Kalpa Publications in Computing

Volume 15, 2023, Pages 43–57

Proceedings of The 2022 International Conference on Digital Transformation and Innovation Technology

# Análisis de Alerta Temprana de la extinción de aves, usando Modelos de Machine Learning

Hernán Jácome Paneluisa, Estevan Gómez-Torres, Edgar Fernando Solis Acosta,Wellington Ernesto Valdivieso and Omar Baldeón

University of Army Forces, ESPE.

hjacome@espe.edu.ec.com, ergomez@espe.edu.ec, efsolis@espe.edu.ec, wevaldivieso@espe.edu.ec, aobaldeon@espe.edu.ec

**Abstract**

This work presents a model to realize early warning of the extinction of birds in Pichincha, considering critical factors as a climatic environmental intervention human and contamination by public and private organisms, opting for wildlife conservation plans. We have achieved a significant improvement in the methodology of trends and patterns of bird extinction with respect to manual projection processes. By using Machine Learning techniques, it is possible to obtain predictive results for early decision making. It describes the theoretical foundation used and the CRISP-DM methodology that was applied; analyze the data base to optimize the data base gradually and finally determine the model more optimally to carry out the prediction of the extinction of birds, validating the predictive model based on the red list of species of endangered birds of Ecuador emitted by the ministry of the environment; we present several conclusions.

**Keywords:** Big Data, Machine Learning , Base de datos, CRISP-DM, Minería de Datos, modelo de predicción, redes neuronales.

## 1 Introducción

La minería de datos permite realizar la extracción de conocimiento valioso que se encuentra oculto en grandes bases de datos que no han sido explorados y que permanece oculto almacenados en bases de datos históricas (Hand, 2013) (Juan, 2019) .

Lo que motiva al investigador es el descubrimiento de información que permite predecir eventos futuros. Como investigadores buscamos información nueva, innovadora, valiosa que ayude en a la toma de decisiones en varios campos como la predicción de extinción de aves. Es decir, una vez generado la predicción de nuevos resultados el investigador estará en la posibilidad de analizar los mismos. Dependiendo del tamaño de la base de datos surge el problema en depurar la misma hasta llegar a optimizar para procesar utilizando técnicas de minería de datos como la regresión linear y las redes neuronales (ALANIA RICALDI, 2019).

Los Modelos predictivos son una alternativa razonable para obtener resultados fiables a medio plazo usando diversas herramientas estadísticas, informáticas y geográficas sobre la información biológica disponible para elaborar predicciones razonables que permitan estimar la distribución de la diversidad biológica en ausencia de datos exhaustivos (Hortal & Lobo, 2003).

La presente investigación utiliza herramientas de minería de datos como regresión lineal y redes neuronales para predecir la extinción de aves, usando Modelos de Machine Learning, a fin de ayudar en la toma de decisiones gerenciales (Lorente Leyva, 2019)

# 2 Trabajos Relacionados

En esta sección se analizan algunos temas relacionados a la investigación. (Santos, Spadon, Rodriguiez, Goncalves, & Machado, 2018), identifican el Pantanal en Brasil como un ecosistema mega-diverso y toman en consideración a 8 especies en peligro de extinción. El estudio radica en realizar miles de tomas fotográficas en el Pantanal, para luego pasar por algoritmos de reconocimiento de imágenes, que una vez identificados, son sometidos bajo el algoritmo SLIC de clasificación para predecir si las especies en cuestión están aumentando o decreciendo en el habitat. Por otra parte , (Omer, Mutanga, Abdel, & Adam, 2015), trabajaron sobre una extensa área de vegetación donde existen tipos de árboles que están en peligro de extinción. Estas especies arbóreas han sido un tanto esquivas, sin embargo, al someter imágenes de alta resolución a algoritmos como máquinas de soporte vectorial y redes neuronales, han conseguido determinar la degradación del área de especies de árboles en peligro de desaparecer.

Mediante el uso de múltiples fuentes de datos y catálogos por más de tres décadas de recolección de información, tiempo en el cual se han recogido variables de vegetación de todo el globo, (Flanklin, Diaz, Syphard, & Regan, 2017), demuestran que se puede hacer predicciones de temple global donde las tecnologías de big data, han sido capaces de soportar esta carga, para analizar millones de datos de comunidades de plantas que han sufrido cambios importantes debido al cambio climático mundial utilizando series temporales. Condiderando las poblaciones de aves de Norte América que habitan según la temporada invernal o veraniega. (Distler, Schuetz, Velásquez, & Langham, 2017), buscan es determinar si el cambio climático afecta la reproducción de aves en estas 2 temporadas donde las aves aumentan o disminuyen drásticamente. Se utilizaron encuestas de aves de los dos países y se sometieron a algoritmos con datos de comportamientos o patrones históricos y los de cambio climático, sin embargo, los resultados no fueron alentadores en esta investigación y las predicciones de nacimientos de polluelos para conservar las familias de aves que residen temporalmente en los dos países, no fueron los esperados.

En su estudio “Updating Known Distribution Models for Forecasting Climate Change Impact on Endangered Species”, (Román, Márquez, & Real, 2013), propusieron actualizar los modelos de distribución de aves para predecir especies en peligro de extinción enfocándose en el cambio climático, sin embargo, los investigadores encontraron que el cambio climático no es decisivo al momento de estas predicciones, es más sugirieron que el cambio climático no es una amenaza y lo aplicaron al águila de Bonelli de España, una especie en peligro de extinción, dando como resultado que estaría volando en el siglo XXI, ellos utilizaron los modelos de distribución y propusieron mejoras optimizando los algoritmos de distribución de especies de aves con nuevos modelos de emisiones.

Finalmente, (McKeea, Sciullia, Foocea, & waitea, 2017), han tomado parámetros de cambio climático para predecir el estatus de nichos de aves en el estado de Filipinas, se consideraron 6 especies que están en peligro, y luego de someter estos parámetros climáticos a dos modernos algoritmos entrópicos como MAXENT y algoritmos genéticos como GARP, tuvieron una alta aceptación como modelos predictivos. Para mejorar los modelos introdujeron variables no climáticas de comportamiento

humano, donde se conoció una mejora en los algoritmos, dando a MAXENT la mejor calificación de predicción para nichos de aves.

# 3 Metodología

Para el desarrollo de esta investigación se utiliza la Metodología CRISP-DM , la cual se resume en las siguientes fases: Comprensión del negocio: Se realiza un estudio de la situación actual del negocio para conocer detalladamente los temas más profundos del negocio. Con estos datos, se realiza la planificación del proyecto detallando las actividades y tareas a ejecutar con los datos sobre los que se trabajará.

Comprensión de los datos: esta fase es muy importante ya que se trabaja directamente con los datos, se enfocará en la captura, exploración y calidad.

Preparación de los datos: Implica datos limpios, es decir el juego de datos que utilizarán los modelos. También se realizan estadísticas y procesos de reducción de dimensionalidad si el caso amerita.

Modelado: pone de manifiesto que el fin del modelado es cumplir tanto los objetivos del proyecto de minería como los objetivos del negocio. Se aplicarán diferentes técnicas como el uso de algoritmos de Machine Learning, es decir, se utilizarán según sea la dinámica del negocio, la segmentación, cauterización, clasificación, predicción, etc., haciendo uso de redes neuronales, máquinas de soporte vectorial, k-nn, árboles de decisión, etc., en esta fase se prueban diferentes algoritmos para obtener los mejores resultados en lugar de elegir uno solo y aumentar la confiabilidad del modelo.

Evaluación: una vez construido el modelo, se lo pone a prueba para descartar o mejorar o utilizar el modelo; este debe ser sometido a una batería de pruebas para comprobar la estabilidad y robustez del modelo a aplicar, el modelo debe cumplir con los objetivos del negocio.

Despliegue: en este punto todas las tareas destinadas al despliegue deben estar detalladas dentro del plan de despliegue; se hará un plan de mantenimiento y seguimiento del proceso. Otro paso importante es la forma de presentar los resultados, se trabaja en la distribución de resultados a los usuarios interesados. Se establecen medidores de beneficio, para demostrar la eficacia del modelo y cumplimiento de objetivos (Mingillón, Caihuelas, & Gironés, 2017) .

# 4 Resultados

A continuación, se describen los procedimientos efectuados en este trabajo de investigación:

1. Compresión del negocio: El estudio de la comprensión del negocio está enfocado en el problema de buscar un modelo que permita conocer la desaparición temprana de las especies de aves en la provincia de Pichincha debido a factores externos que alteran el equilibrio de su hábitat.
2. Comprensión de los datos: Se analizan los datos y variables relacionadas con algunas de las causas por la desaparición de especies de aves en la provincia de Pichincha. Se seleccionaron varias bases de datos abiertas proporcionadas por los organismos gubernamentales y privadas para realizar un análisis apropiado en busca de un patrón que permita discernir como los factores externos están afectando el equilibrio del ecosistema en la provincia y cuales especies de aves son las más susceptibles a estos cambios. Los pasos principales son la recolección, exploración y limpieza de datos.
3. Recolección de datos: Para la recolección de datos, se hizo una búsqueda en repositorios abiertos, descargando la información necesaria y suficiente para proceder con el

almacenamiento. La búsqueda de los datos abiertos se basó en función de los factores críticos como intervención humana, cambio climático, contaminación y en las observaciones de aves en el sitio.

4. Arquitectura de la solución: Una vez recopilada la información, se diseñó la arquitectura que operará todo el proceso de minería de datos. La especificación de la arquitectura está basada en 5 etapas que acompañan al proceso desde la carga de datos hasta la visualización de los mismos y son: Repositorio stage, Data mart, Json Data mart, Minería de datos, Exploración y visualización, como se muestra en la Figura 1.

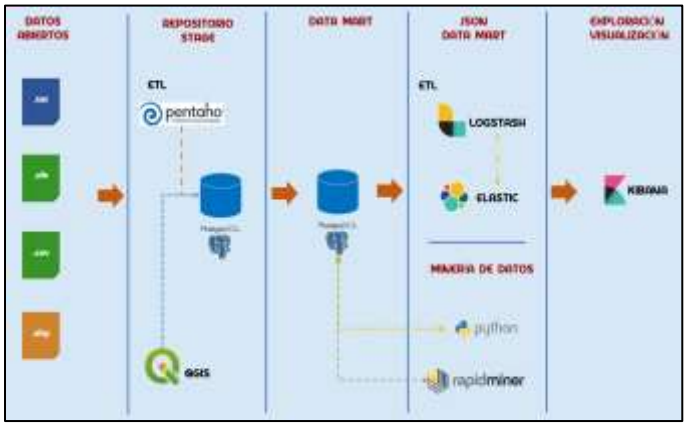


**Figura: 1** Arquitectura para la solución de la investigación

A continuación, se detalla cada una de las etapas en que está estructurada la arquitectura para la obtención del modelo de predicción.

1. Repositorio Stage: Son bases de datos que tienen información que viene de los archivos de bases de datos abiertas y son transferidos mediante un proceso ETL (con la herramienta Pentaho), en este primer proceso el ETL, solamente se encarga de pasar la información desde los archivos fuente hacia los repositorios Stage
2. Data Mart: Es un repositorio con una estructura multidimensional o modelo estrella donde mediante un proceso ETL (con la herramienta Pentaho), carga la información del repositorio Stage, transforma, estandariza los datos y finalmente carga la información en este repositorio, de manera que los datos están limpios y son el insumo para el proceso de minería de datos
3. Json Data Mart: Es un repositorio de datos en formato Json y GeoJson, el cual es alimentado por un proceso ETL (con la herramienta Logstash) que lleva información del repositorio Data Mart, transforma los datos a formato Json y GeoJson y lo carga en este repositorio (Elasticsearch) .
4. Minería de datos: Los datos de las variables preseleccionadas son consumidas directamente del repositorio Data Mart para iniciar el proceso de minado de los datos (con la herramienta Rapid Miner) en busca del modelo de predicción de alerta temprana para detectar cuando una especie de ave estaría en riesgo, dentro de la provincia de Pichincha.
5. Exploración y Visualización: La información es cargada directamente del repositorio Json Data Mart para la exploración de datos y realizar visualizaciones de los resultados obtenidos

del proceso, y la forma cómo los factores críticos han intervenido en el equilibrio del ecosistema habitados por las diferentes especies de aves de la provincia.

Cuando finalmente los datos inalterados son desembarcados en el repositorio Stage, se inicia un proceso de entendimiento de cada una de las variables que llenan el repositorio de datos, al cual se denomina exploración de datos. Se analizó la estructura de los datos, tipo de datos, y sus relaciones con otras fuentes de datos. Este proceso lleva a comprender cómo las variables se relacionan dentro del marco del contexto del problema.

Luego de tener un entendimiento de las variables en juego, se transforman en insumo para la construcción del modelo multidimensional donde residirán los datos seleccionados para el proceso de minería. Las estructuras de datos del modelo se describen a continuación:

- Dimensión especie: tiene la información específica de todas las especies que habitan la provincia de Pichincha.
- Dimensión tiempo: tiene información de fechas y el desglose en años, semestres, meses, días, etc.
- Hechos de Observación: la información es específica a los avistamientos, el lugar y fecha de esas tomas.
- Hechos de data set aves: contiene la información de las variables candidatas y observaciones que son el insumo para el trabajo de minería de datos.
- Hechos de modelos evaluación: tiene la información de los resultados de las predicciones del modelo ejecutado.
- Hechos de lista roja: tiene información del nivel de riesgo de las especies del Ecuador.

Se diseña un proceso ETL que, en su fase de transformación, realizará todas las validaciones y transformaciones de datos necesarias para llegar a la calidad del dato que se necesita para cargar la información al modelo dimensional, donde estarán disponibles para la preparación de datos.

Se prepara los datos al punto de generar una data set o conjunto de datos que serán utilizados como insumo en el modelamiento de datos, para iniciar un proceso de construcción de modelos de predicción que puedan dar el resultado propuesto en la hipótesis.

El conjunto de datos de avistamientos tiene variables espaciales que se pueden representar en un mapa. Los datos de observaciones tienen un sesgo que está limitado a la fecha en la que se tiene la observación y la especie que se está avistando, por lo que los avistamientos no son diarios, y está a discreción del observador. Esta data set tiene muchas observaciones en 0, lo que significa que, en el momento de las lecturas, no se obtuvo avistamiento de la especie en el momento, como se muestra en la Figura 2.

**Figura: 2** Distribución física de las especies de aves en hábitat

Una revisión rápida puede suponer que los datos son correctos, sin embargo, los datos deben ser sometidos a un proceso de estadística descriptiva para analizar los datos. La exploración indica que hay outliers en la variable “observación”, es decir hay valores extremos que salen de la escala. El valor más extremo indica que se trata de la especie cuyo nombre científico es "Bubulcus ibis" del orden las "Pelecaniformes", como se muestra en la Figura 4.

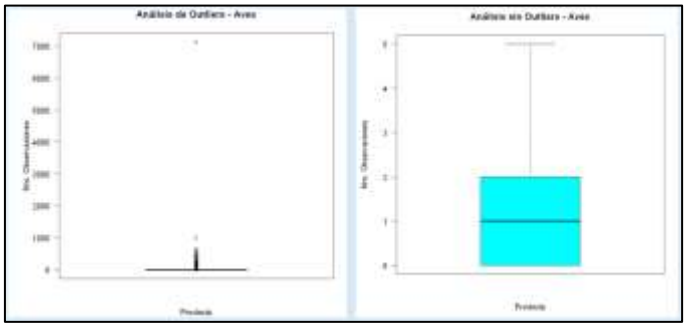

**Figura: 3** Valores atípicos en la data set de especies

La correlación entre variables, permite conocer la relación directa o inversa que tienen todas las variables entre sí. Mientras más alto es el valor de la correlación, es decir mientras más cercano a 1 es el coeficiente de correlación, más fuertemente relacionadas las variables se encuentran, y esto determina la condición de excluir o no del modelo una u otra variable. Esto significa que se tendría el mismo resultado utilizando una o las dos variables. En la gráfica se muestra variables con un coeficiente de correlación elevado, como se muestra en la Figura 4.

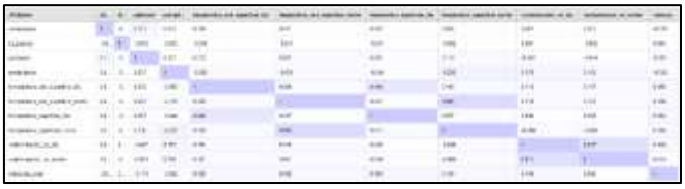

**Figura: 4** Matriz de correlación de variables

Posteriormente se construyen variables a partir de otras, en este caso se tomaron en cuenta variables ambientales que son críticas para el equilibrio del ecosistema en que viven las diferentes especies de aves en la provincia. Es decir que se ha estructurado el conjunto de datos según los factores críticos como intervención humana, cambio climático, contaminación. Los cruces de información de forma general se dieron mediante variables propias de observaciones de aves, que son el objetivo de estudio, las variables de ambiente, de población y contaminación. Esta data set o conjunto de datos, es el seleccionado para trabajar el modelamiento de datos. Se pude observar que se encuentran las variables correlacionadas, sin embargo, son omitidas en esta fase de modelamiento, dado que son variables que no aportan valor al modelo.

Se utiliza Rapidminer para el desarrollo de la investigación (RapidMiner, 2021), Una de las funcionalidades de gran ayuda es el “auto model”, el cual analiza la data set seleccionado y sugiere cuales modelos son los mejor puntuados para predecir en función de las variables y datos entregados. En este apartado, se utilizó esta funcionalidad para conocer rápidamente que modelos son los más mocionados para trabajar con la data set propuesto y afinar los modelos. Cada método asigna un puntaje o peso a cada variable, si considera que será importante como variable independiente del modelo. El

resultado de la ejecución del modelo, indica cuales son las variables adecuadas y cuáles son inadecuadas. Se retornan valores 1 y 0 respectivamente, como se muestra en la Figura 5.

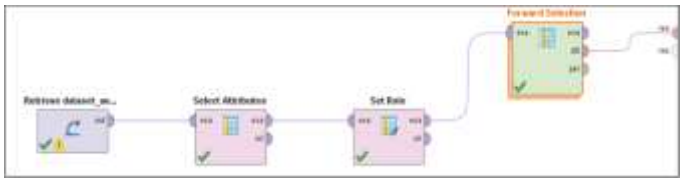

**Figura: 5** Método Forward para selección de variables

Este modelo da importancia a las variables de población, contaminación en el día, radiación solar, temperatura en la superficie en el día, y la radiación ultravioleta. El método indica que el resto de las variables solo provocarán ruido o impactarán negativamente en el desempeño del modelo de predicción.

Con el método Backward, solo el atributo que da la menor disminución de rendimiento finalmente se elimina de la selección. Luego se inicia una nueva ronda con la selección modificada" (RapidMiner, 2021). Las variables con alto rendimiento tomaran el peso de 1, el resto con peso 0. El resultado de tomar todas las variables y compararlas al inicio de las iteraciones, muestra que todas las variables, excepto una, son candidatas idóneas para formar parte de los modelos de predicción.

El Método Optimize Selection (Brute Force), trabaja en el conjunto de potencia del conjunto de atributos, tiene un tiempo de ejecución exponencial". Se debe utilizar con prudencia este método ya que puede tardar exageradamente mucho tiempo, mientras más variables se utilicen al comparar, adicionalmente, tiene la misma respuesta en cuanto a que variables tienen buenos rendimientos, respecto al modelo backward.

Método Optimize Selection (Evolutionary), optimiza el algoritmo de selección para las variables elegidas, es decir "utiliza heurística de búsqueda que imita el proceso de evolución natural. Esta heurística se utiliza habitualmente para generar soluciones útiles a los problemas de optimización y búsqueda (RapidMiner, 2021).

Una vez definidos los modelos y/o afinados, estas variables, pasarán a formar parte de los modelos finales, el resultado de la ejecución de los métodos de la selección se presenta en la Tabla 1:

| Variable | Campo | Estado |
|---|---|---|
| **Temperatura de la superficie en el día** | temperatura_superficie_dia | Aceptada |
| **Radiación ultravioleta** | radiacion_ultravioleta | Aceptada |
| **Radiación solar** | radiacion_solar | Aceptada |
| **Contaminación** | contaminacion_co_dia | Condicionada |
| **Población** | poblacion | Condicionada |

**Tabla 1:** Variables Seleccionadas Para Los Modelos

Las variables condicionadas, participarían en los modelos, siempre que no alteren el comportamiento de los modelos. En este análisis, han sido previamente probados con la herramienta Rapidminer (RapidMiner, 2021), que permite someter los datos a varios algoritmos para establecer de manera más efectiva cuáles son los mejores algoritmos de predicción que se adapten al conjunto de datos. Los modelos seleccionados se escogieron por cada tipo de relación, dado que se adaptan en similares condiciones y se escogerá uno de cada categoría, lo cual se presenta en la Tabla 2.

| Modelo | Relación | Selección |
|---|---|---|
| **Deep Learning** | Neurona | Si |
| **Generalized Linear Model** | Ecuación | Si |
| **Gradient Boosted Tree** | Árboles | Si |
| **Random Forest** | Árboles | No |
| **Decision Tree** | Árboles | No |

**Tabla 2:** MODELOS DE PREDICCIÓN

La construcción genérica de los modelos está conformada por procesos concatenados para lograr un modelo adecuado. El data set de entrada sirve tanto para entrenamiento y pruebas para el aprendizaje de los modelos, en este caso la división será 70/30, 70% de los datos para entrenamiento y 30% para pruebas. El siguiente proceso realiza una mezcla de los datos, con el fin de barajar muy bien los datos. Luego, con una data set de variables con distintas unidades de medida, se normalizan los datos y se seleccionan las variables que entrarán a formar parte del modelo. Después para los datos de entrenamiento, se establece un rol del campo al ser predicho, en este caso, para todos los modelos, es la variable observación. El siguiente paso es el uso del modelo seleccionado, cuya salida será directamente al modelo dimensional. A continuación, se muestran en la Figura 6, el diseño de los modelos realizados para la investigación.

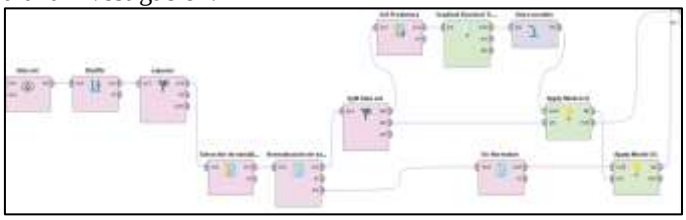

**Figura: 6** Diseño del modelo de regresión Gradient Boosted Tree

El modelo GBT, fue optimizado con un número de 100 árboles, según la especie y la media, dado que valores bajos en la media, exigen mayor afinamiento. Una profundidad de 4, con un learning rate de 0.1 y una función de distribución de Poisson lo cual nos dará mejores resultados. Posteriormente el modelo GLM, fue optimizado usando una distribución de Poisson. A continuación, en el modelo Deep Learning se usa la distribución de Poisson y una función de activación Tanh, con 3 capas ocultadas de 30 neuronas cada capa, una rho de 0.999.

En esta fase se somete a prueba cada uno de los modelos y las variables asociadas para determinar la robustez de cada uno de los modelos seleccionados, usando Deep Learning, Modelo Lineal Generalizado, Árboles potenciados por gradiente (RapidMiner, 2021).

A continuación, se presenta uno de los modelos probados con una especie elegida aleatoriamente, en la Figura 7, se muestra la Comparación de ajuste-efectividad entre modelos

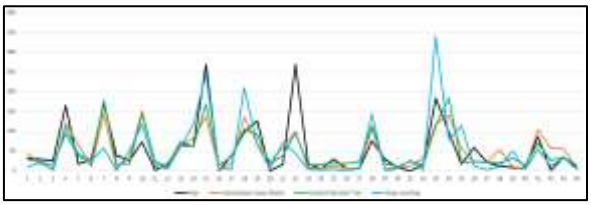

**Figura: 7** Comparación de ajuste-efectividad entre modelos

La mayoría de las especies probadas con los modelos seleccionados, dan el mismo comportamiento que se observa en las gráficas. Se puede ver que la curva de color negro está graficada con los datos de prueba, mientras que el resto de curvas, corresponde a los valores predictivos lanzados por cada modelo de Machine Learning. Para conocer cual se ajusta más a la curva real, es necesario conocer el error que cada una presenta, utilizando lo modelos GLM, Deep Learning, GBT.

El resultado de la evaluación del modelo Deep Learning, indica un error bajo tanto MSE y el RMSE, un R-cuadrado muy bueno como candidato para la ejecución del modelo final, mostrado en la Figura 17.

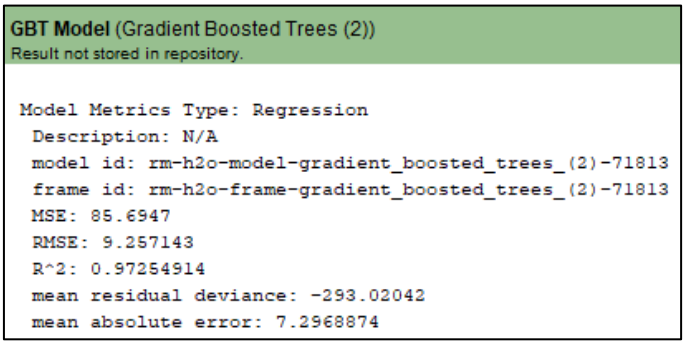


**Figura: 8** Resultado de pruebas del modelo GBT

El resultado del GBT, muestra al igual que el modelo Deep Learning, datos bajos de error, haciendo de estos modelos eficaces para los objetivos de la investigación. Para evaluar los modelos de regresión en aprendizaje supervisado, se utiliza el error cuadrático medio MSE, como medida de evaluación, la raíz del error cuadrático medio RMSE y el R-cuadrado.

Cálculo de error cuadrático medio y raíz del error cuadrático medio

$$MSE = \frac{1}{|D|}\sum_{d \in D}(f(d) - h(d))^2 \qquad RMSE = \sqrt{MSE}$$

Nota: Fórmula de cálculo del error cuadrático. (Mingillón, Caihuelas, & Gironés, 2017)

Estos cálculos son generados directamente por la herramienta Rapidminer. Los resultados se presentan en la Tabla 3

| Criterio | GLM | GBT | DL |
|---|---|---|---|
| Error cuadrático medio | 866.38 | 85.69 | 628.01 |
| Raíz del error cuadrático medio | 29.43 | 9.25 | 25.08 |
| R-cuadrado | 0.72 | 0.97 | 0.80 |
| Error absoluto medio | 20.90 | 7.29 | 15.90 |

**Tabla 3**: EVALUACIÓN DE LOS MODELOS DE REGRESIÓN CANDIDATO

Se opta por tomar el de menor raíz error cuadrático medio RMSE, y un alto R-cuadrado. siendo el modelo GBT en este caso, que tiene una precisión de (100-7.2, el valor del error absoluto), es decir 92,7%, que es el candidato para trabajar con los objetivos planteados en esta investigación. Las variables

seleccionadas que se ajustan mejor al conjunto de datos que darán los mejores resultados de predicción, quedando el resto de las variables fuera del estudio, son:

- Temperatura de la superficie en el día
- Radicación solar
- Radiación ultravioleta

Con el modelo utilizado, se generó una base de datos de predicciones que muestran la tendencia de la existencia de presencia o ausencia de especies de aves en la zona de Pichincha, a lo largo de 14 años a futuro, con capacidad para responder la hipótesis planteada de alertar tempranamente, cuando una especie estaría en riesgo, según la tendencia en las predicciones.

El despliegue de esta solución sigue en todo aspecto en función de la arquitectura propuesta y los pasos a realizar en cada uno de ellos. La arquitectura está dividida en 5 grandes procesos o capas principales.

1. Capa de datos: La recopilación de la información se basa en los criterios de recopilación definidos en este documento.
   - Datos de temperatura, radiación, contaminación y precipitación fue obtenida de la plataforma Geovanni de la Nasa en formato csv.
   - Datos de especies, se obtuvo de la base mundial Avibase en formato txt.
   - Datos de Pichicha geográficos y de población, se obtuvo del INEC de Ecuador, en formato shape y csv
2. Capa Repositorio Stage: Esta capa es importante ya que maneja los procesos ETL que además de cargar información de una fuente y depositarla en otra, sirve para tratar la información antes de ser despachada. Los procesos ETL fueron desarrollados con la herramienta Spoon de Pentaho. Estos procesos suben los archivos txt, json y csv., que fueron obtenidos en la capa de recopilación, a la base de datos denominada Stage. La base de datos que maneja esta capa es PostgreSQL v12
3. Capa Data Mart: Esta capa, maneja un modelo dimensional donde los datos tratados son subidos para mantener el histórico de la data que maneja la investigación. Esta base de datos tiene dimensiones como variables especies, ambiente, geográficas. Y los hechos, son tablas que contienen información de la data set tratado que es consumido por los modelos de Machine Learning. La base de datos que maneja el modelo dimensional es PostgreSQL v12
4. Capa Json Data Mart : Esta capa contiene otra base de datos en formato Json que obedece tanto a la arquitectura como la solución Big Data propuesta. Esta capa maneja una solución Big Data de Elasticsearch, donde se almacena toda la información que viene del Data mart, pero en formato json.
5. Capa Minería de Datos: Esta capa es la más importante de la investigación. En esta capa se encuentran los modelos de datos creados, que toman los datos de las variables seleccionadas y permiten entregar los resultados de predicción de los modelos y descargarlos directamente hacia el Data mart, para luego ser transportados a la solución de Big Data. Los modelos de predicción fueron creados utilizando la herramienta Rapidminer con la versión de evaluación académica propuesto por la misma empresa de Rapidminer.
6. Capa Visualización: Los resultados y los datos que acompañan a toda la información de la investigación son volcados a una interfaz de visualización de datos, donde la información es presentada en forma gráfica. Un usuario final puede interpretar de forma clara y rápida los

resultados de la investigación y centrarse en tareas que van más allá de toda la investigación, como es alertar tempranamente si alguna especie de ave, pude estar o no en peligro, o tomar medidas adecuadas con la sociedad.

Todos los procesos que acompañan a esta solución pueden desplegarse en un ambiente de producción, y se puede proponer al menos 2 nodos para la solución Big Data, 1 para la solución Stage y Datamart, y finalmente 1 de almacenamiento o de recopilación de información de data no procesada.

# 5 Discusión

El objetivo de esta investigación es tener un algoritmo que permita alertar tempranamente cuando una especie estaría en peligro, para tomar medidas a tiempo. Se decide realizar una proyección de datos de las variables independientes hasta el año 2030, como instrumento para la prueba de hipótesis, mediante algoritmos de regresión, para obtener el nuevo data set requerido.

La Provincia de Pichincha tiene un área biodiversa muy rica en especies de todo tipo de animales, solamente en aves se tiene más de 700 especies diferentes. En esta investigación se seleccionan, por la magnitud de especies, una muestra, para la evaluación del modelo de regresión, y dado que se tienen coeficientes diferentes por cada una de las especies de aves. Las especies seleccionadas se listan en la Tabla 4.

| Nombre Científico | Nombre Común |
|---|---|
| **Morphnarchus princeps** | Gavilán Barreteado |
| **Spizaetus isidori** | Águila Andina |
| **Lafresnaya lafresnayi** | Colibrí Terciopelo |
| **Vultur gryphus** | Cóndor Andino |
| **Cephalopterus penduliger** | Pájaro Paraguas Longuipéndulo |

**Tabla 4:** ESPECIES SELECCIONADAS PARA EJECUCIÓN DE MODELO

El modelo predictivo para el gavilán barreteado muestra una gráfica con una curva constante. Desde el 2017 al 2030, las predicciones oscilan entre 0 y 30 ejemplares. Al comparar los datos históricos, la curva histórica muestra una tendencia al alta especialmente los últimos 4 años, mientras que la curva predictora muestra una tendencia más constante, pero con mayores avistamientos respecto al histórico, hasta finales de esta década, mostrado en la Figura. 18.

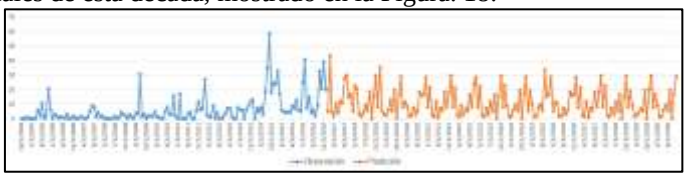

**Figura: 9** Curva de predicción para Gavilán Barreteado

Una comparativa de la estadística básica de las curvas histórica y predicha, muestran que, a futuro en promedio, se tendrán más avistamientos y con mayor número en cada observación, con tomas menos dispersas y un pico máximo de 43 ejemplares. Una comparativa de la estadística básica de las curvas histórica y predicha, muestran que, a futuro en promedio, se tendrán más avistamientos y con mayor número en cada observación, con tomas menos dispersas y un pico máximo de 43 ejemplares, mostrado en la Tabla 5, una estadística comparativa entre el histórico y la predicción para el Gavilan barreteado.

| Medida | Histórico | Predicción |
|---|---|---|
| **Media** | 6.29 | 11.35 |
| **Mediana** | 3 | 8.26 |
| **Desviación estándar** | 9.57 | 2.29 |
| **Min** | 0 | 0.01 |
| **Max** | 59 | 43.47 |
| **Rango** | 59 | 43.46 |
| **Cantidad** | 924 | 1907.61 |

**Tabla 5:** COMPARATIVA PARA EL GAVILÁN BARRETEADO

La predicción del modelo indica que en promedio se mantendrán observaciones constantes desde el 2017 al 2030, el pico más alto es de 4 ejemplares, se estima observaciones hacia finales del 2030, en cuanto a águilas andinas, en cuanto a la ejecución del modelo Lafresnaya lafresnayi para colibrí terciopelo, la curva del modelo tiene una tendencia con menos picos que la histórica, la curva se observa más constante, pero tiene medidas poco más altas entre 2021, 2022 y 2027.

La estadística indica que en promedio se tendrá menos observaciones de ejemplares por año que el histórico; observaciones con menor número de ellos en cada toma, menos dispersos por año, con observaciones de tendencia decreciente, pero constante hasta el fin de la década y con menor número de individuos contabilizados frente al histórico.

El análisis de los resultados se contrasta con la lista roja de aves del Ecuador (Freile, J. F., T. Santander G., G. Jiménez-Uzcátegui, L. Carrasco & E. A. Guevara, 2019) el libro categoriza a las especies de aves en función de la evaluación del riesgo de valuación. El libro, indica mediante el tipo de amenaza de cada especie, donde 1 es la de mayor preocupación y 7 de menor preocupación; para los tipos 8 y 9, para esta investigación no se tomará en cuenta, dado que los modelos de machine Learning, necesitan de un número mínimo de datos, mostrado en la Tabla 6.

| Categoría | Tipo | Amenazada |
|---|---|---|
| **Regionalmente Extinta** | RE | 1 |
| **Críticamente Amenazada-Posiblemente Extinta** | CR-PE | 2 |
| **Críticamente Amenazada** | CR | 3 |
| **En Peligro** | EN | 4 |
| **Vulnerable** | VU | 5 |
| **Casi Amenazada** | NT | 6 |
| **Preocupación Menor** | LC | 7 |
| **Datos Deficientes** | DD | 8 |
| **Especies no Evaluables** | NE | 9 |

**Tabla 6:** CATEGORÍAS DEL LIBRO ROJO DE AVES

Al cruzar las especies seleccionadas con los datos de la lista roja, se tiene la evaluación del riesgo que acompaña a cada especie, según lo cual el Cóndor Andino cuyo nombre científico es Vultur gryphus está en peligro, además el Águila Andina cuyo nombre científico es Spizaetus isidori se encuentre críticamente amenazada con un nivel de valoración 3. a diferencia del colibrí terciopelo que con una valoración de 7 es de menor preocupación.

Los resultados muestran además que la relación por años del comportamiento histórico de avistamientos y la tendencia del modelo de Machine Learning aplicado, donde los datos históricos

comprenden los años desde el 2004 hasta el 2016 y la predicción de los modelos comprenden los años desde 2017 hasta 2030, como se puede observar en la Figura 10, un análisis comparativo de la evolución histórica y la predictiva anual de las especies seleccionadas a través de los años.

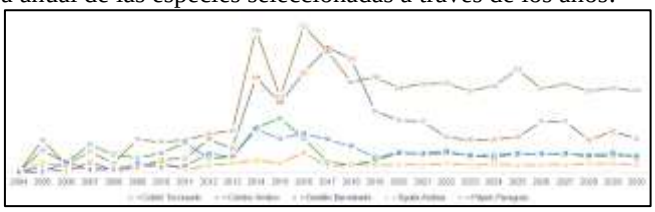

**Figura: 10** Evolución histórica y la predictiva anual de las especies

Según la gráfica, la especie con la curva más plana se da para el águila, mientras que la más dinámica es la del gavilán, todas las curvas muestran que habrá avistamientos de ejemplares hasta el final de esta década. Se observa que en promedio todas tienen un considerable número de avistamientos históricos en los años 2013, 2014, 2015 y 2016, es decir la evolución histórica anual de avistamientos y predicciones para el Águila Andina mostrado en la Figura. 11.

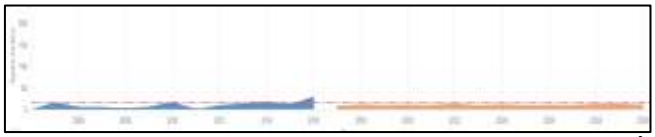

**Figura: 11** Evolución histórica anual de avistamientos y predicciones para el Águila Andina

# 6 Resultados

Los modelos de predicción ayudan a la ciencia y la sociedad a responder cuestiones de fines específicos y que probabilidades podrían acercarse más a la realidad en el futuro. En este caso pronosticar el número de avistamientos de ejemplares en el hábitat de las especies observadas hasta el 2030.

La técnica utilizada para este fin fue considerar algoritmos de Machine Learning, para la investigación se usó Gradient Boosted Tree, una técnica de regresión para el modelo final de alerta temprana. Las variables independientes que conforman las modelos son temperatura en el día a nivel de superficie, radiación solar y radiación ultravioleta.

La infraestructura utilizada se basa en el diseño de una arquitectura para Big Data con capacidad de escalabilidad si el requerimiento existiese.

Los resultados del modelo de alerta temprana se cruzaron con las categorizaciones dadas por la última publicación del libro rojo de especies del Ecuador, publicado en 2019 y se determinó que las especies conservan su estatus de riesgo actual hasta el fin de esta década.

Resumiendo, los resultados del modelo, se puede decir que las variables utilizadas arrojan tendencias dentro de un marco normal. De las 5 especies seleccionadas como objetivo, todas centran las predicciones sobre la misma categoría de riego con data histórica. Se mantienen las mismas alertas para las especies investigadas, a excepción del colibrí terciopelo cuyo modelo anticipa una menor frecuencia de avistamientos de ejemplares y del gavilán barreteado que muestra una tendencia creciente a diferencia del histórico. La solución se completa con los resultados arrojados por el modelo de predicción a un dashboard o visualización donde se muestran las predicciones para las diferentes especies, donde un usuario, puede observar el comportamiento de especies a futuro y tomar decisiones que ayuden a recuperar a especies en riesgo, conocer cuales tienen alguna tendencia a reducir la población.

# 7 Conclusiones y Trabajos Futuros

Este trabajo se basó en la metodología de caso de uso para toda la investigación, y para la parte de minería de datos, se utilizó la metodología de CRISP-DM, específicamente para este proceso. La parte de minería de datos siguió el diseño de la arquitectura propuesta para entornos Big Data.

La selección de los datos se obtuvo de diferentes fuentes como Avibase, Satélite de la NASA mediante web Geovanni e INEC, estos datos fueron cargados a un repositorio Stage mediante procesos ETL, que finalmente desembarcaron hacia el Data mart. El modelo dimensional se realizó para almacenar los datos limpios como insumo para los modelos de Machine Learning.

Las variables independientes que mostraron mejor capacidad de predicción para el conjunto de datos de observaciones de aves son temperatura de la superficie en el día, radiación solar y radiación ultravioleta, como resultado de la ejecución de métodos de selección de variables.

El modelo que mostró el mejor ajuste y menor error en las pruebas de rendimiento de modelos fue Gradient Boosted Tree (GBT) con el 92% de efectividad.

El modelo predictivo se comparó con la condición de riesgo de las especies, la que está catalogada en el libro rojo de especies de Ecuador, donde el modelo predictivo utilizado en esta investigación para las especies seleccionadas no mostró alteraciones o cambios en la categoría de riesgo a ninguna especie, al contrario, mostró una curva predictiva por años, basada en las observaciones de especies, acorde a los datos históricos como insumo de alerta temprana.

Los resultados obtenidos de la ejecución del modelo de predicción para el mismo periodo de 13 años que el histórico, desde inicios del 2017 hasta final del 2029, muestran que para el águila andina para este periodo, subirá un 2.8% más de avistamientos, en el caso de cóndor, se avistarán un 93% más ejemplares, el pájaro paraguas se avistará un 37% mayor, el colibrí terciopelo tendrá una reducción del -30% en los avistamientos, mientras que el gavilán barreteado podrá observarse un 93% más hasta finales del 2029.

Al final del 2030 se estima que el número de avistamientos en ese año, será para el águila andina de 11 avistamientos de estos ejemplares, unas 51 observaciones de cóndores andinos, 124 avistamientos del gavilán barreteado, 22 de pájaros paraguas y 25 del colibrí terciopelo. Estos datos indican que al final del 2030 si se contará con estas especies, pero con tendencia a la baja, sin embargo, la especie que tiene un número muy bajo de avistamientos es el águila andina donde el modelo no muestra un incremento de avistamientos y arroja una tendencia similar al histórico.

Se recomienda utilizar el estudio de esta investigación como base para futuros análisis como por ejemplo la proyección de extinción de aves en diversas provincias del Ecuador.

Generar proyecciones futuras de la extinción de los diversos tipos de aves en el Ecuador que permita al ministerio del medio ambiente tomar decisiones a tiempo.

# Referencias

ALANIA RICALDI, P. F. (10 de Enero de 2019). *Aplicación de técnicas de minería de datos para predecir la deserción estudiantil* . (Universidad Nacional Daniel Alcides Carrión) Obtenido de http://repositorio.undac.edu.pe/handle/undac/829

Distler, Schuetz, Velásquez, T., & Langham. (2017). *Modelos de distribución de especies*.

Flanklin, Diaz, S., Syphard, & Regan. (2017). *Big data para pronósticar los cambios globales*. Recuperado el 5 de 10 de 2021

Hand, D. (15 de Enero de 2013). *Minería de Datos*. Recuperado el 15 de Julio de 2019, de https://onlinelibrary.wiley.com/doi/abs/10.1002/9780470057339.vad002.pub2

Hortal, J., & Lobo, J. (Enero de 2003). *Modelos predictivos: Un atajo para describir la distribución de la diversidad biológica.* Obtenido de https://www.researchgate.net/publication/26495204_Modelos_predictivos_Un_atajo_para_describir_la_distribucion_de_la_diversidad_biologica

Juan, U. N. (Abril de 2019). *Minería de datos, minería de textos y Big Data*. Obtenido de http://sedici.unlp.edu.ar/handle/10915/77022

Lorente Leyva, L. L. (26 de Marzo de 2019). *Modelo para el pronóstico de la demanda de la empresa Dipac Manta S.A*. Obtenido de http://repositorio.utn.edu.ec/handle/123456789/8938

McKeea, Sciullia, Foocea, & waitea. (2017). *Modelo de nicho de aves Pilipinas en peligro*.

Mingillón, J., Caihuelas, R., & Gironés, J. (2017). *Minería de datos.* Oberta UOC.

Omer, Mutanga, Abdel, R., & Adam. (2015). *Desempeño de maquinas de vectores de oporte y redes neuronales artificiales.*

RapidMiner, R. (2021). *Manual de RapidMiner*. Obtenido de https://academy.rapidminer.com/learning-paths/get-started-with-rapidminer-and-machine-learning

Román, Márquez, & Real. (2013). *Modelos de distribución para pronósticar el impacto del acmbio climático en especies*.

Santos, Spadon, Rodriguiez, Goncalves, & Machado. (2018). *Reconocimiento de especies animales del Pantanal en peligro de extinción.* Recuperado el 10 de 01 de 2022