.
<b>XGBoost Best</b>	0.00983	0.00033	0.97958	BIO10, BIO14, BIO15	-0.0005	0.0003	-2.2653	17.2793
<b>XGBoost</b>	0.01122	0.0004	0.97527	BIO10, BIO14, BIO15	0.0001	0.0004	-2.3972	21.291
<b>Random Forest Best</b>	0.01179	0.00053	0.96779	BIO10, BIO14, BIO5	-0.0015	0.0005	-3.5488	24.9988
<b>Random Forest</b>	0.01199	0.00054	0.96718	BIO10, BIO14, BIO5	-0.0015	0.0005	-3.4512	24.0623
<b>MLP (NN)</b>	0.01836	0.00066	0.95948	BIO11, BIO5, BIO12	0.0003	0.0007	-0.0753	4.681
<b>Decision Tree</b>	0.01973	0.00125	0.92382	BIO10, BIO15, BIO5	-0.0023	0.0012	-0.8696	10.6396
<b>RegresiónRidge Best</b>	0.02096	0.00091	0.94465	BIO5, BIO10, BIO8	0.0004	0.0009	-0.8439	6.9366
<b>Regresión Ridge</b>	0.02096	0.00091	0.94461	BIO5, BIO10, BIO8	0.0004	0.0009	-0.8512	6.9324
<b>Regresión lineal</b>	0.02097	0.00091	0.94467	BIO16, BIO17, BIO15	0.0004	0.0009	-0.8353	6.8769
<b>ElasticNet</b>	0.02112	0.00093	0.94336	BIO5, BIO10, BIO2	0.0006	0.0009	-0.9743	6.8689
<b>SVM</b>	0.03503	0.00189	0.88434	BIO10, BIO6, BIO9	0.0105	0.0018	-0.1316	0.3856
<b>Regresión Lasso</b>	0.09456	0.01641	-0.00305	None	-0.0071	0.0164	-0.8723	1.8581
<b>BNM</b>	0.1542	0.0028	-1.402	NA	-0.1241	0.0226	-0.1964	-0.0196

Tabla 4: Diferentes métricas y medidas para cada uno de los modelos en test.

Si bien desde el punto de vista de la métrica que implementamos (MAE) podría concluirse que casi todos los algoritmos dieron resultados casi igualmente buenos, la Figura 11 nos permite explorar con más detalle las diferencias entre el desempeño de los distintos métodos. Si bien la mayoría de los mapas del nicho predicho por los algoritmos son bastante similares al nicho real, cabe resaltar que los algoritmos de mayor complejidad, como es el caso del MLP, el Random Forest y el XGBoost, son los que mejor capturan la dinámica de la especie en ambientes de baja idoneidad (color azul) y en algunas zonas de mediana idoneidad (colores verde y amarillo). Este comportamiento puede constituirse en una gran ventaja sobre los otros algoritmos en determinados contextos de modelado que deberían explorarse en trabajos futuros.



## 7. Implicaciones éticas

Es necesario tener en cuenta que el alcance de las técnicas de predicción de nicho basadas en datos de solo presencia se extiende más allá del campo de la ecología. De manera general, estas técnicas pueden implementarse para determinar funciones de pertenencia a una clase, likelihood o probabilidad de manifestación de cierto fenómeno a partir de variables descriptoras que recopilen bajo qué circunstancias el fenómeno de interés ocurrió. En este orden de ideas, es posible que las técnicas discutidas en el presente trabajo y su posterior desarrollo tengan aplicaciones más allá del campo de la ecología que puedan afectar negativamente seres vivos a nivel individual o colectivo, por ejemplo, en el campo del pronóstico de crímenes.

Restringiéndonos al campo de acción en ecología, podemos mencionar que los investigadores estudian el nicho ecológico de una especie a través de observaciones en campo, experimentos en laboratorio y análisis de datos. Por lo tanto, al determinar el nicho de una especie a través de un modelo que presenta buenos resultados teniendo en cuenta únicamente las observaciones en campo se disminuirán considerablemente los trabajos destinados a los investigadores, ya que no se necesitará de experimentos en laboratorio. Aunque claramente esto tiene implicaciones positivas y

negativas, debido a que por un lado facilita el trabajo de los investigadores, esto implica también que ahora un modelo realizará de forma automática el trabajo diario de miles de personas. Adicionalmente, si el modelo se pone en producción y se entrenó con datos que contienen sesgos, por ejemplo, datos de muestreo no representativos o incompletos, puede conducir a resultados erróneos, lo que podría llevar a decisiones incorrectas en la conservación de las especies; igualmente este producto también podría llegar a las manos equivocadas, ya que por ejemplo cazadores o personas interesadas en comercializar fauna silvestre podrían usar estos modelos para determinar en qué lugar se encuentra determinada especie y de este modo, contribuir a la extinción de las mismas, de esta manera se debe tener cuidado con la distribución de este producto. Finalmente, los usuarios deben entender las limitaciones del modelo y no basar sus decisiones únicamente con sus resultados.

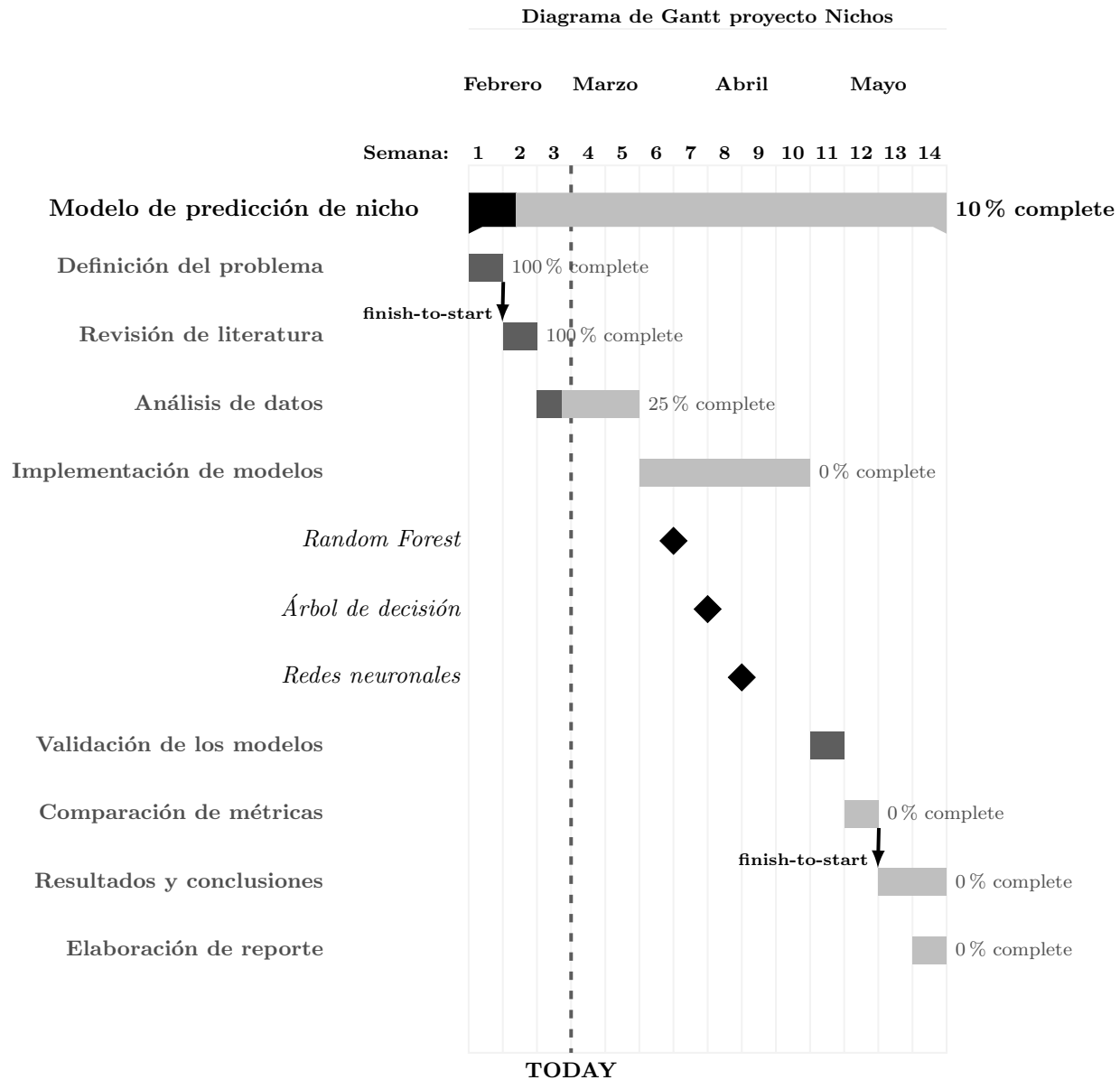
## 8. Aspectos legales y comerciales

Los datos para estas investigaciones serán simulados o serán extraídos de fuentes abiertas. En este caso particular se utilizó este repositorio abierto de Github  donde se encuentran los pasos esenciales de generación de especies virtuales y modelado de nicho BNM. Estos datos no contienen información confidencial que deba ser protegida o anonimizada. La versión final del código producido durante este proyecto junto con el conjunto de datos están publicados en el siguiente repositorio público de Github .

En primera instancia, los mayores interesados en posibles desarrollos en el campo de los SDMs son los ecólogos y biólogos de la conservación. Sin embargo, también se han evidenciado aplicaciones del modelado de nicho en epidemiología para determinar aquellas zonas con potencial para albergar especies propagadoras de enfermedades considerando la influencia del cambio climático bajo cierto horizonte temporal. Dicho lo anterior, creemos que el mayor potencial de negocio en este campo sería el desarrollo de una plataforma o software donde se pueda acceder a un modelo preentrenado desarrollado mediante aprendizaje automático que pueda, mediante algunas observaciones de una especie, predecir automáticamente el nicho de la misma. De esta forma el conocimiento necesario para construir y validar un SDM sería mucho menor y permitiría que un público más amplio pudiera acceder fácilmente a los beneficios de la técnica. Esto tendría potenciales interesados en el ámbito de la salud pública, cambio climático e impacto ambiental de proyectos. El modelo de predicción podría comercializarse como un plug-in para software especializado en información geográfica.

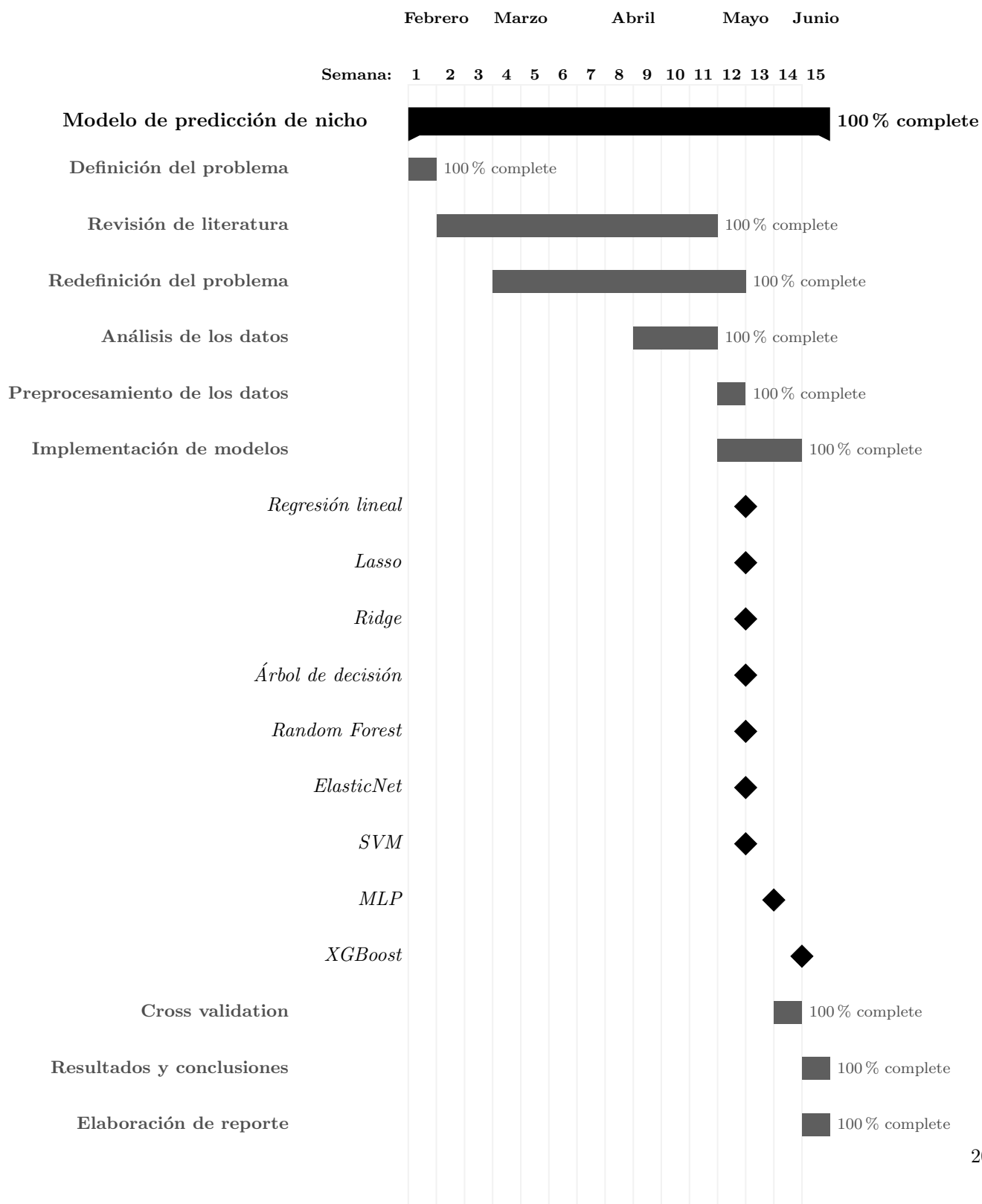
## 9. Ejecución del plan

A continuación se presenta el cronograma de actividades inicialmente planteado:



Ahora, luego de ejecutar y finalizar el proyecto, se presenta el cronograma real

Diagrama de Gantt proyecto Nichos





Es posible evidenciar que nos hizo falta un poco de organización y responsabilidad con el cumplimiento de cada una de las tareas planteadas en el cronograma de actividades inicial, las actividades que más nos retrasaron fueron la revisión de literatura y la redefinición del problema debido a que encontramos mayor complejidad de la esperada en para la consecución de nuestro objetivo original. Otra situación que contribuyó al desarrollo tardío del proyecto fue la espera de la retroalimentación de la primera entrega del anteproyecto dado que debimos haber continuado trabajando mientras llegaba la retroalimentación pero inconscientemente decidimos esperarla, y también porque durante dichas semanas debido a otras ocupaciones estuvimos un poco tranquilos. Sin embargo, en general, se cumplieron cada una de las tareas, con una mayor intensidad en las últimas semanas, pero aún así logramos el cumplir con el objetivo básico del trabajo que habíamos propuesto.

## 10. Conclusiones y trabajo futuro

El presente trabajo se realizó con el objetivo de encontrar los mejores modelos de aprendizaje estadístico, para predecir por medio de regresiones, la idoneidad ambiental asociada a una especie estudiada para una localidad de interés, en este caso, Sudamérica. Para ello, se utilizaron datos simulados de una especie virtual y se tomó como algoritmo del estado del arte al BNM, que presenta un desempeño de  $MAE=0.1542$ . Los resultados obtenidos evidenciaron que los modelos con mejor desempeño para realizar la tarea de predicción del nicho fueron XGBoost ( $MAE = 0.01122$ ) y random forest ( $MAE = 0.01199$ ). En cuanto a los modelos de regresión básica el de mejor desempeño fue la regresión Ridge con un  $MAE=0.02096$ , todos estos con un mejor resultado respecto al BNM. Vale la pena mencionar que nuestra hipótesis inicial, basada en los resultados de correlaciones obtenidos en la Figura 3, resultó comprobarse verdadera: la variable *BIO10*, que presentó la mayor correlación con nuestra variable objetivo (si bien no fue una correlación fuerte), fue la variable más relevante para predecir el nicho de acuerdo a todos los modelos donde podía definirse alguna medida de feature importance. Esto permite concluir que los análisis previos son bastante útiles para definir adecuadamente el preprocesamiento y para ganar intuición sobre el comportamiento de los datos y la naturaleza del problema.

Adicionalmente, con el objetivo de mejorar los modelos obtenidos, se realizó un proceso de validación cruzada para mejorar el MAE y así mismo evitar el sobreajuste o subajuste de los modelos. De este modo, los modelos obtenidos después de dicho proceso fueron satisfactorios logrando para XGBoost y para Random Forest una mejora de su exactitud del 0.17% y del 0.05% respectivamente. Finalmente, revisando los residuales obtenidos de los modelos realizados, encontramos que los mejores modelos (XGBoost y Random Forest) tienen un valor de curtosis cercana a 20 y una asimetría entre -2 y -4. El valor de curtosis indica que los residuales encontrados pertenecen todos a una misma distribución. Sin embargo, la asimetría negativa corrobora que hay predicciones que generaron residuales atípicos localizados a la izquierda de la media muestral. Estos resultados dejan abierta una puerta de mejora para futuras investigaciones, donde se puedan encontrar qué razones producen estos atípicos en los residuales y plantear opciones para una mejor predicción.

El mayor reto que encontramos en el desarrollo del presente trabajo fue la necesidad de plantear marcos de trabajo específicos a la hora de implementar algoritmos de aprendizaje profundo o automático para SDMs, tal y como se evidencia en los trabajos de Deneu, Servajean et al. (2021), Seo et al. (2021) y Deneu, Joly et al. (2022). Debido a la anterior dificultad, nos vimos en la necesidad de restringir nuestras metas para encajar dentro del ámbito del aprendizaje supervisado. En la práctica, no sería posible implementar este tipo de modelos porque el problema de SDMs no se puede entender, de forma inmediata, como un problema de regresión. A partir de la revisión de literatura que realizamos identificamos que los artículos con resultados más relevantes en años recientes en el campo de SDMs han optado por el uso de redes neuronales convolucionales en un contexto de clasificación utilizando múltiples especies simultáneamente, mirar por ejemplo Bourhis et al. (2023). Nosotros creemos que esta no es la mejor manera de extraer todo el potencial de las técnicas puesto que, lo que debería aprender la red convolucional estaría relacionado

a cuál es la especie que tiene mayor probabilidad de observación dado un conjunto de variables ambientales. De forma tal, ante una nueva especie cuyo nicho quiera determinarse sería necesario re-entrenar el modelo teniendo en cuenta todas las otras especies, lo cual, tal y como señalan Seo et al. (2021), puede ser computacionalmente muy costoso.

Dada la experiencia que hemos ganado durante el desarrollo del presente trabajo, creemos que una muy buena opción para permitir la implementación de técnicas de aprendizaje profundo en el ámbito de los SDMs enfocados a la predicción de nicho es el aprendizaje reforzado. En principio, debería ser posible, a partir de un generador de especies virtual, medir el rendimiento de algoritmos de SDM a través de su potencial para recuperar el nicho de la especie mediante un muestreo del mismo. Claramente, las etiquetas del nicho no deben ser accedidas durante el proceso de entrenamiento del algoritmo, es decir, el algoritmo debe producir un modelo de nicho infiriéndolo a partir de las observaciones. Posteriormente, es posible contrastar los resultados del nicho que ha sido predicho por el algoritmo contra el nicho real generado de manera virtual y utilizar la diferencia entre los resultados para generar reglas de aprendizaje en el algoritmo. Repitiendo dicho proceso sobre un conjunto suficientemente grande de especies virtuales, cada una con su respectivo nicho, generadas mediante un algoritmo que conserve un sentido biológico pero que pueda producir especies lo suficientemente extrañas y diversas entre sí, entonces debería ser posible producir un modelo que haya aprendido a extraer las posibles interacciones entre las variables ambientales sobre cierto mapa dada la presencia de la especie en algunas localidades para predecir correctamente el nicho potencial de la especie. Creemos que este sería un enfoque prometedor y novedoso ya que no hemos encontrado literatura relacionada con la aplicación de aprendizaje reforzado al modelado de nicho ni al SDM.

## Referencias

- Austin, M.P., L. Belbin, J.A. Meyers, M.D. Doherty y M. Luoto (nov. de 2006). "Evaluation of statistical models used for predicting plant species distributions: Role of artificial data and theory". En: *Ecological Modelling* 199.2, págs. 197-216. DOI: 10.1016/j.ecolmodel.2006.05.023. URL: <https://doi.org/10.1016/j.ecolmodel.2006.05.023> (vid. pág. 1).
- Bauer, Silke et al. (2019). "The grand challenges of migration ecology that radar aeroecology can help answer". En: *Ecography* 42.5, págs. 861-875. DOI: 10.1111/ecog.04083 (vid. pág. 2).
- Beery, Sara, Elijah Cole, Joseph Parker, Pietro Perona y Kevin Winner (jun. de 2021). "Species Distribution Modeling for Machine Learning Practitioners: A Review". En: *ACM SIGCAS Conference on Computing and Sustainable Societies (COMPASS)*. ACM. DOI: 10.1145/3460112.3471966. URL: <https://doi.org/10.1145/3460112.3471966> (vid. págs. 1, 2).
- Botella, Christophe et al. (2018). "A deep learning approach to Species Distribution Modelling A deep learning approach to Species Distribution Modelling [Preprint] A deep learning approach to Species Distribution Modelling". En: pág. 978. DOI: 10.1007/978-3-319-76445-0\_10. URL: <https://hal.science/hal-01834227> (vid. pág. 2).
- Bourhis, Yoann, James R. Bell, Chris R. Shortall, William E. Kunin y Alice E. Milne (2023). "Explainable neural networks for trait-based multispecies distribution modelling—A case study with butterflies and moths". En: *Methods in Ecology and Evolution* 2023.June 2022, págs. 1-12. DOI: 10.1111/2041-210X.14097 (vid. págs. 3, 21).
- Deneu, Benjamin, Alexis Joly, Pierre Bonnet, Maximilien Servajean y François Munoz (mayo de 2022). "Very High Resolution Species Distribution Modeling Based on Remote Sensing Imagery: How to Capture Fine-Grained and Large-Scale Vegetation Ecology With Convolutional Neural Networks?" En: *Frontiers in Plant Science* 13. DOI: 10.3389/fpls.2022.839279 (vid. págs. 3, 21).

- Deneu, Benjamin, Maximilien Servajean, Pierre Bonnet, Christophe Botella, François Munoz y Alexis Joly (abr. de 2021). "Convolutional neural networks improve species distribution modelling by capturing the spatial structure of the environment". En: *PLOS Computational Biology* 17.4. Ed. por Adam C. Martiny, e1008856. DOI: 10.1371/journal.pcbi.1008856. URL: <https://doi.org/10.1371/journal.pcbi.1008856> (vid. págs. 3, 21).
- Elith, Jane, Steven J. Phillips, Trevor Hastie, Miroslav Dudík, Yung En Chee y Colin J. Yates (nov. de 2010). "A statistical explanation of MaxEnt for ecologists". En: *Diversity and Distributions* 17.1, págs. 43-57. DOI: 10.1111/j.1472-4642.2010.00725.x. URL: <https://doi.org/10.1111/j.1472-4642.2010.00725.x> (vid. págs. 2, 3).
- Estopinan, Joaquim, Maximilien Servajean, Pierre Bonnet, François Munoz y Alexis Joly (abr. de 2022). "Deep Species Distribution Modeling From Sentinel-2 Image Time-Series: A Global Scale Analysis on the Orchid Family". En: *Frontiers in Plant Science* 13. DOI: 10.3389/fpls.2022.839327. URL: <https://doi.org/10.3389/fpls.2022.839327> (vid. págs. 1, 2).
- Fick, Stephen E. y Robert J. Hijmans (oct. de 2017). "WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas". En: *International Journal of Climatology* 37.12, págs. 4302-4315. DOI: 10.1002/joc.5086. URL: <https://doi.org/10.1002/joc.5086> (vid. págs. 5).
- Fitzpatrick, Matthew C., Nicholas J. Gotelli y Aaron M. Ellison (mayo de 2013). "MaxEnt versus MaxLike: empirical comparisons with ant species distributions". En: *Ecosphere* 4.5, págs. 1-15. DOI: 10.1890/es13-00066.1. URL: <https://doi.org/10.1890/es13-00066.1> (vid. págs. 1).
- Garzon-Lopez, Carol X., Lucy Bastin, Giles M. Foody y Duccio Rocchini (jun. de 2016). "A virtual species set for robust and reproducible species distribution modelling tests". En: *Data in Brief* 7, págs. 476-479. DOI: 10.1016/j.dib.2016.02.058. URL: <https://doi.org/10.1016/j.dib.2016.02.058> (vid. págs. 3).
- Gavilán, Ignacio G.R. (jun. de 2021). *Metodología Para Machine Learning (i): CRISP-DM*. URL: <https://ignaciogavilan.com/metodologia-para-machine-learning-i-crisp-dm/> (vid. págs. 4).
- Grimmett, Liam, Rachel Whithed y Ana Horta (feb. de 2021). "Creating virtual species to test species distribution models: the importance of landscape structure, dispersal and population processes". En: *Ecography* 44.5, págs. 753-765. DOI: 10.1111/ecog.05555. URL: <https://doi.org/10.1111/ecog.05555> (vid. págs. 3).
- Hastie, Trevor, Robert Tibshirani y J H Friedman (dic. de 2009). *The elements of statistical learning*. en. 2.<sup>a</sup> ed. Springer series in statistics. New York, NY: Springer (vid. págs. 11).
- Hutchinson, G. E. (ene. de 1957). "Concluding Remarks". En: *Cold Spring Harbor Symposia on Quantitative Biology* 22.0, págs. 415-427. DOI: 10.1101/sqb.1957.022.01.039. URL: <https://doi.org/10.1101/sqb.1957.022.01.039> (vid. págs. 2).
- Lee, Wang-hee, Jae-woo Song, Sun-hee Yoon y Jae-min Jung (2022). "applied sciences Spatial Evaluation of Machine Learning-Based Species Distribution Models for Prediction of Invasive Ant Species Distribution". En: (vid. págs. 2).
- Leroy, Boris, Christine N. Meynard, Céline Bellard y Franck Courchamp (jun. de 2015). "virtualspecies, an R package to generate virtual species distributions". En: *Ecography* 39.6, págs. 599-607. DOI: 10.1111/ecog.01388. URL: <https://doi.org/10.1111/ecog.01388> (vid. págs. 4).
- (2016). "virtualspecies, an R package to generate virtual species distributions". En: *Ecography* 39.6, págs. 599-607. DOI: 10.1111/ecog.01388 (vid. págs. 3).
- Merow, Cory y John A. Silander (2014). "A comparison of Maxlike and Maxent for modelling species distributions". En: *Methods in Ecology and Evolution* 5.3, págs. 215-225. DOI: 10.1111/2041-210X.12152 (vid. págs. 2, 3).
- Meynard, Christine N., Boris Leroy y David M. Kaplan (jul. de 2019). "Testing methods in species distribution modelling using virtual species: what have we learnt and what are we missing?" En: *Ecography* 42.12, págs. 2021-2036. DOI: 10.1111/ecog.04385. URL: <https://doi.org/10.1111/ecog.04385> (vid. págs. 1).
- Mouton, Ans M., Bernard De Baets y Peter L.M. Goethals (ago. de 2010). "Ecological relevance of performance criteria for species distribution models". En: *Ecological Modelling* 221.16, págs. 1995-2002. DOI: 10.1016/j.ecolmodel.2010.04.017. URL: <https://doi.org/10.1016/j.ecolmodel.2010.04.017> (vid. págs. 3, 4).

- Phillips, Steven J., Robert P. Anderson y Robert E. Schapire (2006). "Maximum entropy modeling of species geographic distributions". En: *Ecological Modelling*. DOI: 10.1016/J.ECOLMODEL.2005.03.026 (vid. pág. 2).
- Pichler, Maximilian y Florian Hartig (2023). "Machine learning and deep learning—A review for ecologists". En: *Methods in Ecology and Evolution* 2023.August 2022, págs. 994-1016. DOI: 10.1111/2041-210X.14061 (vid. pág. 1).
- Rew, Jehyeok, Yongjang Cho y Eenjun Hwang (abr. de 2021). "A Robust Prediction Model for Species Distribution Using Bagging Ensembles with Deep Neural Networks". En: *Remote Sensing* 13.8, pág. 1495. DOI: 10.3390/rs13081495. URL: <https://doi.org/10.3390/rs13081495> (vid. págs. 1-3).
- Seo, Eugene et al. (feb. de 2021). "StatEcoNet: Statistical Ecology Neural Networks for Species Distribution Modeling". En: arXiv: 2102.08534. URL: <http://arxiv.org/abs/2102.08534> (vid. págs. 2, 21, 22).
- Thuiller, Wilfried, Sandra Lavorel, Miguel B. Araújo, Martin T. Sykes e I. Colin Prentice (mayo de 2005). "Climate change threats to plant diversity in Europe". En: *Proceedings of the National Academy of Sciences* 102.23, págs. 8245-8250. DOI: 10.1073/pnas.0409902102. URL: <https://doi.org/10.1073/pnas.0409902102> (vid. pág. 1).
- Walther, Gian-Reto et al. (dic. de 2009). "Alien species in a warmer world: risks and opportunities". En: *Trends in Ecology & Evolution* 24.12, págs. 686-693. DOI: 10.1016/j.tree.2009.06.008. URL: <https://doi.org/10.1016/j.tree.2009.06.008> (vid. pág. 1).