ESPE

UNIVERSIDAD DE LAS FUERZAS ARMADAS

INNOVACIÓN PARA LA EXCELENCIA

Desarrollo de un modelo de predicción de extinción de aves en Pichincha - Ecuador mediante técnicas de minería de datos

Coyago Remache, Diego Fernando

Vicerrectorado de Investigación, Innovación y Transferencia de Tecnología

Centro de Posgrados

Maestría en Gestión de Sistemas de Información e Inteligencia de Negocios

Trabajo de titulación, previo a la obtención del título de Magíster en Gestión de Sistemas de Información e Inteligencia de Negocios

Msc. Jácome Paneluisa, Hernán

21 de agosto del 2021

Informe de originalidad

NOMBRE DEL CURSO
Revisión Tesis_1

NOMBRE DEL ALUMNO
DIEGO FERNANDO COYAGO REMACHE

NOMBRE DEL ARCHIVO
DIEGO FERNANDO COYAGO REMACHE - Tesis

SE HA CREADO EL INFORME
25 nov 2021

Firmado electrónicamente por: HERNAN JACOME

Resumen

Fragmentos marcados	9	1 %
Fragmentos citados o entrecomillados	2	0,2 %

Coincidencias de la Web

audubon.org	1	0,2 %
dataone.org	1	0,2 %
unir.net	1	0,2 %
wiley.com	1	0,1 %
slideplayer.es	1	0,1 %
nih.gov	1	0,1 %
ieee.org	1	0,1 %
researchgate.net	1	0,1 %
semanticscholar.org	1	0,1 %
books.google.com	1	0,1 %
plos.org	1	0,1 %

1 de 11 fragmentos
Fragmento del alumno MARCADO

Trabajo de titulación, previo a la obtención del título de Magíster en Gestión de Sistemas de Información e Inteligencia de Negocios

Mejor coincidencia en la Web

TRABAJO DE TITULACIÓN, PREVIO A LA OBTENCIÓN DEL TITULO DE MAGISTER EN GESTIÓN DE SISTEMAS DE INFORMACIÓN E INTELIGENCIA DE NEGOCIOS *ANÁLISIS PARA PREDICCIÓN.

trabajo de titulación, previo a la obtención del título de magister en ... https://slideplayer.es/slide/13381355/

---

2 de 11 fragmentos
Fragmento del alumno MARCADO

Recognition of Endangered Pantanal Animal Species using Deep Learning Methods.

Mejor coincidencia en la Web

Recognition of Endangered Pantanal Animal Species using Deep Learning Methods. Abstract: Pantanal is one of the most important biomes of the world, with a large number of wild animal species, some of...

Recognition of Endangered Pantanal Animal Species using Deep ... https://ieeexplore.ieee.org/document/8489369

---

3 de 11 fragmentos
Fragmento del alumno CITADO

Artificial neural network (ANN), **Dukuduku, endangered tree** species (ETS), **indigenous forest, support vector** machines (SVM).

Mejor coincidencia en la Web

Index Terms—Artificial neural network (ANN), **Dukuduku,. endangered tree** species (ETS), **indigenous forest, support vector.** machines (SVM).

(PDF) Performance of Support Vector Machines and Artificial Neural ... https://www.researchgate.net/publication/281208945_Performance_of_Support_Vector_Machines_and_Artificial_Neural_Network_for_Mapping_Endangered_Tree_Species_in_Dukuduku_Forest_South_Africa

---

4 de 11 fragmentos
Fragmento del alumno MARCADO

VICERRECTORADO DE INVESTIGACIÓN, INNOVACIÓN Y TRANSFERENCIA DE TECNOLOGÍA

CENTRO DE POSGRADOS

CERTIFICACIÓN

Certifico que el trabajo de titulación, "**Desarrollo de un modelo de predicción de extinción de aves en Pichincha - Ecuador mediante técnicas de minería de datos**" fue realizado por el señor Coyago Remache, Diego Fernando el mismo que ha sido revisado en su totalidad, analizado por la herramienta de verificación de similitud de contenido; por lo tanto cumple con los requisitos teóricos, científicos, técnicos, metodológicos y legales establecidos por la Universidad de las Fuerzas Armadas ESPE, razón por la cual me permito acreditar y autorizar para que lo sustente públicamente.

Sangolquí, 21 de agosto de 2021

Firma:

firmado electrónicamente por:
HERNAN JACOME

.....

Jácome Paneluisa, Hernán
C.C.: 1707493159

VICERRECTORADO DE INVESTIGACIÓN, INNOVACIÓN Y TRANSFERENCIA DE TECNOLOGÍA

CENTRO DE POSGRADOS

RESPONSABILIDAD DE AUTORÍA

Yo, ***Coyago Remache, Diego Fernando***, con cédula de ciudadanía n° 1712254307, declaro que el contenido, ideas y criterios del trabajo de titulación***: Desarrollo de un modelo de predicción de extinción de aves en Pichincha - Ecuador mediante técnicas de minería de datos*** es de mi autoría y responsabilidad, cumpliendo con los requisitos teóricos, científicos, técnicos, metodológicos y legales establecidos por la Universidad de las Fuerzas Armadas ESPE, respetando los derechos intelectuales de terceros y referenciando las citas bibliográficas.

Consecuentemente el contenido de la investigación mencionada es veraz.

Sangolquí, 21 de agosto de 2021

Firma

Firmado electrónicamente por:
DIEGO FERNANDO
COYAGO REMACHE

.....

Coyago Remache, Diego Fernando

C.C.: 1712254307

VICERRECTORADO DE INVESTIGACIÓN, INNOVACIÓN Y TRANSFERENCIA DE TECNOLOGÍA

CENTRO DE POSGRADOS

AUTORIZACIÓN DE PUBLICACIÓN

Yo, ***Coyago Remache, Diego Fernando,*** con cédula de ciudadanía n° 1712254307 *autorizo a la Universidad de las Fuerzas Armadas ESPE publicar el trabajo de titulación:* ***Desarrollo de un modelo de predicción de extinción de aves en Pichincha - Ecuador mediante técnicas de minería de datos*** *en el Repositorio Institucional, cuyo contenido, ideas y criterios son de mi responsabilidad.*

Sangolquí, 21 de agosto de 2021

Firma

Firmado electrónicamente por:
DIEGO FERNANDO
COYAGO REMACHE

.....

Coyago Remache, Diego Fernando

C.C.: 1712254307

Dedicatoria

Yo, dedico esta investigación a los entusiastas a favor de la biodiversidad, para que cuenten con una ayuda más en la prevención para la conservación de las aves que se encuentran en peligro o posible vía de extinción.

Agradecimiento

Yo, Agradezco a Dios por, sobre todo, también a mis padres, en especial a mi madre Margarita, que fue de ella donde nació la idea de realizar este aporte a la comunidad. Agradezco a mi familia por las contribuciones y aliento en la consecución de este trabajo y a mi esposa Mariuxi por al apoyo y paciencia otorgados.

ÍNDICE GENERAL

Certificación ***3***

Responsabilidad de Autoría ***4***

Autorización de Publicación ***5***

Dedicatoria ***6***

Agradecimiento ***7***

Resumen ***16***

Abstract ***17***

Capítulo I ***18***

Problema ***18***

Antecedentes **18**

Planteamiento del Problema **19**

Descripción del Problema 19

Objetivos de la Investigación **21**

Objetivo General 21

Objetivos Específicos 22

Justificación e Importancia del Proyecto **22**

Hipótesis **23**

Capítulo II ***25***

Marco Referencial ***25***

Marco Teórico **25**

Fundamentación de la variable independiente **25**

Big Data 25

Machine Learning 26

Minería de datos 26

Aprendizaje supervisado 27

Modelos predictivos 27

Fundamentación de la variable dependiente 27

Gobiernos 27

Grupos de interés 28

Investigadores u Ornitólogos 29

Conservación de especies de aves 29

Estado del Arte 29

Definición del Objetivo 29

Criterios de inclusión y exclusión 30

Criterios de inclusión 30

Criterios de exclusión 30

Definición de la estrategia de búsqueda 30

Grupo de control 31

Construcción de la cadena de búsqueda 32

Estudios de control 33

Conclusión 36

Capítulo III 37

Marco Metodológico 37

Introducción 37

Estudio de Caso 37

Fases de la metodología de Estudio de Caso 37

Capítulo IV 43

Desarrollo de la Investigación 43

Introducción 43

Planificación del proyecto de minería de datos 43

Fase 1. Compresión del negocio 45

Fase 2. Comprensión de los datos 45

Recolección de datos 46

Arquitectura de la solución 48

Exploración de datos 50

Calidad de datos 52

Fase 3. Preparación de los datos 53

Exploración de datos 53

Calidad de datos 55

Construcción de datos 58

Integración de datos 59

Fase 4. Modelamiento 61

Selección de variables 62

Selección de modelos 71

Diseño del modelo 72

Fase 5. Evaluación 75

Fase 6. Despliegue 80

Capa de datos 80

Capa Repositorio Stage 80

Capa Data Mart 81

Capa Json Data Mart 81

Capa Minería de Datos 81

Capa Visualización 82

Capítulo V 83

Discusión de Resultados 83

Evaluación de resultados obtenidos 83

Ejecución del modelo para Morphnarchus prínceps - Gavilán Barreteado 84

Ejecución del modelo para Spizaetus isidori - Águila Andina.....85

Ejecución del modelo para Lafresnaya lafresnayi Colibrí Terciopelo.....86

Ejecución del modelo para Cephalopterus penduliger Pájaro Paraguas Longuipéndulo 88

Ejecución del modelo para Vultur gryphus Cóndor Andino 89

Análisis de resultados..... 90

Análisis comparativo entre las muestras.....91

Informe de Resultados.....96

Capítulo VI 99

Conclusiones y Recomendaciones 99

Conclusiones 99

Recomendaciones 100

Futuros estudios de investigación 102

Referencias Bibliográficas 103

Anexos..... 106

ÍNDICE DE TABLAS

Tabla 1 Artículos del Grupo de Control 31

Tabla 2 Construcción de la Cadena de Búsqueda 32

Tabla 3 Metodología de investigación Estudio de Caso 41

Tabla 4 Contraste de fases entre metodologías Estudio de Caso y CRISP-DM 44

Tabla 5 Planificación del proyecto 44

Tabla 6 Variables candidatas para la investigación 59

Tabla 7 Puntaje dado por los métodos de selección 69

Tabla 8 Variables seleccionadas para los modelos 70

Tabla 9 Modelos de predicción 72

Tabla 10 Evaluación de los modelos de regresión candidatos 79

Tabla 11 Especies seleccionadas para ejecución de modelos 84

Tabla 12 Estadística comparativa entre el histórico y la predicción para el Gavilán Barreteado 85

Tabla 13 Estadística comparativa entre el histórico y la predicción para el Águila Andina 86

Tabla 14 Estadística comparativa entre el histórico y la predicción para el Colibrí Terciopelo 87

Tabla 15 Estadística comparativa entre el histórico y la predicción para el Pájaro Paraguas 88

Tabla 16 Estadística comparativa entre el histórico y la predicción para el Cóndor Andino 89

Tabla 17 Categorías del libro rojo de aves 90

Tabla 18 Nivel de riesgo de las especies seleccionadas 91

Tabla 19 Nivel de riesgo y frecuencia de avistamientos predichos 95

Tabla 20 Resumen de la predicción de alertas para las especies seleccionadas 97

ÍNDICE DE FIGURAS

Figura 1 Comparativa entre variables Independiente y Dependiente 25

Figura 2 Etapas de la metodología CRISP-DM 40

Figura 3 Factores críticos que afectan al hábitat de las especies 47

Figura 4 Arquitectura para la solución de la investigación 49

Figura 5 Diagrama Entidad Relación Multidimensional 52

Figura 6 Distribución física de las especies de aves en hábitat 54

Figura 7 Contabilización del número de observaciones por especie 55

Figura 8 Valores atípicos en el data set de especies 56

Figura 9 Especie Bubulcus ibis con valores atípicos 56

Figura 10 Matriz de correlación de variables 57

F**igura 11** Matriz correlacionada reducida 58

Figura 12 Variables candidatas para el entrenamiento de los modelos 60

Figura 13 Conjunto de datos o data set 61

Figura 14 Lista de modelos de regresión candidatos 62

Figura 15 Método Forward para selección de variables 63

Figura 16 Resultado método Forward 64

Figura 17 Método Backward para selección de variables 64

Figura 18 Resultado método Backward 65

Figura 19 Método Optimize Selection 66

Figura 20 Resultado método Optimize Selection 66

Figura 21 Método Brute Force para selección de variables 67

Figura 22 Resultado del método Brute Force 67

Figura 23 Método Optimize Selection Evolutionary para selección de variables 68

Figura 24 Resultado de método Optimize Selection Evolutionary 68

Figura 25 Variables seleccionadas por los modelos 70

Figura 26 Diseño del modelo de regresión Gradient Boosted Tree 73

Figura 27 Diseño del modelo de regresión Generalized Linear 74

Figura 28 Diseño del modelo de regresión Deep Learning 74

Figura 29 Comparación de ajuste-efectividad entre modelos 76

Figura 30 Perspectiva superior del ajuste entre modelos.....76

Figura 31 Resultado de pruebas del modelo GLM 77

Figura 32 Resultado de pruebas del modelo Deep Learning 77

Figura 33 Resultado de pruebas del modelo GBT.....78

Figura 34 Cálculo de error cuadrático medio y raíz del error cuadrático medio.....78

Figura 35 Curva de predicción para Gavilán Barreteado.....84

Figura 36 Curva de predicción para el Águila Andina.....85

Figura 37 Curva de predicción para el Colibrí Terciopelo.....87

Figura 38 Estadística comparativa entre el histórico y la predicción para el Pájaro Paraguas.....88

Figura 39 Estadística comparativa entre el histórico y la predicción para el Cóndor Andino 89

Figura 40 Comparativo de la evolución histórica y la predictiva anual de las especies seleccionadas a través de los años 92

Figura 41 Evolución histórica anual de avistamientos y predicciones para el Águila Andina 92

Figura 42 Evolución histórica anual de avistamientos y predicciones para el Cóndor Andino.....93

Figura 43 Evolución histórica anual de avistamientos y predicciones para el Pájaro Paraguas.....93

Figura 44 Evolución histórica anual de avistamientos y predicciones para el Colibrí Terciopelo 94

Figura 45 Evolución histórica anual de avistamientos y predicciones para el Gavilán Barreteado 94

Figura 46 Visualización de los resultados del modelo de predicción de alerta temprana 98

Resumen

Alrededor del mundo se han observado varias especies de aves extinguirse y pasar a formar parte de enciclopedias; muchas de las veces la humanidad como tal es reactiva a estos acontecimientos pudiendo hacer muy poco para conservar la fauna. Con el auge de los datos abiertos fue posible obtener múltiples fuentes heterogéneas que sometidas a un proceso de análisis con técnicas de minería de datos se descubrió información relacionada con la tendencia de avistamientos tempranos de especies de aves, para ser más proactivos que reactivos. El objetivo propuesto fue desarrollar un modelo de predicción de alerta temprana a la extinción de especies de aves basado en el cruce de varios factores críticos como cambio climático, intervención humana y contaminación para que organismos públicos y privados, opten por planes de conservación faunística. Para lograrlo *se* utilizó la metodología de Estudio de Caso, para comprender la extinción de especies en la provincia de Pichincha en Ecuador y mediante la metodología CRISP-DM, se diseñó un modelo predictivo de alerta temprana para prevenir la extinción de especies de aves. Los resultados obtenidos en la aplicación del modelo de predicción fueron efectivos en mostrar la tendencia de avistamientos de especies hasta finales de esta década, donde se evidencia la posibilidad de predecir una futura propensión a la extinción de especies. Estos resultados son insumos para grupos de interés que son parte activa de una comunidad de conservación faunística. Se concluye que los datos abiertos y los algoritmos predictivos, dan grandes posibilidades a los investigadores de realizar cruces de variables que antes no eran posibles descubriendo tendencias y resultados más tempranos que ayuden a la comunidad a ser proactivos frente a los acontecimientos irreversibles como es la perdida de cualquier especie, incluidas las más emblemáticas que llenan los escudos y símbolos de las naciones.

PALABRAS CLAVE:

- **MINERÍA DE DATOS**
- **BIG DATA**
- **MODELO DE PREDICCIÓN**
- **APRENDIZAJE AUTOMÁTICO**

Abstract

Around the world, several species of birds have been observed to become extinct and become part of encyclopedias; Many of the times humanity as such is reactive to these events, being able to do very little to conserve fauna. With the rise of open data, it was possible to obtain multiple heterogeneous sources that, subjected to an analysis process with data mining techniques, discovered information related to the trend of early sightings of bird species, to be more proactive than reactive. The proposed objective was to develop an early warning prediction model for the extinction of bird species based on the intersection of several critical factors such as climate change, human intervention and pollution for public and private organizations, opting for wildlife conservation plans. To achieve this, the Case Study methodology was used to understand the extinction of species in the province of Pichincha in Ecuador and through the CRISP-DM methodology, an early warning predictive model was designed to prevent the extinction of bird species. The results obtained in the application of the prediction model were effective in showing the trend of species sightings until the end of this decade, where the possibility of predicting a future propensity to species extinction is evident. These results are inputs for interest groups that are an active part of a wildlife conservation community. It is concluded that open data and predictive algorithms give researchers great possibilities to carry out crossovers of variables that were not possible before, discovering trends and earlier results that help the community to be proactive in the face of irreversible events such as loss. of any species, including the most emblematic ones that fill the shields and symbols of the nations.

KEYWORDS:

- **DATA MINING**
- **BIG DATA**
- **PREDICTION MODEL**
- **MACHINE LEARNING**

Capítulo I

Problema

Antecedentes

El equilibrio ecológico ha ido disminuyendo mientras que la intrusión del hombre va en aumento. Actualmente las personas que habitan el planeta ascienden a 7.53 mil millones (*Población total*, 2019) Las necesidades de la población mundial han aumentado a escalas colosales, donde la industria ha devastado grandes extensiones de bosque tropical, perdiendo en 2017 unas 15.8 millones de hectáreas (Hierro, 2018) con riqueza biológica para producir productos y saciar las necesidades básicas del hombre, principalmente la alimentación. El sector de la construcción ha devastado bosques para levantar casas, edificios, carreteras, puentes, etc. teniendo al 54% de la población mundial viviendo en zonas urbanas en 2014 (FAO, 2016). Sumando a éstas, y no menos invasiva la industria textil, minera, petrolera, energética, de carbono, etc. cada una contribuyendo a la expansión industrial y social, llevando a la reducción de ecosistemas biodiversos y al deterioro de la simbiosis natural.

La contaminación es el subproducto del desarrollo, la polución, agentes químicos, desastres de petróleo, materiales radiactivos, basura, pesticidas, etc. han contaminado la atmósfera, aguas, suelo, subsuelo y casi toda forma de vida. Según la Organización Mundial de la Salud (OMS), 1 de cada 8 muertes humanas son por contaminación (OMS, 2019). Existen pocos lugares en el mundo que aún conservan su estado natural, pero la mayoría de especies convive junto con la contaminación humana.

El cambio climático ha sido afectado por la gran cantidad de gases que suben a la atmósfera reteniendo el calor y son conocidos como gases de efecto invernadero que son responsables del 63% del calentamiento global, según lo menciona la comisión Europea (CE, 2019), y es producido principalmente por las fábricas, los automóviles, las

grandes granjas de ganado y la superpoblación. Una pequeña variación en la temperatura tiene el poder de desatar tormentas, precipitaciones, inundaciones, huracanes, desertificaciones, etc. que azotan a zonas de todo el mundo.

El desequilibrio de los ecosistemas ha incurrido en que varias de las especies animales y plantas que conviven en el planeta hayan desaparecido, migrado, o cambiado su comportamiento; estas condiciones producidas por la influencia el hombre, ponen en peligro de forma acelerada, la supervivencia de las especies en general.

Las aves son una de las especies que ha evolucionado desde la era de los dinosaurios, hacen más de 60 millones de años teniendo un antecesor común, el Archaeopteryx, que, según registros fósiles, tenía plumas. Las aves han reducido su tamaño desde entonces y han dominado casi todo el territorio del planeta, siendo los bosques tropicales el mayor albergue de aves, citando que en Ecuador hay 1690 especies de aves (Freile, J. F., 2018). La belleza, colorido y vocalización de las aves, han sido catalogados como íconos representativos desde milenios, como lo vemos hoy en día en emblemas, pinturas o grabados en piedra de culturas antiguas. Existen miles de variedades de especies de aves en el mundo, cuya existencia se siente amenazada por actividades humanas, el aumento de la población, la contaminación y cambio climático.

Planteamiento del Problema

Descripción del Problema

El desarrollo humano nos ha permitido un progreso acelerado y sin fronteras, dejando grandes beneficios, pero también varias consecuencias como la contaminación, que han apresurado el cambio climático normal del planeta y la degradación del ecosistema biodiverso de casi todo el planeta Ver Anexo 1 (Diagrama Causa-Efecto).

La agricultura junto con la tala de árboles, son las causas más importantes en la extinción de especies, dado que se han talado millones de hectáreas de bosques que

son el refugio de miles de especies de aves y donde se encuentra la mayor concentración de aves por Km^2. A estas causas se suma la caza indiscriminada de aves exóticas, la introducción de especies foráneas como gatos, perros, ratas, ratones, y demás animales domésticos, han reducido aún más el número de especies de aves. Con el impulso de la industrialización, la contaminación ha aumentado aceleradamente, la excesiva polución originada por autos, fábricas y las grandes ganaderías, están llenando la atmósfera con gases de efecto invernadero que provocan cambios en la temperatura global del planeta (BirdLife International, 2018).

Todas estas causas están reduciendo el número de especies de aves por todo el mundo, en la actualidad hay un 14% de especies de aves amenazadas en todo el planeta según un estudio de BirdLife (BirdLife International, 2018).

La provincia de Pichincha en Ecuador es el mayor centro administrativo, económico y financiero del país, por su localización estratégica y geopolítica. Pichincha es una zona altamente fértil, localizada en la zona interandina con una población mayor de 2 millones y medio de habitantes según el censo 2010 y 9.612 km2 de extensión, asentada en el valle de Quito rodeada de cordillera a una altura media de 2.816 metros sobre el nivel del mar con una herencia histórica de varios miles de años. Tiene un clima desde tropical a glacial, teniendo únicamente 2 estaciones, húmeda y seca con una temperatura media de entre 8 y 24 grados centígrados, características que son idóneas para una rica y variada biodiversidad de especies entre ellas las aves, dado que tiene extensas áreas de bosque tropical, hogar de cientos de especies de aves (*Gobierno de* Pichincha, 2019).

Todas estas características importantes de la provincia de Pichincha, han proliferado una fuerte migración, aumentando radicalmente la población, características que han llevado a una producción más industrial y menos agrícola; factores que inciden

en el desequilibrio del ecosistema. Las actividades humanas, la tala de los bosques, la introducción de especies foráneas, la industrialización, la polución y el cambio en el clima están deteriorando el ecosistema de una extensa zona biodiversa cuya variedad de especies de aves endémicas están reduciendo en número. El caso de Eriocnemis godini, una especie de ave endémica de la localidad de Perucho localizada al noroccidente de la provincia, se cree que estaría en extinción absoluta (Freile, J. F., 2018).

El estudio de las aves y el registro de las mismas ha tenido un gran esfuerzo en todo el mundo, ya que investigadores, entusiastas, grupos sociales y gubernamentales han trabajado en la conservación de especies de aves del mundo y cuentan con datos históricos de registros de especies de aves de todo el planeta recopilados desde hace más de 150 años. Estos registros históricos se tomaron de la observación y conteo del número de aves en su hábitat, por años y por diferentes ornitólogos y demás grupos interesados alrededor del mundo. Las variables obtenidas han indo en aumento hasta la posición geográfica. Gracias a estos datos y variables obtenidas, se puede alertar cuando una especie está en peligro, sin embargo en el Ecuador hay dependencia de fuentes e informes internacionales que alertan cuando una especie está amenazada, pero no se ajustan a la realidad y cambios en el entorno de la provincia de Pichincha, es por esto que deben considerarse el uso de nuevas variables críticas que permitan predecir la extinción de especies de aves como el aumento de la población, las actividades humanas, contaminación y cambio climático, factores propios de la provincia que aceleran el ritmo de supervivencia de las especies de aves endémicas.

Objetivos de la Investigación

Objetivo General

Desarrollar un modelo de predicción de extinción de especies de aves endémicas de la provincia Pichincha en Ecuador, usando técnicas de Machine Learning que permita

alertar tempranamente a organizaciones sociales y gubernamentales de conservación de aves afines a la ornitología cuando una especie se encuentre en etapas iniciales de extinción.

Objetivos Específicos

- Realizar una revisión de literatura inicial sobre extinción de especies de aves y las mejores técnicas de Maching Learning sobre modelos de predicción enfocadas a modelos de extinción.
- Realizar una recolección de fuentes abiertas sobre actividades humanas, población, contaminación y cambio climático y seleccionar las mejores herramientas para limpieza, análisis y Maching Learning.
- Desarrollar un modelo predictivo de extinción de especies de aves que permita obtener resultados tempranos para los grupos de interés, usando técnicas de Machine Learning.
- Validar el modelo predictivo en base a la lista roja de especies de aves en peligro de extinción del Ecuador emitido por el ministerio del ambiente.

Justificación e Importancia del Proyecto

Muchos investigadores enfocados al campo de la ornitología, han realizado importantes hallazgos en este campo, basado en la observación de aves, y con la ayuda de las nuevas tecnologías han profundizado y reforzado sus investigaciones. Ahora el nuevo reto es trabajar en un modelo de predicción de extinción de aves con variables ajustadas a la realidad local y con la ayuda de data proveniente de diferentes fuentes y varios colaboradores que exponen como datos abiertos, que son bases de datos publicadas en Internet por cada país para el público en general. La información se encuentra diseminada por toda la nube digital, lo cual permite a los investigadores hacer uso efectivo de ellas, sin embargo, se puede encontrar desde fuentes pequeñas hasta

fuentes de gran tamaño. Existen tecnologías para almacenar, procesar y analizar gran cantidad de información y varias de ellas tienen licencias de uso de software de forma libre, sin embargo, tienen un grado alto de complejidad. Hay datos abiertos de especies de animales con miles de datos de información, con una riqueza histórica muy importante, e invaluable para los investigadores. Varios de los sitios donde se publican datos abiertos, ofrecen como servicio estadísticas y consultas de datos en línea, manteniéndose en el marco de los datos abiertos que se ofrece. Sin embargo, para muchos investigadores es preciso realizar cruce de variables con otras fuentes de datos con información de otros contextos para enriquecer y potenciar las investigaciones. La importancia de una plataforma que pueda agrupar estos diferentes contextos de información en un silo de datos centralizado que permita descubrir patrones e incluso someter a analítica de datos para predicciones que son necesarias para poder actuar antes que sucedan acontecimientos como la reducción de miembros de una determinada especie que pueda desaparecer o estar en peligro de extinción.

Hipótesis

La implementación de un modelo de predicción permitirá alertar a organizaciones gubernamentales y sociales cuando una especie de ave se encuentra en peligro de extinción.

En este contexto se lo hará mediante un enfoque dependiente e independiente.

Variable dependiente, Identificación temprana de especies de aves en peligro de extinción.

Variable independiente, Modelo predictivo para evaluar las especies de aves que podrían estar en peligro de extinción.

Para comprobar esta hipótesis, será necesario en primera instancia contar con múltiples fuentes de datos que potencien la principal fuente que tiene información sobre

especies de aves. Después de someter la información relacionada a un tratamiento de calidad de datos, se estima elaborar un modelo de predicción en base al sometimiento de una batería de algoritmos de predicción utilizados en minería de datos, hasta encontrar el que entregue mayor confianza en los resultados y será el candidato para el modelo final que entregue una alta probabilidad de encontrar especies en peligro de extinción.

Capítulo II

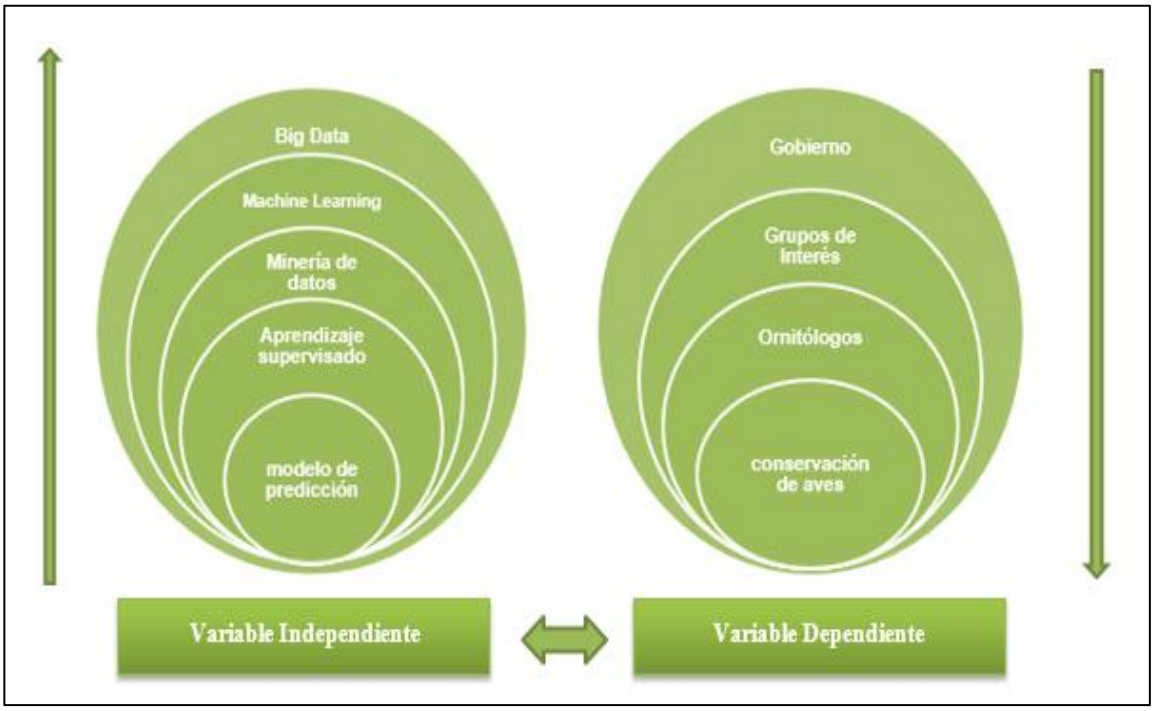
Marco Referencial

Marco Teórico

El estudio teórico realizado adopta dos perspectivas, explicadas o fundamentadas en las variables dependientes e independientes. Esta categorización permite separar la corriente teórica en estas dos caracterizaciones que justifican de qué manera teórica abordaremos el problema planteado y cómo llegaremos a concluir el estudio.

Figura 1

Comparativa entre variables Independiente y Dependiente



Nota: caracterización de la variable dependiente e independiente.

Fundamentación de la variable independiente

Big Data

La inmensidad de datos que circulan por la red, la enorme cantidad de datos generados en investigaciones científicas, la masiva cantidad de transacciones en las

organizaciones, y más ejemplos de este tipo, son hoy en día, la realidad del mundo globalizado y gracias al Internet, cada 2 años se duplica la información mundial. En este contexto, esta monstruosa cantidad de datos, solo puede ser tratada con una máquina igualmente grande. Por máquina me refiero a un conjunto de tecnologías, metodologías, procesos, técnicas y profesionales multidisciplinarios que en su conjunto pueden tratar toda esa masiva información que se genera día a día desde diferentes fuentes.

Cuando se habla de Big Data, los proyectos que encajan en este segmento, generalmente tienen asociadas la Velocidad, Variedad y Volumen (Mayer-Schonbrger & Cukier, 2013). Las cuales indican la rapidez con que crecen los datos, la variedad en el tipo de datos pudiendo ser estructurados o no estructurados y volumen, la acumulación masiva de datos, respectivamente.

Machine Learning

La aplicación de las técnicas de minería de datos, frecuentemente usa algoritmos de inteligencia artificial, los que pueden aprender de datos históricos suministrados pudiendo ser de dos corrientes, supervisados o no supervisados. La primera se enfoca en el aprendizaje de una variable objetivo o clase conocida previamente, y la segunda no cuenta con esta clase objetivo, sino que intenta aprender por sí mismo mediante el uso de técnicas de clasificación o agrupaciones. La combinación de varios de los algoritmos de Machine Learning potencian las analíticas, este proceso es conocido como Deep Learning, o aprendizaje profundo, el cual permite obtener mejores resultados que con la aplicación de algoritmos individuales (Gironés et al., 2017).

Minería de datos

Es un campo de muchas disciplinas como la estadística, el álgebra, el aprendizaje automático, etc., metodologías, procesos, modelización, técnicas matemáticas para descubrir en los datos relaciones, patrones, comportamiento, tendencias, optimizaciones,

etc. con la ayuda de algoritmos de la rama de la inteligencia artificial como clasificación, predicción, segmentación, series temporales, reglas de asociación (Gironés et al., 2017), etc. que dentro del contexto en el que son aplicados, se alinean con objetivos establecidos y es un pilar fundamental para la obtener conocimiento y toma de decisiones.

Una de las metodologías más utilizadas en el medio, es CRIPS-DM (Gironés et al., 2017) que permiten concluir con proyectos de minería de datos de principio a fin, siguiendo una serie de pasos o etapas bien diferenciadas como la comprensión del negocio y de los datos, preparación y modelado, evaluación y despliegue.

Aprendizaje supervisado

Se trata de algoritmos que aprenden en base a datos históricos, es decir, se tiene un juego de datos con ocurrencias pasadas que permiten al algoritmo aprender de entre un conjunto de datos con variables correctamente etiquetadas, para luego ser probados con un conjunto nuevo y tratar de predecir una variable objetivo en función de lo aprendido con el primer conjunto de entrenamiento.

Modelos predictivos

Hay modelos que tienen parámetros climáticos, como los más generales, sin embargo, nuevos estudios indican que el clima no es un factor de peso que amenaza a especies de aves, sino más bien la agricultura con un descomunal 74%, según estudio de BirdLife International (BirdLife International, 2018). Los factores críticos como la agricultura, contaminación atmosférica, minería, industria del papel, etc., podrían ser parte de estos modelos de predicción con resultados positivos y con menor sesgo en la aplicación de los mismos.

Fundamentación de la variable dependiente

Gobiernos

Los gobiernos como estrategia tienen varios temas de interés, entre los cuales la conservación de los ecosistemas a nivel mundial ha tenido una importancia vital, y es que varios factores han afectado el equilibrio terrestre, donde los humanos somos los principales involucrados en este desastre. La UNESCO por ejemplo ha declarado muchos sitios como patrimonio natural para evitar la destrucción y favorecer la conservación de ecosistemas megadiversos y sensibles. Ahora son conscientes de que las especies que se pierden no volverán jamás, incluso algunas de ellas son consideradas emblemáticas en varios países como el cóndor andino cuyo territorio sobrepasa fronteras como Colombia, Ecuador, Perú, Chile, Bolivia, Brasil y Argentina.

Grupos de interés

Muchos países y en especial Ecuador, aportan con datos a una base centralizada de observación de especies, entre ellas aves, con una riqueza histórica de más de una siglo, que ahora son datos abiertos donde miles de especies son catalogadas, es así que en un esfuerzo de décadas, Bird Life International, The Cornell Lab of Ornithology, Avibase, Unión Internacional para la conservación de la Naturaleza (IUCN) y asociaciones de cada país alrededor del mundo, han proporcionado a la comunidad una de las más grandes fuentes de datos abiertos de ornitología de todo el mundo, para que investigadores, gobiernos, organizaciones y todo aquel que requiera de datos, pueda hacer uso bajo licencia de los datos proporcionados.

Las instituciones gubernamentales del medio ambiente tienen mucho involucramiento en estudios del ecosistema donde su interés en la preservación de la flora, fauna y ambiente simbiótico muy sensible. A este grupo se suman organizaciones con alto sentido de conservación de ecosistema global que luchan por los derechos de quienes no tienen vos en este mundo, en este caso las aves. Este entrono que circunda a los humanos, quien ha devastado por milenios, pero ha sido en este último siglo que

ha socavado muy profundo, aunque existen proyectos de conservación de la naturaleza a favor del equilibrio ecológico del planeta.

Investigadores u Ornitólogos

Son investigadores en el campo de la ornitología dedicados al estudio de las aves, que trabajan para observar, estudiar y conservar especies de aves de todo el mundo, sin embargo, el esfuerzo de estos investigadores es investigar las causas externas de la amenaza o extinción de especies de aves, y en especial las que tienen gran riqueza histórica y patrimonial como el cóndor andino.

Conservación de especies de aves

La conservación de especies, en especial de aves, va más allá de la responsabilidad gubernamental, es una responsabilidad en defensa contra el desequilibrio y destrucción del ambiente; muchas organizaciones y voluntarios luchan por la conservación y prevención de la desaparición temprana de especies de aves, luchan contra de leyes contradictorias que favorecen la minería, industrias madereras, asentamientos ilegales, etc. las aves son más que un símbolo, son el equilibrio sostenido del ecosistema y vida de los bosques tropicales y su importancia en la biodiversidad del planeta.

Estado del Arte

Dado la problemática y su hipótesis, se plantea un estudio sistemático de literatura SMS, el cual permite conocer el estado actual de investigación sobre un problema específico. Para ello, se ha recurrido a varias fuentes digitales académicas como IEEExplore, ACM Digital Library, Google Schoolar, Cielo, Elsevier y Springer que recogen los trabajos de investigación realizados en el mundo, dando a la comunidad un acercamiento a las últimas investigaciones realizadas.

Definición del Objetivo

El estudio del estado del arte, tiene un enfoque revelador sobre la hipótesis y el problema planteado, cuya consecuencia es conocer cuanta investigación se ha realizado hasta el momento en el mundo.

Criterios de inclusión y exclusión

Las búsquedas sobre el tema investigado en las diferentes bases académicas digitales retornan una gran cantidad de artículos relacionados, que sería casi imposible estudiarlos todos, sin embargo, para reducir y acercarse al tema revisado es preciso adoptar ciertas reglas para minimizar y seleccionar los documentos de interés que son objeto de estudio.

Criterios de inclusión

- Artículos que sean publicados a partir del año 2012.
- Artículos que hablen sobre temas de extinción, conservación o peligro de supervivencia de especies en general debido al cambio climático, contaminación, superpoblación.
- Artículos que hablen sobre modelos predictivos enfocados a biología y relacionados con especies.
- Artículos que hablen sobre metodologías de sistemas de recomendación.
- Documentos, libros o revistas que tengan carácter de artículo científico.

Criterios de exclusión

- Artículos con metodologías distintas a sistemas de recomendación.
- Artículos con temas de entornos biológicos sin relación a extinción o peligro de especies.
- Artículos escritos en otros idiomas fuera del inglés y el español.

Definición de la estrategia de búsqueda

La estrategia de búsqueda utilizada en esta revisión inicial de literatura comprende de lo siguientes etapas realizadas:

- Revisión inicial para realizan búsquedas de literatura en las bases de datos académicas digitales que tienen relación con las preguntas de investigación.
- Validación cruzada de estudios que permitan descartar todos los artículos que caen dentro de los criterios de exclusión y que permiten el desarrollo de las siguientes fases; se añaden estos artículos al listado de control.
- Integración del grupo de control de todos los artículos que han superado los criterios de inclusión y exclusión forman el grupo de control, de los cuales se realizó un estudio de ítems como son el título, introducción y conclusiones cuyo análisis está en armonía con las preguntas de investigación.

Grupo de control

Todos los estudios que han cumplido con las características de la investigación forman el grupo de control seleccionado.

Tabla 1

Artículos del Grupo de Control

Grupo de control	Título	Palabras clave
EC1	Recognition of Endangered Pantanal Animal Species using Deep Learning Methods.	Animals, Computer architecture, Image color analysis, Electronic mail, Image segmentation, Clustering algorithms, and convolutional neural networks.
EC2	Performance of Support Vector Machines and Artificial Neural Network for Mapping Endangered Tree Species Using WorldView-2	Artificial neural network (ANN), Dukuduku, endangered tree species (ETS), indigenous forest, support vector machines (SVM).

Grupo de control	Título	Palabras clave
	Data in Dukuduku Forest, South Africa.	
EC3	Big data for forecasting the impacts of global change on plant communities.	Database, environmental maps, geospatial, informatics, remote sensing, species occurrences, uncertainty, vegetation.
EC4	Stacked species distribution models and macroecological models provide congruent projections of avian species richness under climate change.	Biodiversity, birds, climate change, conservation, macro ecology, North America, seasonality, species distribution models.
EC5	Updating Known Distribution Models for Forecasting Climate Change Impact on Endangered Species.	Distribution models, forecasting, climate change, endangered species.
EC6	Niche modeling of Endangered Philipine birds using GARP and MAXENT.	Niche, modelling, neural network, endangered birds, GARP, MAXENT.

Nota: Grupo de artículos candidatos para formar grupo de control

Construcción de la cadena de búsqueda

La cadena de búsqueda se construyó en función del número de repeticiones por palabra clave de los estudios de control. En la siguiente tabla se muestra la construcción de la cadena de búsqueda.

Tabla 2

Construcción de la Cadena de Búsqueda

Palabra clave	EC 1	EC 2	EC 3	EC 4	EC 5	EC 6	Repetición
Neural networks	X	X				X	3
Endangered		X			X	X	3

Palabra clave	EC 1	EC 2	EC 3	EC 4	EC 5	EC 6	Repetición
Support vector machines		X					1
Species		X	X	X	X		4
Biodiversity				X			1
Climate change				X	X		2
Distribution models				X	X		2
Birds				X		X	2

Nota: construcción de la cadena de búsqueda en función de las palabras clave de los artículos del grupo de control.

La cadena de búsqueda luego del conteo de palabras clave, se establece en un contexto de similitud con OR y un contexto diferente con AND de tal suerte que la cadena es la siguiente:

((((Endangered) OR (Species)) AND (Biodiversity OR “Climate change”)) OR (Birds)) AND ((Distribution AND models) AND (“Neural networks” OR “Support vector machines”))

((((Endangered) OR (Species)) AND (Biodiversity OR “Climate change”)) OR (Birds)) AND ((“Distribution models”) OR (“Neural networks” AND “Support vector machines”))

Estudios de control

A continuación, se listan los estudios de control seleccionados:

(De Arruda et al., 2018) Recognition of Endangered Pantanal Animal Species using Deep Learning Methods.

En este estudio, los autores identifican el Pantanal en Brasil como un ecosistema megadiverso 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 hábitat.

(Omer et al., 2015) Performance of Support Vector Machines and Artificial Neural Network for Mapping Endangered Tree Species Using WorldView-2 Data in Dukuduku Forest, South Africa.

Los autores del estudio 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.

(Franklin et al., 2017) Big data for forecasting the impacts of global change on plant communities.

Este estudio demuestra 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, esto demuestra 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.

(Distler et al., 2015) Stacked species distribution models and macroecological models provide congruent projections of avian species richness under climate change.

En este estudio abraca las poblaciones de aves de Norte América que habitan según la temporada invernal o veraniega. Lo que buscan los autores es determinar si el

cambio climático afecta la reproducción de aves en estas 2 temporadas donde las aves aumentan o disminuyen drásticamente. Se trabajó en base a encuestas de aves de los dos países y se sometió 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.

(Muñoz et al., 2013) Updating Known Distribution Models for Forecasting Climate Change Impact on Endangered Species.

Los autores de este estudio, se 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 un 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 la mejora no haciendo uno nuevo, sino mejorando los algoritmos de distribución de especies de aves con nuevos modelos de emisiones.

(Montenegro et al., 2017) Niche modeling of Endangered Philipine birds using GARP and MAXENT.

Los investigadores de este estudio, 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.

Conclusión

Los estudios demuestran que la aplicación de diferentes técnicas como los algoritmos de redes neuronales y modelos de distribución aplicados a especies pueden ser muy positivos, sin embargo, otro argumento válido es la fusión de varios modelos que permitan potenciar las predicciones de especies en peligro. Una batería de fuentes heterogéneas reducirá el sesgo referente a si solamente se toma en consideración parámetros climáticos para predecir comportamientos de especies; se pueden potenciar con fuentes biodiversas como vegetación, actividades humanas o emisiones, que permitan aumentar la confiabilidad del modelo. Contando con más variables de diferentes contextos que afectan a los ecosistemas, se puede encontrar relaciones que aumenten la capacidad efectiva de la predicción, cada variable debe aportar un valor determinante que puede llegar a formar un patrón de cambio en el ecosistema. Estas relaciones deben ser analizadas bajo varias técnicas de Machine Learning que permitan construir un modelo ajustado a la realidad de la zona de Pichincha. La influencia de las variables no son las mismas aplicadas a otros entornos donde las circunstancias son diferentes al igual que el deterioro de los ecosistemas.

Capítulo III

Marco Metodológico

Introducción

Estudio de Caso

Para cumplir con los objetivos de este proyecto y tomando en cuenta el enfoque cualitativo de la investigación, se propone la utilización de la metodología de investigación de estudio de caso, que permitirá llevar a cabo todas las actividades propuestas para contestar la pregunta de investigación planteada. Esta es una metodología que es utilizada cuando el investigador realiza un estudio particular (HELEN SIMONS, 2009), el estudio de un caso, y utilizado en ramas como las ciencias sociales y medicina, entre otras. Una definición de esta metodología la dio STAKE (1995) “El estudio de caso es el estudio de la particularidad y la complejidad de un caso, por el que se llega a comprender su actividad en circunstancias que son importantes”. A partir de esta clara definición, se tomará como metodología para la presente investigación.

Fases de la metodología de Estudio de Caso

Diseño del estudio de caso

En esta fase, se realizan diferentes tareas como la identificación de la problemática, se plantean los objetivos y se plantean las preguntas de investigación en relación al problema, hace un estudio de literatura sobre la extinción de especies, no necesariamente dirigido a las aves, sino ampliado a otras especies de animales o plantas que incurren en la misma problemática de extinción y modelos aplicados en función de variables independientes que ocasionan vulneración o reducción en el número de miembros de una determinada especie que lo pone en peligro o situación de riesgo. Para ello se realiza una investigación sobre los trabajos más recientes en temas de extinción

manejados con el uso de tecnologías de minería de datos y algoritmos de Machine Learning.

3.1.2.2. Marco teórico

Se revisan las referencias empíricas dentro del contexto de la investigación y las referencias teóricas que acompañan el estudio del caso; en esta fase se realiza un estudio de literatura inicial reciente que permita una comprensión global del fenómeno o problema de estudio.

3.1.2.3. Recolección de datos

En esta fase se definen las herramientas de recolección de datos y la determina el investigador en función de su criterio de cuáles serían las mejores técnicas de recolección adecuadas. Existen varias técnicas como la entrevista, observación, grupos focales, entre otros, los cuales permiten recabar datos del fenómeno investigado.

3.1.2.4. Análisis de la información

Luego de recoger la información, ésta debe ser analizada de manera que pueda ser estudiada a profundidad aplicando diferentes técnicas. Es importante el contraste empírico con el teórico para alienarse con la investigación científica realizada.

El desarrollo del presente proyecto de investigación requiere de un trabajo especializado en el área de la minería de datos, por lo que requiere de una metodología especial que permita la consecución de un proyecto de predicción basada en datos históricos y la culminación en sintonía con los objetivos planteados. Este desarrollo, seguirá la metodología CRISP-DM, que hoy es un referente en proyectos de minería de datos. Está compuesta de seis fases cuyo objetivo es guiar proyectos de minería de datos hacia la obtención de conocimiento. Sus fases son:

Comprensión del negocio: Esta fase tiene un enfoque de negocio y alineado con los objetivos, permitiendo obtener una visión global y establecer límites y alcances

del proyecto y centrarlo en ruta más objetiva. Se hace un estudio de la situación actual del negocio para conocer detalladamente temas más profundos del negocio. Con estos datos, se realiza la planificación del proyecto detallando las actividades y tareas a realizar.

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. En la captura se identifican todas las fuentes necesarias, en la exploración se realiza un estudio donde se observan datos, estructura, las condiciones de los datos, los errores o inconvenientes que tienen; finalmente se trabaja en la calidad de los datos, siendo este un proceso crucial en esta fase.

Preparación de los datos: Implica datos limpios, es decir el juego de datos que utilizarán los modelos. Estos datos son necesarios y suficientes para alcanzar los objetivos, y es por este motivo que deben alcanzar una fase de calidad y la adopción de criterios de inclusión y exclusión de datos que serán determinantes para alcanzar los objetivos. 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.

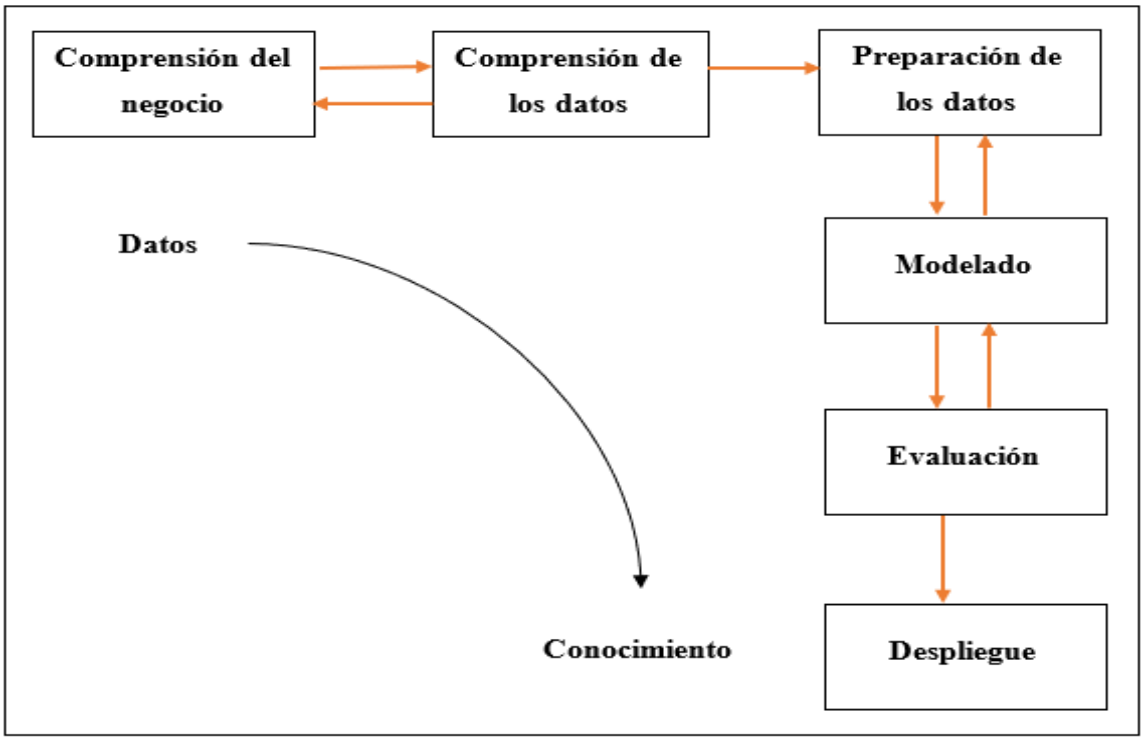
Evaluación: Una vez construido el modelo, se lo pone a prueba para descartar o mejorar o utilizar el modelo; este debe ser sometido a una batería de pruebas para

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.

Figura 2

Etapas de la metodología CRISP-DM



Nota: flujo de procesos en la metodología CRISP-DM, Fuente (Gironés et al., 2017)

3.1.2.5. Resultados

El estudio de caso determina la elaboración de informe en el que se documenta los hallazgos encontrados en la investigación; dentro de este informe, se presentará un estudio exploratorio, descriptivo e inferencial de lo investigado. Estos resultados deben definir el caso y su contexto; finalizando con las conclusiones y recomendaciones.

Los objetivos específicos son guiados mediante las fases de la metodología de investigación seleccionada Estudio de Caso; esta relación se puede observar según se muestra en la tabla 3.

Tabla 3

Metodología de investigación Estudio de Caso

Fase	Objetivo	Actividades relacionadas
1, 2	**OE1**: Realizar una revisión de literatura inicial sobre extinción de especies de aves y las mejores técnicas de Maching Learning sobre modelos de predicción enfocadas a modelos de extinción.	Revisión de literatura sobre extinción de especies, contaminación, población, intervención humana en el ecosistema mundial, cambio climático y modelos predictivos. Revisión de literatura sobre técnicas y algoritmos de Machine Learning que permitan determinar las características críticas de extinción. Revisión teórica de conceptos y fundamentos aplicados en la investigación de estudio de caso.
3	**OE2:** Realizar una recolección de fuentes abiertas sobre actividades humanas, población, contaminación y cambio climático y seleccionar las mejores herramientas para	Recolectar fuentes de datos abiertas de ornitología, población, intervención humana, contaminación y cambio climático relacionadas a la provincia de Pichincha-Ecuador.

Fase	Objetivo	Actividades relacionadas
	limpieza, análisis y Maching Learning.	
4	**OE3:** Desarrollar un modelo predictivo de extinción de especies de aves que permita obtener resultados tempranos para los grupos de interés, usando técnicas de Machine Learning.	Aplicar la metodología de minería de datos CRISP-DM en el desarrollo del modelo predictivo y entrenamiento del mismo mediante el aprendizaje supervisado.
5	**OE4:** Validar el modelo predictivo en base a la lista roja de especies de aves en peligro de extinción del Ecuador emitido por el ministerio del ambiente.	Evaluar el modelo predictivo en función de la Lista roja de especies en peligro de extinción. Elaborar un informe con los resultados encontrados y generar conclusiones y recomendaciones en base a la investigación realizada.

Nota: Establecimiento de actividades por objetivos

Capítulo IV

Desarrollo de la Investigación

Introducción

El estudio de investigación es el caso de la extinción de las especies de aves que habitan dentro de la provincia de Pichincha en Ecuador, las características específicas y los cambios externos que suceden en esta provincia, afectan el hábitat de todas las especies endémicas. Estos factores externos disminuyen la población de algunas especies que son susceptibles a cambios y no logran a largo plazo sobrevivir en un medio cambiante como es Pichicha que, por su ubicación y avance social, financiero, tecnológico, etc., es de rápido crecimiento. Ver capítulo 1 para mayor comprensión de la problemática.

El resultado de este estudio permitirá determinar un modelo que permita anticipar que especie de ave podría estar amenazada o en peligro de extinción, para tomar medidas tempranas en la prolongación de la especie en el tiempo.

El proyecto estará sujeto a la metodología de Estudio de Caso y a la metodología CRISP-DM, descritas en el capítulo 2 y 3. Con esta guía se realizarán todos los temas teóricos y el desarrollo tecnológico implícito correspondiente a este caso. Se realizará una comparativa entre las dos metodologías, aunque la metodología de caso es más general y la CRISP-DM es más específica, ambas pueden ser correspondidas entre sus diferentes fases en el marco del mismo proyecto.

Planificación del proyecto de minería de datos

La planificación está relacionada en la forma cómo se aborda el proyecto. Las metodologías guían al proyecto a su consecución. Sin embargo, las dos tienen fases similares que para el presente proyecto no amerita duplicar las fases. A continuación, se

realiza una correspondencia entre la similitud de las fases y las tareas que se realizan en una u otra fase sin duplicar la información.

Tabla 4

Contraste de fases entre metodologías Estudio de Caso y CRISP-DM

Caso de estudio	CRISP-DM	Referencia
Diseño del estudio de caso	Comprensión del negocio	Capítulo 1 Capítulo 2
Marco teórico		Capítulo 3
Recolección de datos	Comprensión de datos	Capítulo 4
Análisis de la información	Preparación de datos Modelado Evaluación	Capítulo 4
Resultados	Despliegue	Capítulo 5

Nota: Actividades relacionadas a cada fase de las metodologías utilizadas.

Con el contraste realizado entre metodologías, la planificación inicial del proyecto de minería juntamente con las actividades relacionadas en cada metodología, se observa en la tabla 5 a continuación.

Tabla 5

Planificación del proyecto

ACTIVIDAD	MES							
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Análisis SMS	x							
Análisis de metodologías	x							
Elaboración planificación de proyecto de minería		x						
Comprensión del negocio		x						

ACTIVIDAD	MES						
Recopilación de fuentes abiertas de especies de aves, cambio climático, población, contaminación, división política.	x						
Diseño y creación del repositorio de datos	x	x	x				
Comprensión de datos		x	x				
Preparación de datos			x	x			
Modelamiento y evaluación				x	x	x	
Despliegue					x	x	x
Conclusiones y Recomendaciones							x

Nota: Tiempo planificado para alcanzar los objetivos propuestos.

Fase 1. Compresión del negocio

El estudio de la comprensión del negocio realizado en el presente proyecto está descrito en el Capítulo 1 y 2 de este documento y 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. El otro factor determina los estudios similares en otros lugares con características diferenciadoras que definen de alguna manera la supervivencia o no de la especie.

Fase 2. Comprensión de los datos

En este estudio de caso 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 en para realizar un análisis de los mismos 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 datos para este proyecto tienen que pasar por un proceso antes de la fase de preparación de datos, cuyo propósito es tener una base de datos de calidad para el proceso de minería de datos. Los pasos principales son la recolección, exploración y limpieza de datos.

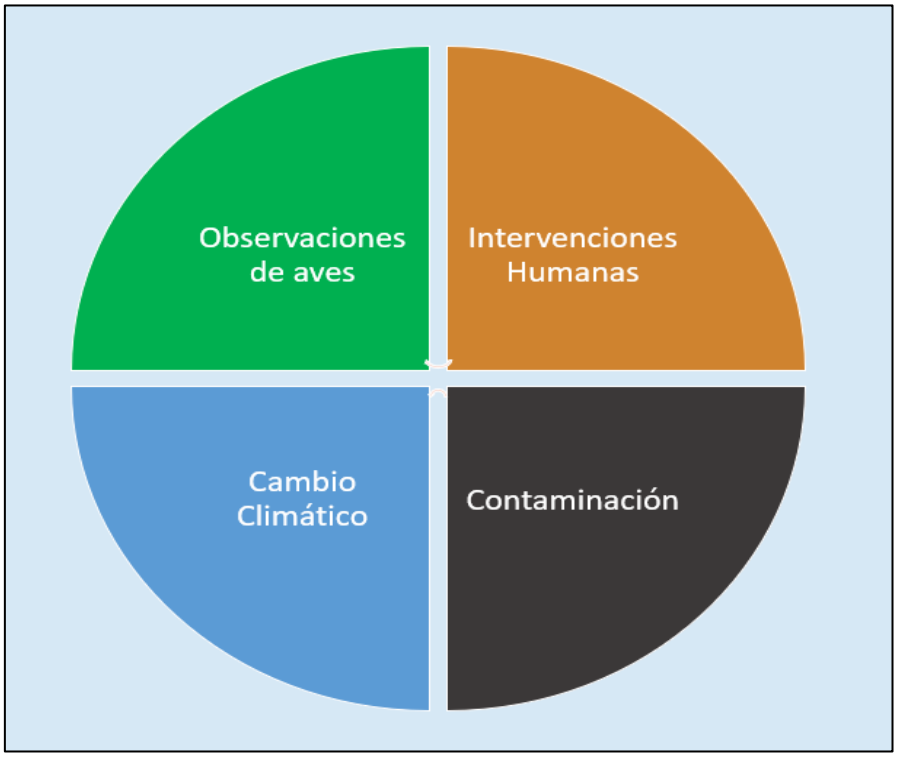
Recolección de datos

Para la recolección de datos, se hizo una búsqueda en repositorios abiertos, descargando la información necesaria y suficiente para proceder con el almacenamiento. Las bases de datos abiertas, son archivos en formato tipo texto, xls, csv y shape. En este estado la información no puede ser procesada, cruzada y entendida de manera global en el contexto del problema. Para entender el contenido de la información, se plantea una arquitectura que soporte la carga de datos y sea lo suficiente flexible para realizar el proceso de minería de datos.

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 sitio.

Figura 3

Factores críticos que afectan al hábitat de las especies



Nota: Factores que afectan el equilibrio de los ecosistemas.

Las bases de datos abiertas se buscaron en fuentes de organizamos gubernamentales o de organizaciones privadas. La base de datos de aves se obtuvo de Avibase[1], mediante el sitio de BirdLife. Los datos de contaminación y cambio climático se obtuvieron directamente del sitio web de satélites de la NASA Geovanni [2](NASA, 2020), y los datos de población y cartografía se obtuvieron del sitio web del INEC [3](INEC, 2019).

las descargas de los archivos fueron recopiladas en diferentes formatos. A continuación, se explica la forma en que la información de los archivos se encuentra contenida:

[1] Una base de datos mundial de especies de aves
[2] Servicio de descarga de parámetros geofísicos de la NASA (Agencia espacial norteamericana)
[3] Instituto Nacional de Estadísticas y Censos del Ecuador

- Tipo Texto, la información contenida está en formato texto separado por tabulaciones.
- Tipo Excel, la información está en formato xls en forma de filas y columnas y en formato csv, cuyos datos se encuentran separados por comas.
- Tipo Shape[4], la información alfanumérica y gráfica se guarda en este tipo de archivos que son leídos en programas especiales como Qgis[5] o ArcGis.

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:

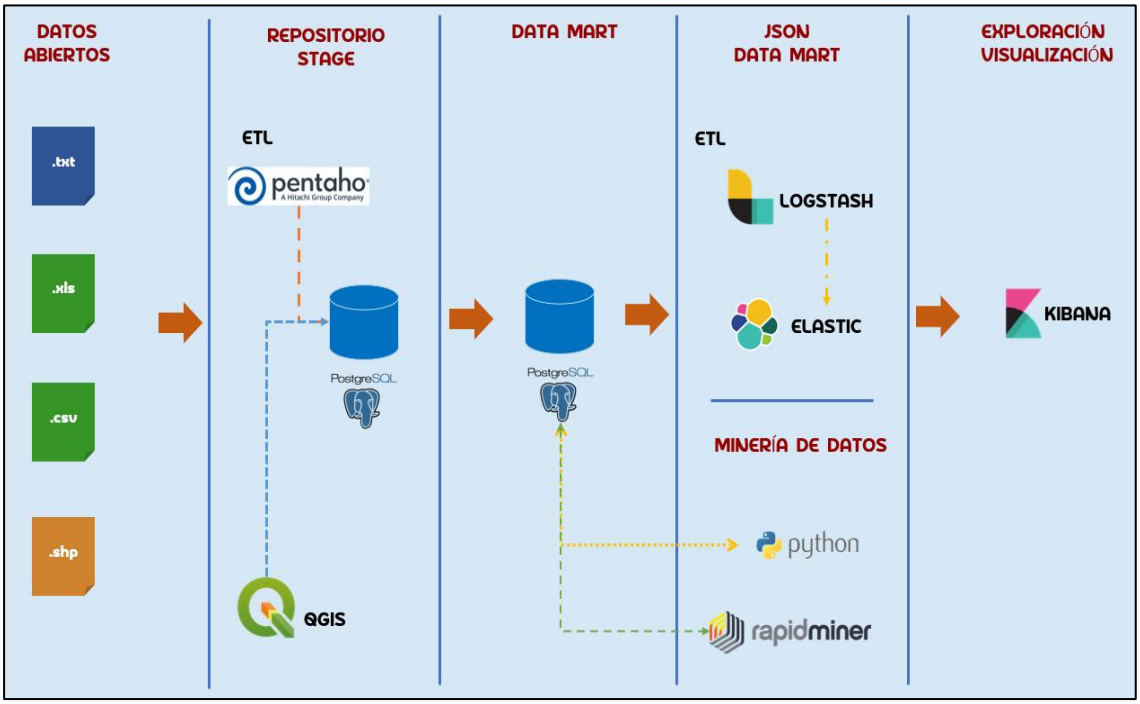
- Repositorio stage
- Data mart
- Json Data mart
- Minería de datos
- Exploración y visualización

La arquitectura de la solución está representada en la figura 4.

[4] Extensión o formato de un archivo con datos espaciales
[5] Herramientas de trabajo con archivos espaciales

Figura 4

Arquitectura para la solución de la investigación



Nota: La arquitectura es el diseño que seguirán los datos desde la recopilación hasta la presentación de resultados.

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

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. Dependiendo del tamaño, estructura y la forma de exploración de los datos, se cargan en fuentes y son enviados al repositorio Sql para su almacenamiento.

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.

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).

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.

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.

Exploración de datos

Cuando finalmente los datos inalterados son desembarcados en el repositorio Stage, se inicia un proceso de entendimiento de cada una de variables que llenan el repositorio de datos. Se analiza 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.

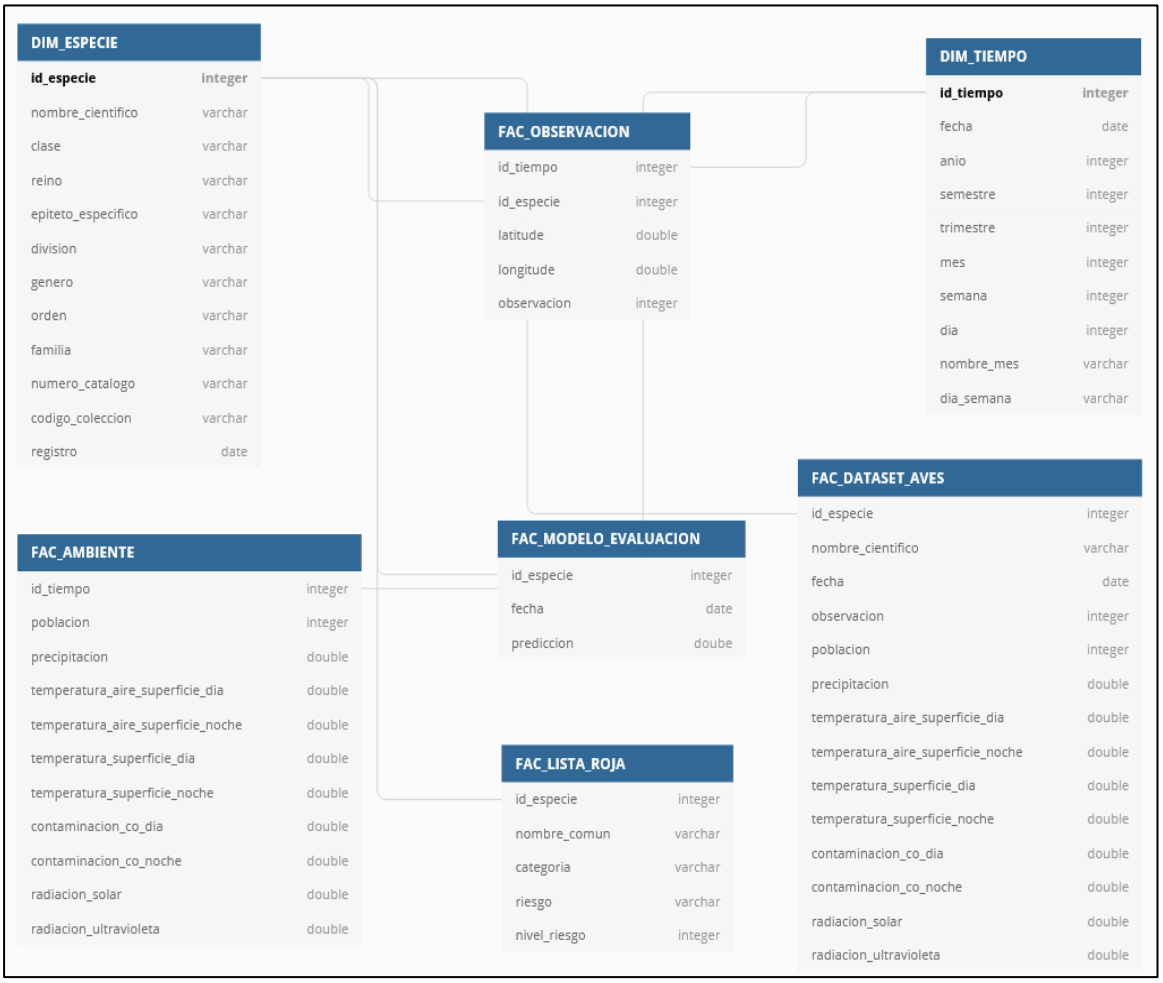
Hechos de data set aves: contiene la información de las variables candidatas y observaciones que son el insumo para el trabajo de minería de datos.

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.

Figura 5

Diagrama Entidad Relación Multidimensional



Nota: Representa el modelo adecuado para el análisis de los datos de la investigación.

Calidad de datos

El proceso de calidad de datos, es un insumo indispensable y obligatorio, no solo para un proceso de minería de datos, sino para cualquier proyecto de análisis de datos. En esta fase, se revisa que los datos estén completos, se eliminan las ambigüedades entre variables, se establecen reglas de estandarización de datos, se especifican los nuevos tipos de datos para las variables y se eliminan todo tipo de datos considerados basura que pueden causar ruido o distorsión en el proceso de minería.

Para este propósito, 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.

Fase 3. Preparación de los datos

En esta fase, los datos han pasado por un proceso de limpieza, estandarización y selección de variables importantes, cuyo objetivo es tener datos de calidad para el trabajo de investigación a realizar. En esta fase los datos se encuentran en los repositorios Data Mart y Json Data Mart, en este estado los datos pueden ser explorados visualmente y utilizar estadística descriptiva para conocer el estado de los mismos y sus relaciones.

El objetivo de esta fase es preparar los datos al punto de generar un 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.

Exploración de datos

Dado que el conjunto principal de datos son las observaciones de aves en campo; una visualización de cada especie pude ser mostrada en un mapa. El mapa corresponde a la provincia de Pichincha, claramente delimitada, con todas las observaciones de los investigadores realizadas durante años. Se diferencian algunas especies por el “orden” o clasificación de la especie.

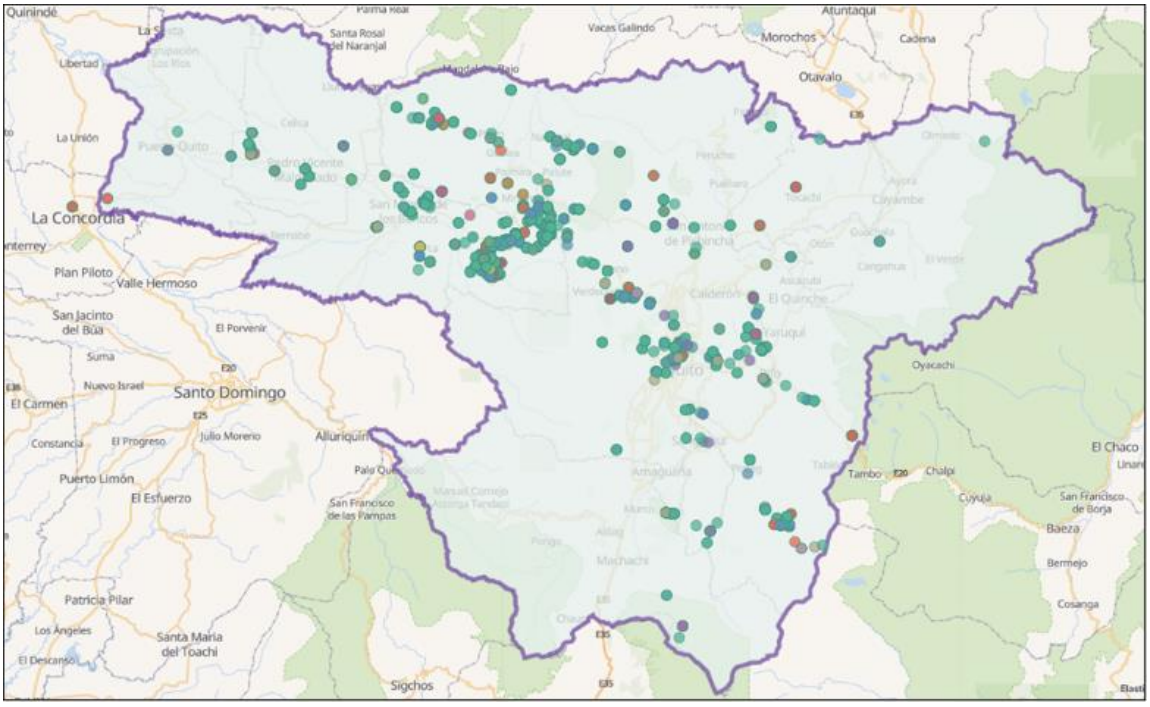
El conjunto de datos de avistamientos tiene variables espaciales que se pueden representar en un mapa. Los datos de observaciones tienen un sesgo que está limitado a la fecha en la que se tiene la observación y la especie que se está avistando, por lo que los avistamientos no son diarios, y está a discreción del observador. Este 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.

Los datos descargados de ambiente, tienen datos constantes, sin embargo, hay ciertos datos que serán tratados en una imputación más adelante, en el proceso de calidad de datos.

Figura 6

Distribución física de las especies de aves en hábitat



Nota: Muestra la dispersión de las tomas de los avistamientos de las especien en campo.

La cantidad de observaciones numéricas anuales, dan una clara idea sobre cuántas veces se toman muestras de las aves en campo y cuando no se ha tomado ninguna muestra, según el siguiente gráfico, donde se realiza un conteo por “orden” de la especie.

Figura 7

Contabilización del número de observaciones por especie

Passeriformes	2,706	2,571	11,365	19,564	21,799	22,925	20,845	26,149	29,901	28,292	32,044	28,077	43,433	38,418	98,302	123,780	160,718
Coprinuglriformes	948	775	2,578	4,329	3,467	8,180	6,829	7,545	10,040	12,313	16,049	12,203	20,304	16,791	48,332	48,825	62,783
Columniformes	203	426	761	1,071	956	1,567	1,303	1,457	2,257	2,774	2,823	2,354	2,925	3,914	15,852	16,221	14,649
Piciformes	139	175	762	1,214	1,410	1,494	1,680	1,982	1,922	1,657	2,161	1,759	2,623	2,557	6,284	8,083	11,414
Pelecaniformes	139	222	888	1,666	1,511	1,538	2,474	3,004	2,457	2,098	2,807	1,668	2,708	2,812	7,375	7,129	8,846
Cathartiformes	94	150	217	237	161	977	179	842	1,282	1,330	1,511	1,654	3,050	1,286	5,754	7,491	7,960
Accipitriformes	64	110	181	193	180	340	410	433	418	573	621	592	854	825	2,067	2,640	3,393
Charadriiformes	63	25	11	10	8	153	276	258	788	529	232	130	351	159	1,545	2,727	1,800
Trogoniformes	47	48	182	334	277	361	275	548	476	495	522	343	420	601	1,057	1,681	1,601
Cuculiformes	33	44	108	90	126	254	302	251	311	241	302	334	390	445	1,050	1,303	1,861
Galliformes	31	29	165	457	491	419	334	712	609	266	334	497	615	646	1,020	1,643	2,342
Falconiformes	25	30	36	71	49	147	80	158	362	414	433	197	405	289	1,023	1,750	1,512
Galbuliformes	14	3	9	14	9	44	88	50	39	25	64	62	56	75	167	167	293
Coraciiformes	6	13	33	47	53	80	97	78	98	100	141	180	348	278	842	930	1,269
Strigiformes	6	4	28	40	59	72	71	237	152	90	82	69	102	123	231	262	388
Pelecaniformes	5	13	75	125	126	358	90	627	545	1,148	7,842	592	2,754	1,405	9,664	7,362	6,983
Tinamiformes	4	3	58	29	45	33	115	64	26	15	39	15	53	52	87	120	191
Gruiformes	3	5	8	9	11	52	56	54	46	70	29	24	120	57	262	258	654
Anseriformes	2	26	13	0	4	102	80	149	159	90	17	61	312	102	2,400	3,370	1,613
Eurypygiformes		1	2	5	1	2	3	6	5	5	8	7	13	20	75	35	42
Suliformes		0	0		0	28	17	6	8	19	48	16	24	6	54	42	169
Podicipediformes						8			10	2	30	2	13	0	14	113	668
Phoenicopteriformes									1								

Nota: Exploración de datos de avistamientos de especies de aves.

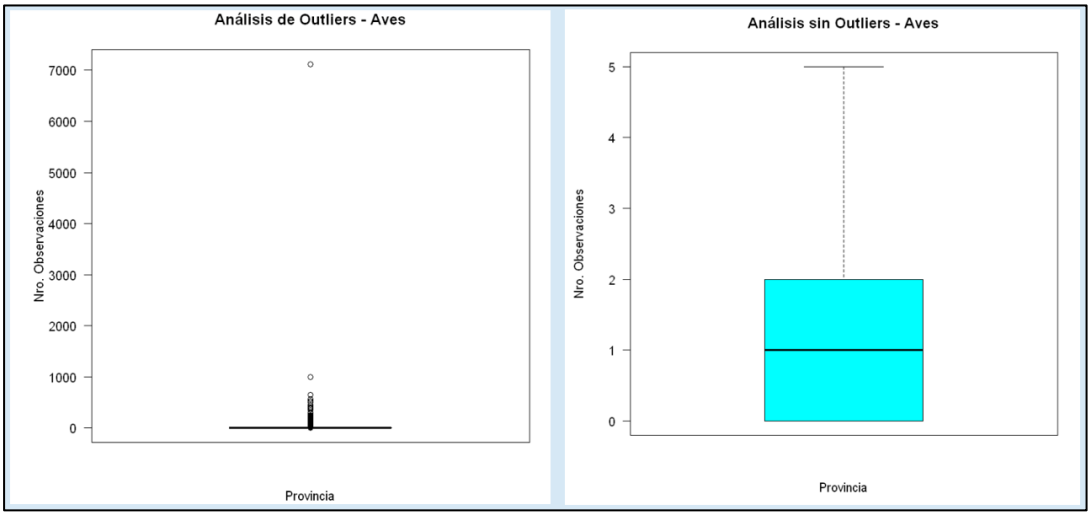
Calidad de datos

En el conjunto de datos preparado, se encontró valores nulos. Para este proceso, se excluyen todos los valores nulos de todas las variables. Previamente se trabajó las variables de forma independientemente, dado que son conjuntos de datos de tiempo. El tratamiento de imputación de datos que se dio a estos conjuntos de datos fue someterlo a un proceso de interpolación de datos orientado a series temporales, debido a la estructura que tienen los conjuntos de datos, es decir que todas las variables fueron recogidas con una temporalidad de un día y los valores son los promedios de los valores diarios, algunas con parciales de día y de noche.

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".

Figura 8

Valores atípicos en el data set de especies



Nota: Eliminación de valores atípicos del conjunto de datos.

Una de las especies con mayor número de valores atípicos, outlier o valor extremo, es la garza de la especie bubulcus, como se puede ver en la gráfica.

Figura 9

Especie Bubulcus ibis con valores atípicos

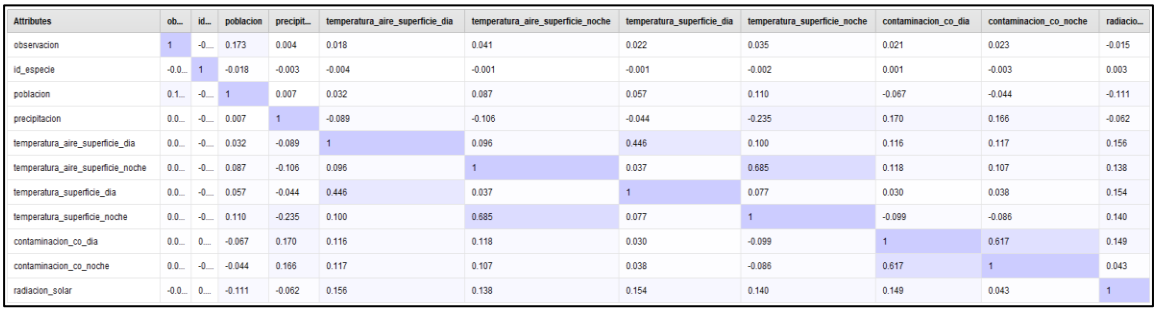
Nota: Ave con mayor número de valores extremos. Fuente (Wikipedia, 2021)

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.

Figura 10

Matriz de correlación de variables



Attributes	ob...	id...	poblacion	precipit...	temperatura_aire_superficie_dia	temperatura_aire_superficie_noche	temperatura_superficie_dia	temperatura_superficie_noche	contaminacion_co_dia	contaminacion_co_noche	radiacio...
observacion	1	-0....	0.173	0.004	0.018	0.041	0.022	0.035	0.021	0.023	-0.015
id_especie	-0.0...	1	-0.018	-0.003	-0.004	-0.001	-0.001	-0.002	0.001	-0.003	0.003
poblacion	0.1...	-0....	1	0.007	0.032	0.087	0.057	0.110	-0.067	-0.044	-0.111
precipitacion	0.0...	-0....	0.007	1	-0.089	-0.106	-0.044	-0.235	0.170	0.166	-0.062
temperatura_aire_superficie_dia	0.0...	-0....	0.032	-0.089	1	0.096	0.446	0.100	0.116	0.117	0.156
temperatura_aire_superficie_noche	0.0...	-0....	0.087	-0.106	0.096	1	0.037	0.685	0.118	0.107	0.138
temperatura_superficie_dia	0.0...	-0....	0.057	-0.044	0.446	0.037	1	0.077	0.030	0.038	0.154
temperatura_superficie_noche	0.0...	-0....	0.110	-0.235	0.100	0.685	0.077	1	-0.099	-0.086	0.140
contaminacion_co_dia	0.0...	0....	-0.067	0.170	0.116	0.118	0.030	-0.099	1	0.617	0.149
contaminacion_co_noche	0.0...	-0....	-0.044	0.166	0.117	0.107	0.038	-0.086	0.617	1	0.043
radiacion_solar	-0.0...	0....	-0.111	-0.062	0.156	0.138	0.154	0.140	0.149	0.043	1

Nota: Correlación entre las variables candidata del modelo.

Se encontró correlación entre ciertas variables. Las variables involucradas en la correlación son:

- temperatura_aire_superficie_dia
- temperatura_aire_superficie_noche
- temperatura_superficie_dia
- temperatura_superficie_noche
- contaminacion_co_dia
- contaminacion_co_noche

Donde la correlación se da entre las siguientes variables:

- temperatura_aire_superficie_dia y temperatura_superficie_dia
- temperatura_superficie_noche y temperatura_superficie_noche
- contaminacion_co_dia y contaminacion_co_noche

En este contexto, se excluyen las variables del proceso de modelado y son:

- temperatura_superficie_dia (se intercambiaría por temperatura de aire día)
- temperatura_superficie_noche (se intercambiaría por temperatura de aire noche)
- contaminacion_co_noche

Figura 11

Matriz correlacionada reducida

Attributes	observacion	poblacion	precipitacion	temperatura_aire_superficie_dia	temperatura_aire_superficie_noche	contaminacion_co_dia	radiacion_solar
observacion	1	0.173	0.004	0.018	0.041	0.021	-0.015
poblacion	0.173	1	0.007	0.032	0.087	-0.067	-0.111
precipitacion	0.004	0.007	1	-0.089	-0.106	0.170	-0.062
temperatura_aire_superficie_dia	0.018	0.032	-0.089	1	0.096	0.116	0.156
temperatura_aire_superficie_noche	0.041	0.087	-0.106	0.096	1	0.118	0.138
contaminacion_co_dia	0.021	-0.067	0.170	0.116	0.118	1	0.149
radiacion_solar	-0.015	-0.111	-0.062	0.156	0.138	0.149	1

Nota: se apartan las variables que tienen correlación alta, dado que hacen ruido en el conjunto de datos.

En la figura se observa que, al excluir las variables correlacionadas, la matriz ya no muestra un coeficiente elevado entre ninguna variable, lo que significa que todas las variables son independientes unas con otras, potenciando al modelo para el proceso de minería de datos.

Construcción de datos

En esta actividad, se construyen variables a partir de otras. Para este proceso, no fue necesario construir nuevas variables, dado que 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. Sin embargo, se ha estructurado el conjunto de datos según los factores críticos como intervención humana, cambio climático, contaminación, según se observa en la siguiente tabla.

Tabla 6

Variables candidatas para la investigación

Variable	Descripción	Perspectiva
Fecha	Fecha correspondiente al día de la observación de una especie particular de ave en la provincia	Común a todas
Observacion	Número individual de una especie en particular de ave observado en campo en una fecha determinada	Observación de Aves
Nombre_cientifico	Nombre científico de cada especie de ave	Observación de Aves
Poblacion	Proyección de población a una fecha determinada	Intervención Humana
Precipitacion	Precipitación tomada como promedio dentro del área de la provincia	Ambiente
temperatura_aire_superficie_dia	Temperatura del aire cerca de la superficie tomada como promedio durante el día	Ambiente
temperatura_aire_superficie_noche	Temperatura del aire cerca de la superficie tomada como promedio durante la noche	Ambiente
contaminacion_co_dia	Contaminación de monóxido de carbono, tomado como promedio durante el día.	Contaminación
radiacion_solar	Radiación solar sobre la superficie de la tierra, tomado como promedio diario.	Ambiente
radiacion_ultravioleta	Radiación ultravioleta tomado como promedio diario	Ambiente

Nota: Descripción de las variables candidatas que formarán parte del modelo de predicción.

Integración de datos

La creación del conjunto de datos como insumo para el modelado, tuvo varias entradas o bases de datos, que giró en torno a 4 aspectos, en este sentido, se trabajó la calidad de cada conjunto de datos de forma independientemente, luego se cruzó cada una de ellas en función de la variable pivote fecha, la cual es común en todos los

conjuntos de datos seleccionados. 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. Este data set o conjunto de datos, es el seleccionado para trabajar el modelamiento de datos. Se pude observar qua se encuentran las variables correlacionadas, sin embargo, son omitidas en esta fase de modelamiento, dado que son variables que no aportan valor al modelo.

Figura 12

Variables candidatas para el entrenamiento de los modelos

Nota: Tabla que almacena las variables que conforman el conjunto de datos.

Figura 13

Conjunto de datos o data set

Row No.	id_especie	observacion	poblacion	temperatura_air...	temperatura_...	contaminaci...	radiacion_solar	temperatura_superfi...	temperatura...	contaminacion_co_...	radiacion_ultraviol...	precipitacion
153262	1136	38	2490909.685	19.900	14.437	119.895	109.828	18.626	9.900	124.998	11.639	4.126
153263	1136	45	2495254.055	19.758	14.384	124.758	110.000	19.020	8.830	113.222	11.398	4.997
153264	1136	57	2499598.425	19.409	12.939	114.882	95.203	18.171	7.232	115.786	10.726	3.596
153265	1136	55	2504080.443	19.583	14.171	119.867	146.661	16.643	9.637	123.249	10.672	8.800
153266	1136	53	2508472.912	19.860	14.846	122.637	153.499	19.906	11.145	120.918	11.635	9.485
153267	1136	50	2512890.511	20.504	14.096	138.310	161.982	22.086	9.172	124.646	12.280	9.749
153268	1136	44	2517346.852	20.358	15.379	122.603	167.111	22.900	11.292	117.667	12.577	7.573
153269	1136	57	2521817.852	19.482	13.994	119.994	153.001	20.383	8.189	110.934	10.578	5.319
153270	1136	53	2526357.093	19.381	14.547	102.390	134.901	19.973	11.273	99.324	9.186	3.896
153271	1136	62	2530828.261	18.944	13.534	98.641	129.539	18.605	10.611	107.684	8.959	2.401
153272	1136	5	2535304.148	19.540	12.905	105.938	153.408	20.514	9.396	107.302	9.979	2.805
153273	1136	0	2539775.148	19.785	13.875	120.995	145.451	20.964	10.680	114.258	10.833	3.101
153274	1136	0	2544246.148	19.081	12.911	120.305	114.155	18.840	9.599	112.081	11.257	4.771
153275	1136	7	2548785.388	20.198	14.019	115.850	93.238	20.599	10.528	117.336	11.243	2.654
153276	1136	3	2553188.148	19.223	14.307	112.690	87.667	18.546	10.072	127.361	9.978	2.671
153277	1136	4	2557814.090	19.004	13.601	125.464	158.325	17.497	7.882	129.291	10.910	8.855
153278	1136	2	2562289.038	19.957	15.123	124.496	170.839	18.591	10.371	129.548	11.509	6.070
153279	1136	2	2566763.986	20.836	14.790	123.467	139.788	21.433	9.256	117.925	11.730	6.994
153280	1136	0	2571390.627	20.222	15.077	134.660	152.917	20.961	11.020	115.831	12.106	4.702
153281	1136	6	2575941.422	20.691	14.295	109.184	146.382	21.139	10.960	108.497	11.330	3.989
153282	1136	0	2580643.910	19.690	14.713	101.774	133.395	19.294	11.603	98.663	9.929	2.095
153283	1136	2	2585270.551	19.906	14.630	105.376	138.015	22.263	12.011	99.671	10.606	1.505
153284	1136	0	2589973.038	20.247	14.304	106.891	159.930	20.347	10.808	108.715	10.449	1.664
153285	1136	3	2594665.064	20.934	14.633	114.810	158.721	23.766	12.157	120.511	12.022	1.169

Nota: data set para entrenamiento de los modelos de Machine Learning.

Fase 4. Modelamiento

Esta sección es compleja, debido a que se toman varios aspectos relacionados con la cantidad de variables, los tipos de datos, los valores de los datos y los diferentes modelos de selección, los cuales son la base de las predicciones. En este apartado, seleccionar las variables y el modelo es muy importante para responder las preguntas de hipótesis y cumplir con el objetivo propuesto.

Una de las herramientas seleccionadas para el modelamiento, debido a la facilidad de uso y variedad de modelos de Machine Learning es “Rapidminer Educational Edition[6]” en su versión Académica. Una de las funcionalidades de gran ayuda es el “auto model”, el cual analiza el data set seleccionado y sugiere cuales modelos son los mejor puntuados para predecir en función de las variables y datos dados. En este apartado, se

[6] Edición de Rapidminer, para investigación académica netamente

utilizó esta funcionalidad para conocer rápidamente que modelos son los más mocionados para trabajar con el data set propuesto y afinar los modelos.

Después de ejecutar el “auto model”, se obtuvo que los modelos que mejor se ajustan o adaptan para trabajar en predicción con el data set proporcionado son Deep Learning, Gradient Boosted Trees y Random Forest, Generalized Linear, Decision Tree. Al someter a pruebas, se definirá la lista final, sin embargo, se muestra los primeros candidatos:

Figura 14

Lista de modelos de regresión candidatos

Nota: Modelos candidatos para el aprendizaje.

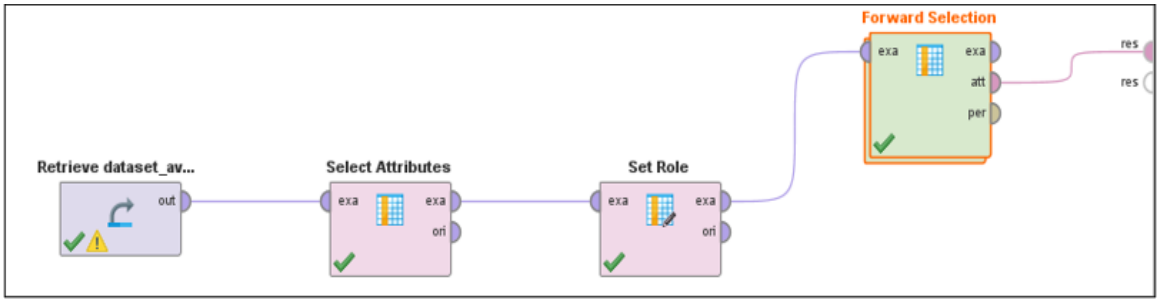
Selección de variables

Previo al diseño del modelo, el data set o conjunto de datos, contiene un número de variables que, por lo general, no siempre es conveniente para la construcción del modelo, debido a que ciertas variables pueden generar ruido en el modelo y, por lo tanto, no es óptimo y puede bajar el rendimiento o precisión del mismo. En este apartado se presentan algunas de las técnicas o métodos de selección de variables para optimizar el modelo y elevar el desempeño del modelo de predicción. La variable fecha se separó en año y mes por facilidad en los procesos. Cada método asigna un puntaje o peso a cada variable, si considera que será importante como variable independiente del modelo.

Método Forward: es un método de selección de variables que “comienza con una selección vacía de atributos y, en cada iteración, agrega cada atributo no utilizado del conjunto de datos. Para cada atributo agregado, el rendimiento se estima utilizando los operadores internos” (RapimdMiner, 2021). El resultado de la ejecución del modelo, indica cuales son las variables adecuadas y cuáles son inadecuadas. Se retornan valores 1 y 0 respectivamente.

Figura 15

Método Forward para selección de variables



Nota: método que compara la eficacia de las variables que tendrían en los modelos de predicción.

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 variables solo provocarán ruido o impactarán negativamente en el desempeño del modelo de predicción.

Figura 16

Resultado método Forward

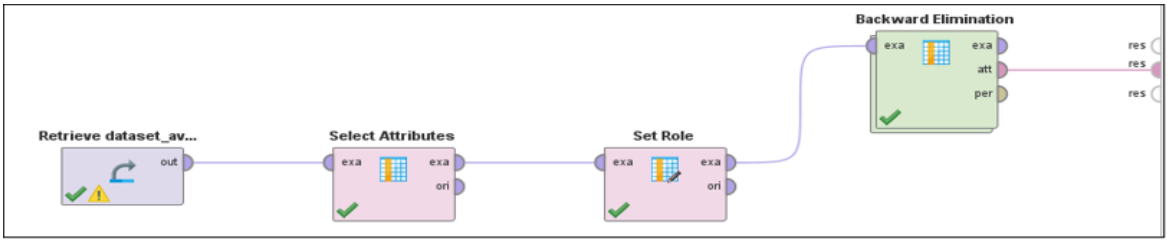
attribute	weight
poblacion	1
temperatura_aire_superficie_dia	0
temperatura_aire_superficie_noche	0
contaminacion_co_dia	1
radiacion_solar	1
temperatura_superficie_dia	1
temperatura_superficie_noche	0
contaminacion_co_noche	0
radiacion_ultravioleta	1
precipitacion	0
year	0
month	0

Nota: Con el método forward, más de la mitad de las variables son descartadas.

Método Backward: este método hace un proceso similar al anterior, pero hacia atrás, es decir “comienza con el conjunto completo de atributos y, en cada iteración, elimina cada atributo restante del conjunto de datos. Para cada atributo eliminado, el rendimiento se estima utilizando los operadores internos, como la validación cruzada. 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” (RapimdMiner, 2021). Las variables con alto rendimiento tomaran el peso de 1, el resto con peso 0.

Figura 17

Método Backward para selección de variables



Nota: Este método backward, solamente descartó una variable del conjunto.

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.

Figura 18

Resultado método Backward

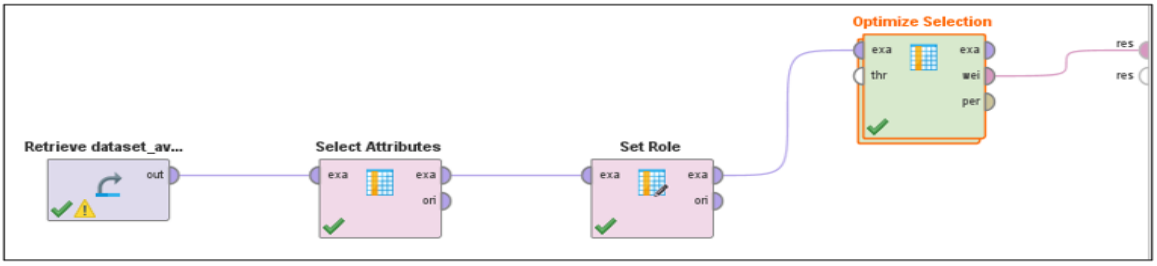
attribute	weight
poblacion	1
temperatura_aire_superficie_dia	1
temperatura_aire_superficie_noche	0
contaminacion_co_dia	1
radiacion_solar	1
temperatura_superficie_dia	1
temperatura_superficie_noche	1
contaminacion_co_noche	1
radiacion_ultravioleta	1
precipitacion	1
year	1
month	1

Nota: Con el método backward, solo una variable es separada del conjunto de datos.

Método Optimize Selection: este método optimiza la selección de las variables, tomando las bondades de los métodos mencionados anteriormente, dado que “selecciona los atributos más relevantes del conjunto de datos. Para la selección de características se utilizan dos algoritmos deterministas de selección de características codiciosos, forward y backward” (RapimdMiner, 2021). Las variables más eficientes obtendrán un peso de 1, el resto un peso de 0.

Figura 19

Método Optimize Selection



Nota: este método es más riguroso con la selección de las variables candidatas.

Este método al combinar 2 métodos de selección, descartó mas de la mitad de las variables, es decir, no aportarían valor al modelo, y le restarían precisión.

Figura 20

Resultado método Optimize Selection

attribute	weight
poblacion	1
temperatura_aire_superficie_dia	0
temperatura_aire_superficie_noche	0
contaminacion_co_dia	1
radiacion_solar	1
temperatura_superficie_dia	1
temperatura_superficie_noche	0
contaminacion_co_noche	0
radiacion_ultravioleta	1
precipitacion	0
year	0
month	0

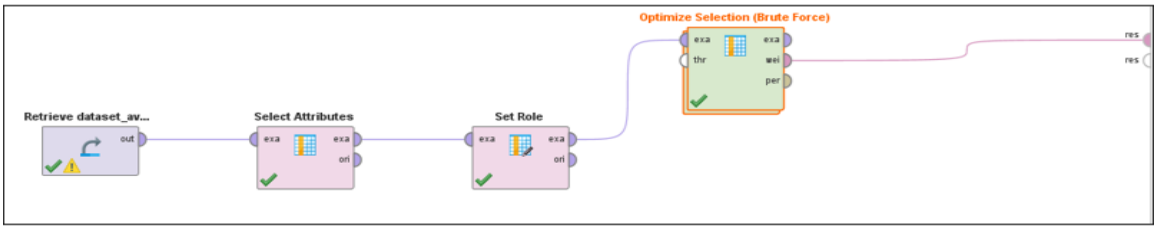
Nota: descartó más variables que el método forward.

Método Optimize Selection (Brute Force): este método alta demanda de procesamiento y memoria, permite encontrar el o los mejores candidatos, y se basas en “seleccionar el mejor conjunto de atributos probando todas las combinaciones posibles de selecciones de atributos. Devuelve el set que contiene el subconjunto de atributos que produjeron el mejor rendimiento. Como este operador 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 entren a comparar.

Figura 21

Método Brute Force para selección de variables



Nota: método que combina todas las posibles combinaciones entre variables.

Este método tiene la misma respuesta en cuanto a que variables tienen buenos rendimientos, respecto al modelo backward. Solo que la variable ineficaz es diferente.

Figura 22

Resultado del método Brute Force

attribute	weight
fecha	1
nombre_cientifico	1
id_especie	1
poblacion	1
precipitacion	0
temperatura_aire_superficie_dia	1
temperatura_aire_superficie_noche	1
contaminacion_co_dia	1
radiacion_solar	1

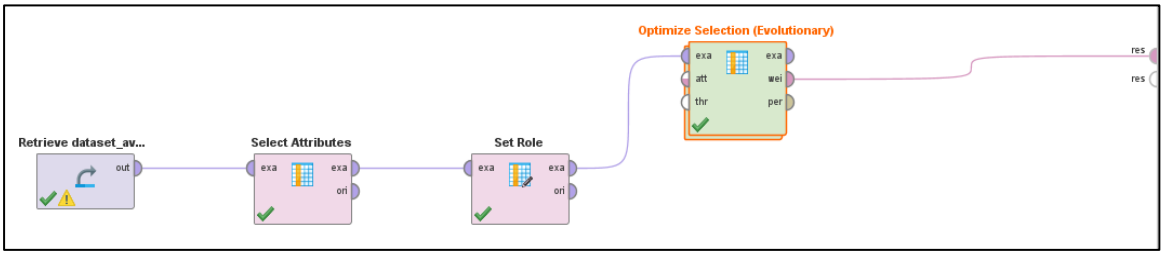
Nota: Este método, encontró que solo la precipitación, tiene un bajo rendimiento.

Método Optimize Selection (Evolutionary): este método 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. Los algoritmos genéticos pertenecen a la clase más amplia de algoritmos evolutivos (EA), que generan soluciones

a problemas de optimización utilizando técnicas inspiradas en la evolución natural, como herencia, mutación, selección y cruce” (RapimdMiner, 2021).

Figura 23

Método Optimize Selection Evolutionary para selección de variables



Nota: Utiliza algoritmos genéticos para seleccionar las variables condidatas.

El método evolucionario, toma mas de la mitad de variables como efectivas para el modelo de predicción. Varias de las variables desechadas por este método, fueron consideradas por los modelos anteriores.

Figura 24

Resultado de método Optimize Selection Evolutionary

attribute	weight
poblacion	0
temperatura_aire_superficie_dia	0
temperatura_aire_superficie_noche	0
contaminacion_co_dia	0
radiacion_solar	1
temperatura_superficie_dia	1
temperatura_superficie_noche	1
contaminacion_co_noche	1
radiacion_ultravioleta	1
precipitacion	1
year	1
month	0

Nota, Este modelo encontró mas variables útiles que ineficaces.

Luego de una batería de algoritmos para establecer los atributos eficientes para ser partícipes activos de los modelos de regresión, se presentan las variables potenciales

que pueden participar en la construcción de los modelos de datos. Una vez definidos los modelos y/o afinados, estas variables, pasarán a formar parte de los modelos finales.

Tabla 7

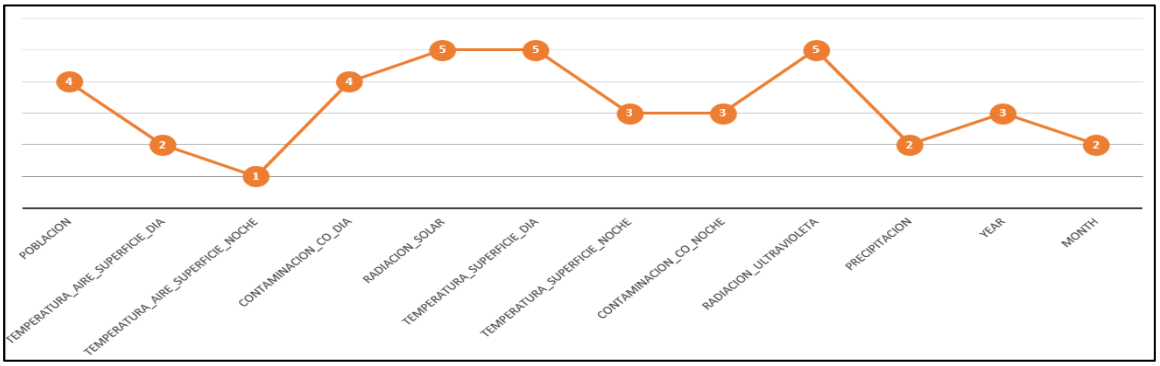
Puntaje dado por los métodos de selección

Variable	Optimize	Backward	Forward	Brute Force	Evolutionary	Peso
poblacion	1	1	1	1	0	4
temperatura_aire_superficie_dia	0	1	0	1	0	2
temperatura_aire_superficie_noche	0	0	0	1	0	1
contaminacion_co_dia	1	1	1	1	0	4
radiacion_solar	1	1	1	1	1	5
temperatura_superficie_dia	1	1	1	1	1	5
temperatura_superficie_noche	0	1	0	1	1	3
contaminacion_co_noche	0	1	0	1	1	3
radiacion_ultravioleta	1	1	1	1	1	5
precipitacion	0	1	0	0	1	2
year	0	1	0	1	1	3
month	0	1	0	1	0	2

Nota: Suma del peso total de las variables candidatas de selección.

Figura 25

Variables seleccionadas por los modelos



Nota: selección de las variables candidatas.

La selección de las variables está dada por la suma de las puntuaciones dadas por cada método de selección, en este caso, se tomarán las variables cuyos datos suman 5 puntos, expuestas en la tabla 8, es decir, aceptadas por cada método. Las variables con puntaje 4, son condicionadas, participarían en los modelos, siempre que no alteren el comportamiento de los modelos. El resultado de la ejecución de los métodos de la selección se presenta en la siguiente tabla:

Tabla 8

Variables seleccionadas para los modelos

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

Nota: Selección de las mejores variables para construir los modelos de regresión.

El resto de variables fueron desestimadas de la selección por tener poco peso dentro del data set. Por lo tanto, no serán incluidas en la construcción de los modelos de regresión.

Selección de modelos

Para la elaboración de modelos, es necesario que la técnica adecuada cumpla con los objetivos propuestos y los datos recopilados que conforman el contexto general del problema, en esta investigación el objetivo es tener una predicción de la supervivencia de las aves que habitan la provincia de Pichincha en Ecuador, mediante el cruce con otras variables que hacen de su habitad el hogar de varias especies, en alrededor de 1600 en todo el Ecuador, en Pichincha rondan alrededor de 776 especies.

Los modelos de predicción requieren que los datos con los que se trabajará, deben tener un mínimo de registros, para que los modelos puedan entrenarse y testearse. Para esta tesis, se elegirán los registros cuyas especies tengas más de 90 registros sin considerar los registros de test o prueba. Generalmente se podría considerar data a partir de las 50 observaciones por fenómeno, sin embargo, los valores de la mediana cuyo valor sea menor de 3 no se considerarán, excepto para realizar un comparativo y comprobar el comportamiento de los modelos, Se seleccionarán como tope mínimo 90. Hay especies que estando en la lista roja de aves en peligro de extinción, no cuentan con el suficiente número de observaciones. Para este análisis no se considerarán estas especies, sin embargo, si el número de observaciones supera los 90 registros en el data set a lo largo de alrededor de 12 años de registros, formarán parte de los modelos.

Los modelos se utilizan como técnicas de minería de datos. En este análisis, han sido previamente testeados con la herramienta Rapidminer 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 escogerán por cada tipo de relación, dado que se adaptan en similares condiciones y se escogerá uno de cada categoría.

Tabla 9

Modelos de predicción

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

Nota: Modelos de predicción candidatos.

Diseño del modelo

El diseño de los modelos y las variables asociadas, gráficamente seguirán el mismo patrón, aplicado para cada modelo de forma individual.

Para el diseño de los modelos se utilizará la herramienta Rapidminer, el cual facilita la construcción gracias a los procesos de componentes gráficos. 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. Este data set, pasó por un proceso de calidad de datos, antes de llegar al Data mart; descrito en la arquitectura de la solución. El siguiente proceso realiza una mezcla de los datos, con el fin de barajar muy bien los datos. Luego, con un 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 “**observacion**”. El siguiente paso es el uso del modelo seleccionado, cuya salida será directamente al modelo dimensional.

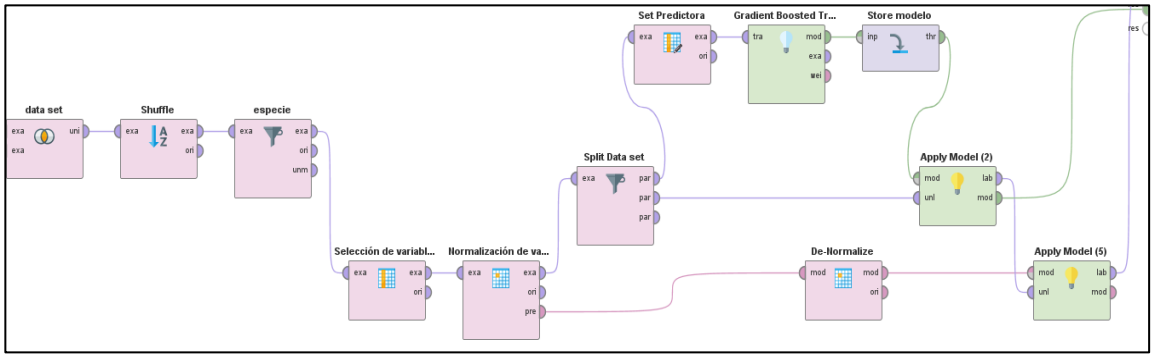
De forma general se puede resumir el proceso de la siguiente manera:

- Entrada del conjunto de datos
- Barajar el conjunto de datos de la especie
- Seleccionar las variables finales para trabajar con el modelo
- Normalizar las variables seleccionadas
- Dividir el conjunto de datos en entrenamiento 70% y prueba 30%
- Seleccionar el rol o variable dependiente, es decir la predictora
- Utilizar el modelo de Machine Learning seleccionado
- Grabar el modelo con los coeficientes devueltos por el modelo
- Desnormalizar el conjunto de datos
- Enviar los resultados a la fuente Stage (no requerido, pero útil para revisión)

A continuación, se muestran los diseños de los modelos realizados para la investigación:

Figura 26

Diseño del modelo de regresión Gradient Boosted Tree

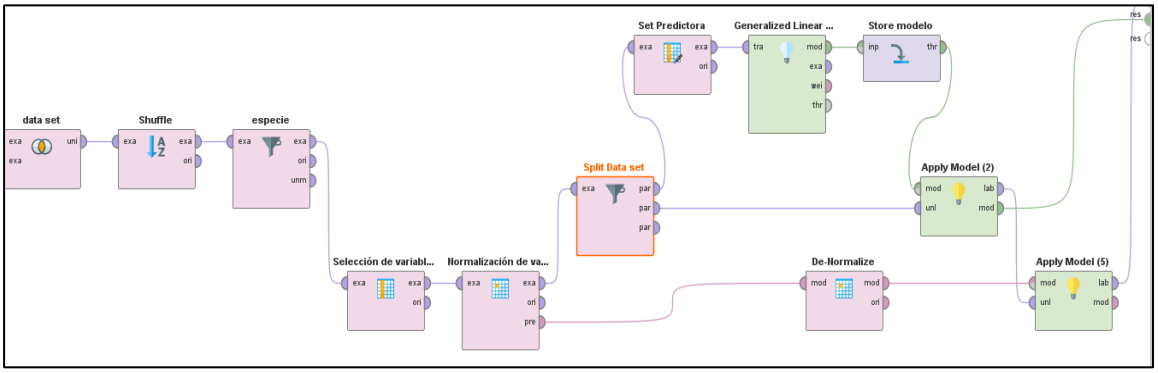


Nota: Proceso del diseño para el modelo GBT.

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 que mejora el modelo. El diseño sigue el procedimiento estándar descrito.

Figura 27

Diseño del modelo de regresión Generalized Linear

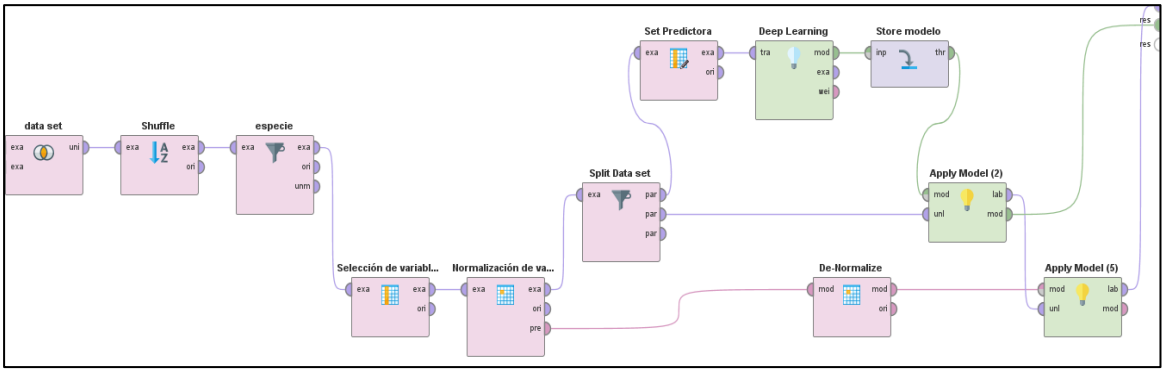


Nota: Proceso del diseño para el modelo GLM.

El modelo GLM, fue optimizado usando una distribución de Poisson para mejorar el modelo. El diseño del modelo sigue el procedimiento estándar descrito.

Figura 28

Diseño del modelo de regresión Deep Learning



Nota: Proceso del diseño para el modelo DP.

El modelo Deep Learning, fue optimizado con una función de activación Tanh, con 3 capas ocultadas de 30 neuronas cada capa, una rho de 0.999 y una función de distribución de Poisson. El modelo sigue el procedimiento estándar descrito.

Fase 5. Evaluación

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. La mejora de los modelos también obedece a modificar los parámetros propios de cada modelo, con el objetivo de mejorar la precisión de cada uno de los modelos aplicados. Las predicciones obtenidas serán validadas y determinar el cumplimiento o no de los objetivos propuestos y las hipótesis de esta investigación. Los modelos definidos en la tabla 9, son los escogidos para la tarea de regresión.

Deep Learning: “se basa en una red neuronal artificial de retroalimentación multicapa que se entrena con descenso de gradiente estocástico mediante retro propagación. La red puede contener una gran cantidad de capas ocultas que consisten en neuronas con funciones de activación” (RapimdMiner, 2021).

Generalized Linear Model: “son una extensión de los modelos lineales tradicionales. Este algoritmo ajusta modelos lineales generalizados a los datos maximizando la probabilidad logarítmica” (RapimdMiner, 2021).

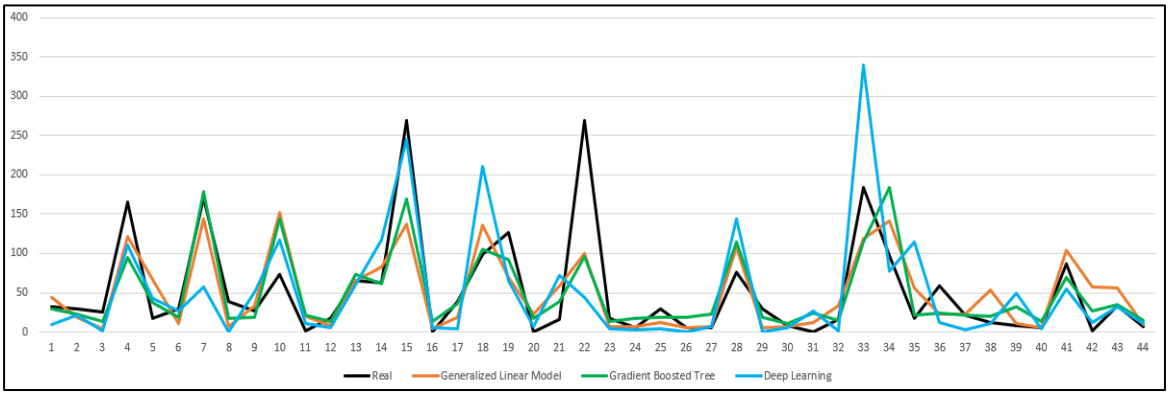
Gradient Boosted Trees: “es un conjunto de modelos de árbol de regresión o de clasificación. Ambos son métodos conjuntos de aprendizaje progresivo que obtienen resultados predictivos a través de estimaciones mejoradas gradualmente. El impulso es un procedimiento de regresión no lineal flexible que ayuda a mejorar la precisión de los árboles” (RapimdMiner, 2021).

Estos tres modelos han sido sometidos a una batería de pruebas, las cuales permitieron mediante las variables seleccionadas, determinar un excelente resultado

para cumplir con el objetivo propuesto. A continuación, se presenta uno de los modelos probados con una especie elegida aleatoriamente.

Figura 29

Comparación de ajuste-efectividad entre modelos

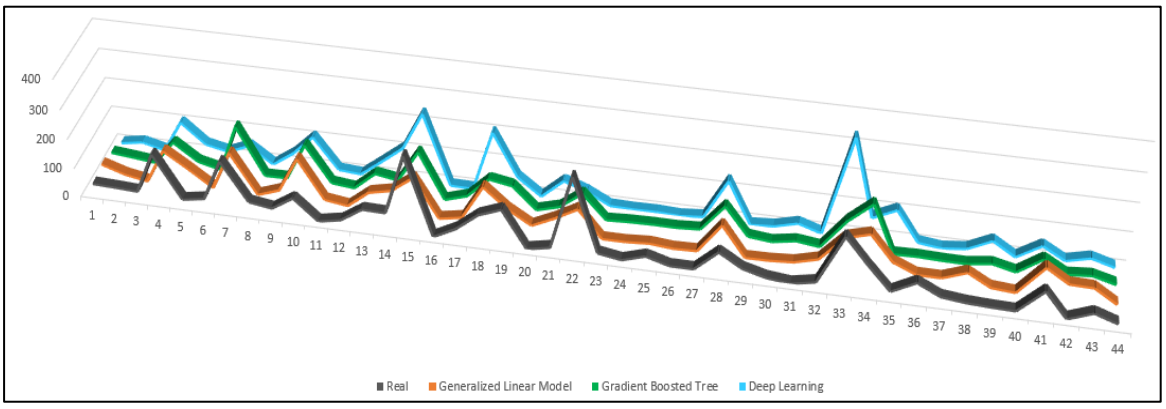


Nota: La efectividad de los modelos candidatos, logran un ajuste mayor al valor real.

El ajuste de los modelos determina la efectividad que tiene cada uno, mientras más ajustado a los datos de prueba, mejores resultados se obtienen en un ambiente real. Una vista superior permite observar el ajuste a mayor detalle.

Figura 30

Perspectiva superior del ajuste entre modelos



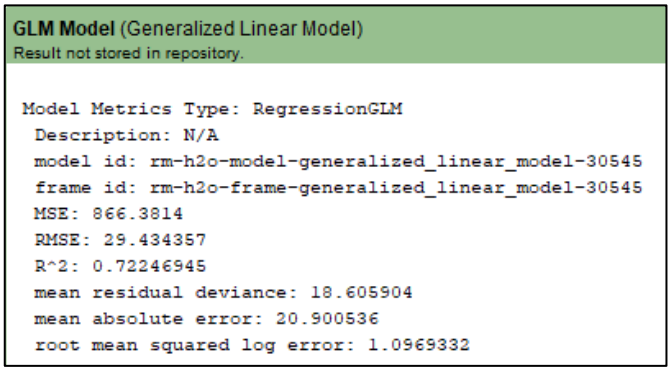
Nota: El ajuste de los modelos es más notorio, en una vista superior.

La mayoría de especies probadas con los modelos seleccionados, dan el mismo comportamiento que se observa en las gráficas. En la figura 29, 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.

Figura 31

Resultado de pruebas del modelo GLM

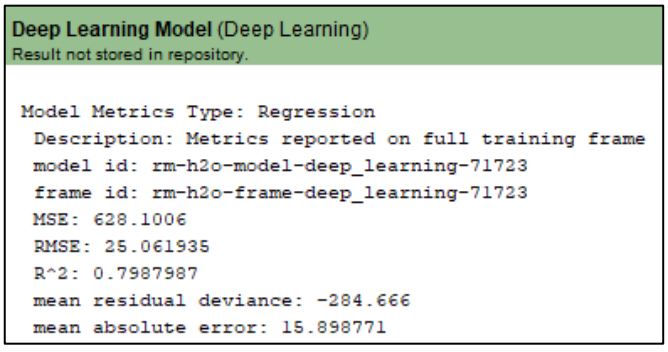


Nota: resultado del modelo GLM afinado.

El resultado de evaluar el modelo GLM, muestra números altos en el error, especialmente con MSE y RMSE, también se observa un R-cuadrado bajo respecto a los otros modelos.

Figura 32

Resultado de pruebas del modelo Deep Learning

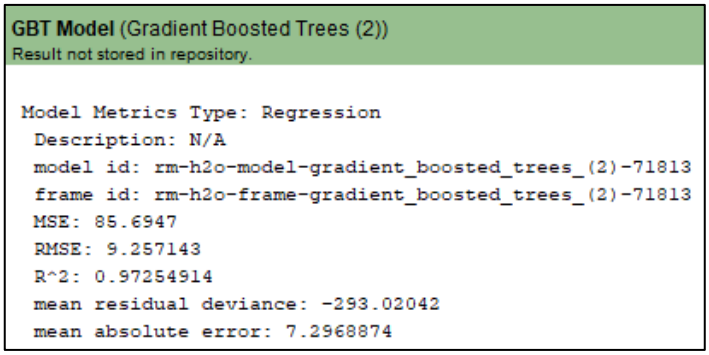


Nota: resultado del modelo DP afinado.

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.

Figura 33

Resultado de pruebas del modelo GBT



Nota: resultado del modelo GBT afinado.

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.

El cálculo para el MSE y el RMSE, están dadas por la siguiente fórmula:

Figura 34

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.

Estos cálculos son calculados directamente por la herramienta Rapidminer con el objetivo de reducir el tiempo de análisis de los modelos. Los resultados de herramienta, se presentan en la siguiente tabla.

Tabla 10

Evaluación de los modelos de regresión candidatos

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

Nota: Modelos candidatos evaluados por error cuadrático y R-cuadrado.

Los modelos candidatos son muy buenos, sin embargo, Gradient Boosted Tree y Deep Learning son muy similares en sus resultados, quedando el modelo lineal por debajo de ambos modelos. 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.

Como principales objetivos propuestos en este documento, es comprobar si las caracterizaciones de variables en relación a Intervención humana, contaminación y cambio climático, permiten demostrar que afectan en algún importante hecho, la desaparición de especies en su entorno.

Las variables seleccionadas que se ajustan mejor al conjunto de datos que darán los mejores resultados de predicción, quedando el resto de 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.

Fase 6. Despliegue

Todos los procesos anteriores se han realizado puntualizando y detallando los aspectos de cada uno, sin embargo, estos procesos tienen una entrada y una salida, convirtiéndolos un eslabón que puede ser encadenado dentro de un proceso mayor. Este encadenamiento es representado como la arquitectura de la solución del apartado 4.4.2. de este documento.

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.

El despliegue seguirá el proceso que se describe a continuación:

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.

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.

Capa Data Mart

Esta capa, maneja un modelo dimensional donde los datos tratados son subidos en para mantener el histórico de 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.

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.

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.

Capa Visualización

Finalmente, 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 Data mart, y finalmente 1 de almacenamiento o de recopilación de información de data no procesada.

Capítulo V

Discusión de Resultados

Evaluación de resultados obtenidos

La evaluación permite determinar el comportamiento de los modelos con nuevos datos, cabe decir que estos datos son los que se van produciendo dentro de una organización, con una frecuencia diaria, semanal, etc., según el ámbito de negocio. Estos nuevos datos son el insumo para mejorar los modelos en el transcurrir del tiempo, incluso los resultados de los modelos, pueden ser analizados de diferentes formas de acuerdo al giro de negocio y los objetivos y visión de las organizaciones.

Para responder la pregunta inicial de esta investigación, fue determinar que variables permiten contestar al objetivo inicial, que es tener un algoritmo que permita alertar tempranamente cuando una especie estaría en peligro, para tomar medidas a tiempo. Dado que no se tiene esta información a futuro, se decide por investigación netamente académica, 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. Se obvia la parte matemática detrás de las proyecciones, insistiendo que tal recurso es con fines de investigación académica y no será tratado en esta investigación. Este nuevo data set, será el insumo para responder las preguntas y objetivos de esta investigación.

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 lista en la siguiente tabla:

Tabla 11

Especies seleccionadas para ejecución de modelos

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

Nota: Selección de especies objetivos para ejecución de modelos de predicción.

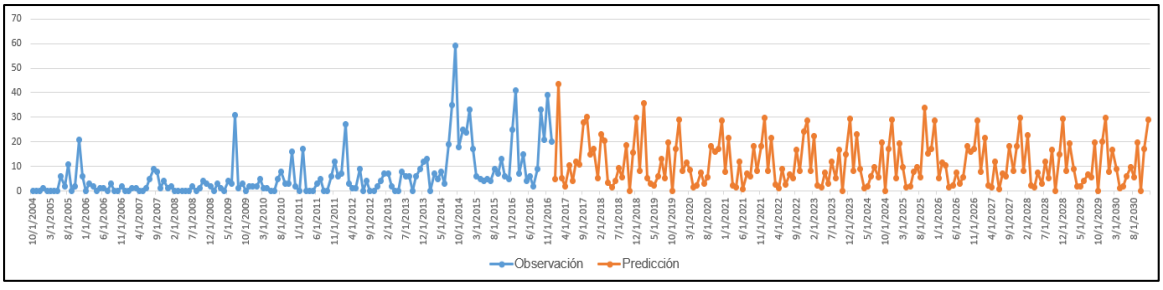
A continuación, se ejecutan los modelos de las especies de aves seleccionadas como objeto de estudio.

Ejecución del modelo para Morphnarchus prínceps - Gavilán Barreteado

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.

Figura 35

Curva de predicción para Gavilán Barreteado



Nota: Resultado de predicción para el 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.

Tabla 12

Estadística comparativa entre el histórico y la predicción para el Gavilán 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

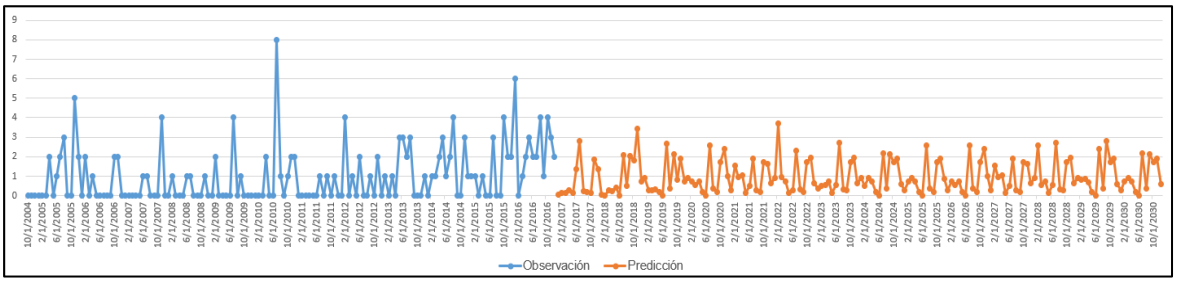
Nota: Resumen estadístico aplicado a los datos históricos de avistamientos frente a los valores predichos para el Gavilán Barreteado.

Ejecución del modelo para Spizaetus isidori - Águila Andina

El Águila Andina tiene un histórico de observaciones muy bajas, sin embargo, la curva predictora mantiene esa tendencia, con avistamientos más frecuentes que el histórico.

Figura 36

Curva de predicción para el Águila Andina



Nota: Resultado de predicción para el Águila Andina.

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.

Tabla 13

Estadística comparativa entre el histórico y la predicción para el Águila Andina

Medida	Histórico	Predicción
Media	0.98	0.95
Mediana	0	0.73
Desviación estándar	1.41	0.83
Min	0	0.01
Max	8	3.71
Rango	8	3.71
Cantidad	144	159.91

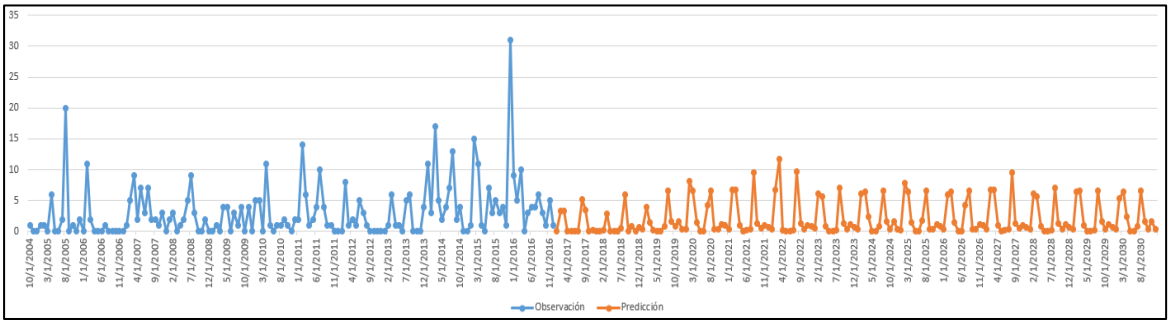
Nota: Resumen estadístico aplicado a los datos históricos de avistamientos frente a los valores predichos para el Águila Andina.

Ejecución del modelo para Lafresnaya lafresnayi 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.

Figura 37

Curva de predicción para el Colibrí Terciopelo



Nota: Resultado de predicción para el Colibrí Terciopelo.

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.

Tabla 14

Estadística comparativa entre el histórico y la predicción para el Colibrí Terciopelo

Medida	Histórico	Predicción
Media	3.11	2.06
Mediana	2	0.85
Desviación estándar	4.39	2.71
Min	0	0
Max	31	11.66
Rango	31	11.66
Cantidad	458	345.88

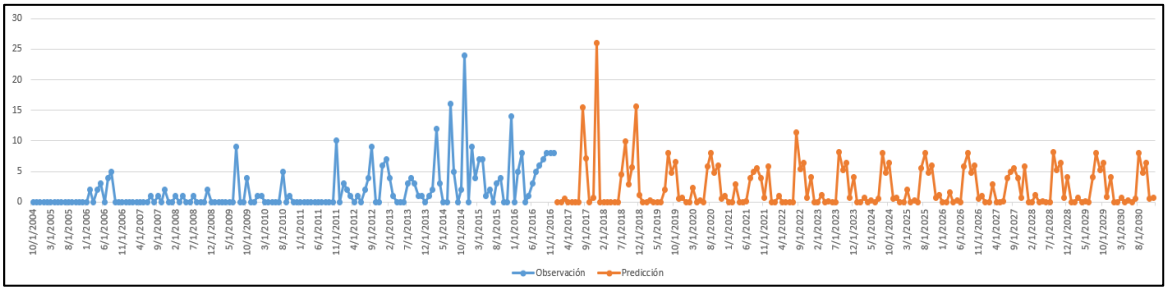
Nota: Resumen estadístico aplicado a los datos históricos de avistamientos frente a los valores predichos para el Colibrí Terciopelo.

Ejecución del modelo para Cephalopterus penduliger Pájaro Paraguas Longuipéndulo

Para esta especie, el modelo indica que se tienen avistamientos muy bajos, pero regulares y se mantendrá así desde 2019 hasta el 2030, con una disminución del 2017 al 2022.

Figura 38

Estadística comparativa entre el histórico y la predicción para el Pájaro Paraguas



Nota: Resultado de predicción para el Pájaro Paraguas.

La estadística muestra que habrá más avistamiento de ejemplares por mes que el histórico, con mayor número por avistamiento, la dispersión de la curva similar a la anterior, un pico máximo de ejemplares avistado de 4 después del 2022.

Tabla 15

Estadística comparativa entre el histórico y la predicción para el Pájaro Paraguas

	Histórico	Predicción
Media	1.93	2.44
Mediana	0	0.64
Desviación estándar	3.56	3.65
Min	0	0
Max	24	26
Rango	24	26
Cantidad	283	410

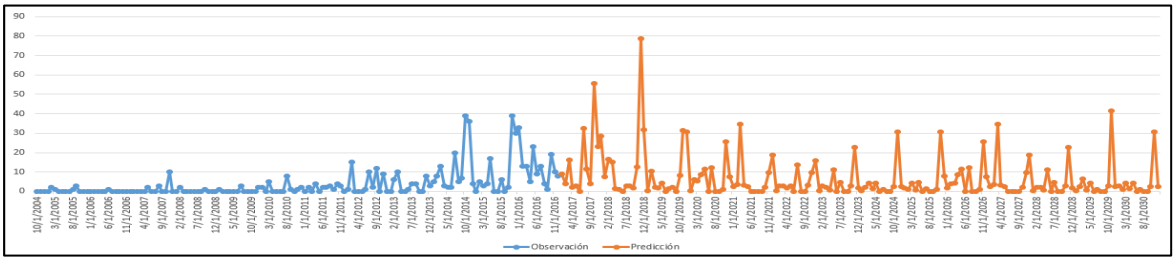
Nota: Resumen estadístico aplicado a los datos históricos de avistamientos frente a los valores predichos para el Pájaro Paraguas.

Ejecución del modelo para Vultur gryphus Cóndor Andino

Esta especie es muy bien conocido en el país, siendo el emblema nacional. La curva muestra una tendencia al alza hasta 2019, con un decrecimiento desde el 2020 hasta el 2023, y ser repetirá entre 2027 y 2029. Sin embargo, los picos indican las temporadas de mayor avistamiento.

Figura 39

Estadística comparativa entre el histórico y la predicción para el Cóndor Andino



Nota: Resultado de predicción para el Cóndor Andino.

La estadística muestra que se tendrá mayor número de ejemplares en promedio, se muestra un incremento en la cantidad de avistamientos, aunque las tomas serán más dispersas con un pico alto en 2018, con un incremento considerable en la cantidad de observaciones hasta el 2030.

Tabla 16

Estadística comparativa entre el histórico y la predicción para el Cóndor Andino

Medida	Histórico	Predicción
Media	3.80	6.73
Mediana	0	2.54
Desviación estándar	7.47	11.18
Min	0	0.01

Max	39	78.73
Rango	39	78.72
Cantidad	559	1131.99

Nota: Resumen estadístico aplicado a los datos históricos de avistamientos frente a los valores predichos para el Cóndor Andino.

Análisis de resultados

En este apartado, se analiza el resultado obtenido de los modelos, es decir numéricamente se revisa que las curvas obtenidas arrojen los resultados que permita conocer si la hipótesis planteada es verdadera o falsa.

El análisis de los resultados, se contrastará 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.

Tabla 17

Categorías del libro rojo de aves

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

Nota: El libro rojo agrupa a las especies por niveles de riesgo.

Análisis comparativo entre las muestras

Al cruzar las especies seleccionadas con los datos de la lista roja, se tiene la evaluación del riesgo que acompaña a cada especie, según se muestra en la tabla.

Tabla 18

Nivel de riesgo de las especies seleccionadas

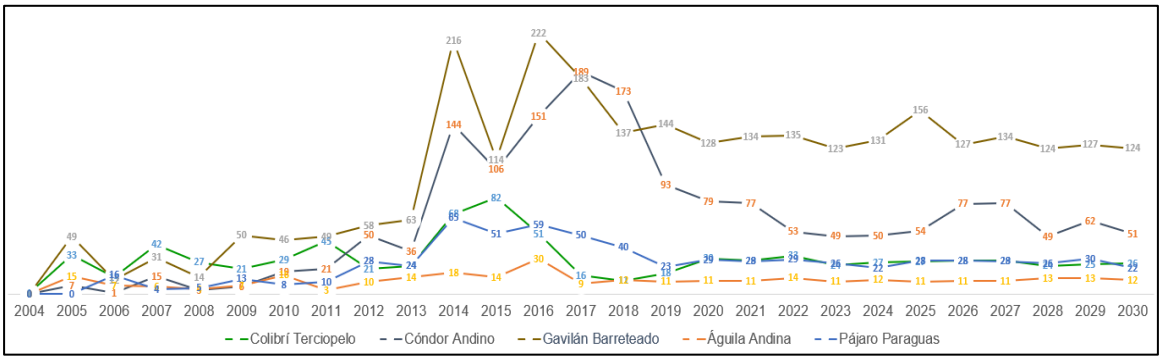
Nombre Científico	Nombre Común	Categoría	Tipo	Tipo Amenaza
Vultur gryphus	Cóndor Andino	En Peligro	EN	4
Morphnarchus princeps	Gavilán Barreteado	Casi Amenazada	NT	6
Lafresnaya lafresnayi	Colibrí Terciopelo	Preocupación Menor	LC	7
Spizaetus isidori	Águila Andina	Críticamente Amenazada	CR	3
Cephalopterus penduliger	Pájaro Paraguas Longuipéndulo	En Peligro	EN	4

Nota: Categorización del nivel de riesgo de las especies objetivo de este análisis.

El gráfico siguiente muestra 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.

Figura 40

Comparativo de la evolución histórica y la predictiva anual de las especies seleccionadas a través de los años

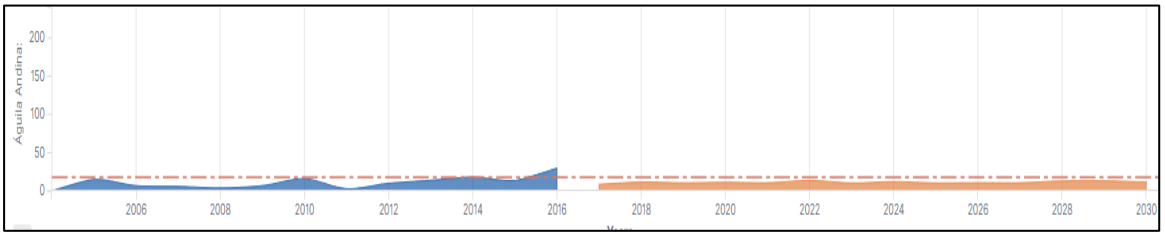


Nota: Resultados predichos de cada especie objetivo frente a los históricos.

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.

Figura 41

Evolución histórica anual de avistamientos y predicciones para el Águila Andina

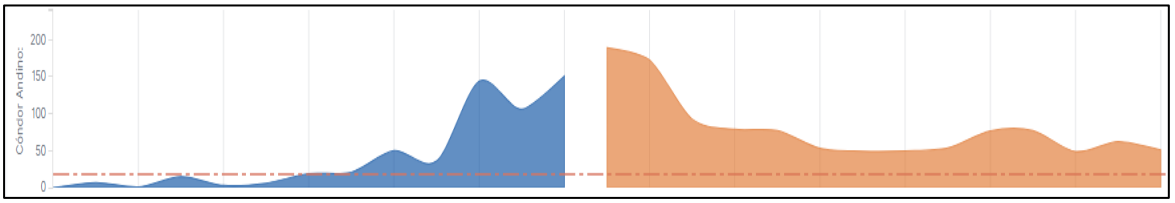


Nota: Tendencia futura en avistamientos anuales para el Águila Andina.

El águila andina, tiene un nivel de riesgo con categoría **críticamente amenazada**, declarada en el libro rojo, donde se tiene un pico de avistamientos en 2016, sin embargo, la tendencia de la curva mantiene un número de ejemplares constante hasta el año 2030, el modelo muestra que esta especie seguirá en peligro, según la tendencia de la curva, con un nivel de avistamientos por debajo de 20 observaciones anuales.

Figura 42

Evolución histórica anual de avistamientos y predicciones para el Cóndor Andino

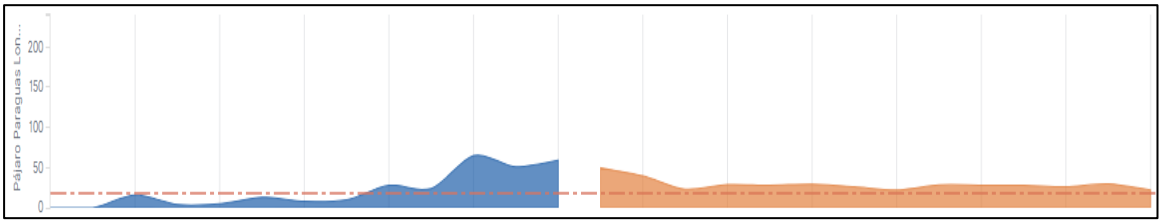


Nota: Tendencia futura en avistamientos anuales para el Cóndor Andino.

El cóndor andino tiene un nivel de riesgo **en peligro** con avistamientos máximos desde 2012 al 2022, después tendrá una población avistada entre 49 y 77 hasta el fin de la década. La curva del modelo mantiene en la misma condición de nivel de riesgo para esta especie.

Figura 43

Evolución histórica anual de avistamientos y predicciones para el Pájaro Paraguas

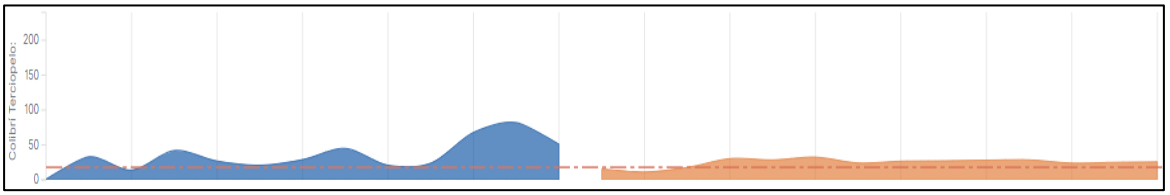


Nota: Tendencia futura en avistamientos anuales para el Pájaro Paraguas.

El pájaro paraguas tiene un nivel de riesgo **en peligro** con pico máximo de avistamientos en 2014, a partir de esta fecha hay una tendencia a la baja hasta el año 2019 y finalmente se tendrán avistamientos de entre 22 y 30 ejemplares hasta el año 2030. La curva del modelo indica mayores avistamientos, pero la mantiene en el mismo nivel de riesgo.

Figura 44

Evolución histórica anual de avistamientos y predicciones para el Colibrí Terciopelo

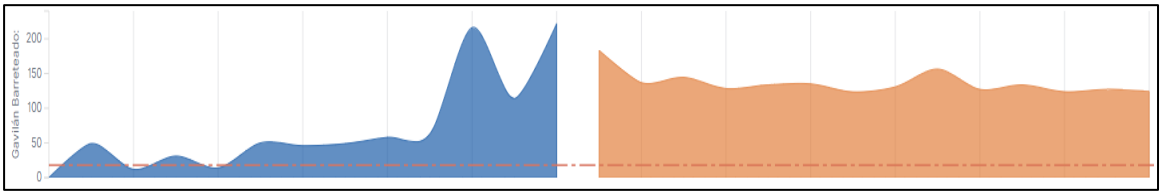


Nota: Tendencia futura en avistamientos anuales para el Colibrí Terciopelo.

El colibrí terciopelo con un nivel de riesgo **preocupación menor** con un máximo pico en el año 2015, tendrá avistamientos bajos entre los años 2017 y 2019, para luego llegar a un rango de avistamientos entre 24 y 33 ejemplares a partir del 2020 hasta final de década. La tendencia mantiene a esta especien en la misma categoría de riesgo.

Figura 45

Evolución histórica anual de avistamientos y predicciones para el Gavilán Barreteado



Nota: Tendencia futura en avistamientos anuales para el Gavilán Barreteado.

El gavilán barreteado es una especie con nivel de riesgo **casi amenazada** con pocos avistamientos de hasta 63 ejemplares anuales, hasta el año 2013, donde se tiene mayores avistamientos hasta el año 2017. Hasta el fin de la década de tiene un alto grado de avistamientos hasta antes del 2015, exceptuando 2014. La tendencia muestra un aumento de ejemplares vistos y con mayor frecuencia.

Tabla 19

Nivel de riesgo y frecuencia de avistamientos predichos

Especie	Nivel Riesgo	Predicción	Ejemplares
Águila Andina	Críticamente amenazada	La curva mantiene la tendencia histórica, y se observa un número invariante de ejemplares hasta el 2030	Histórico 144 Predicho 159.91 + 15.91 al alza. Mayor frecuencia de avistamientos
Cóndor Andino	En peligro	La curva indica un aumento de la especie, pero la estabiliza a partir del 2021.hasta el fin de la década se predice el doble de avistamientos	Histórico 559 Predicho 1131.99 + 572.99 al alza Mayor frecuencia de avistamientos
Pájaro Paraguas	En peligro	La curva mantiene a la especie con poca variación de avistamientos a partir del 2020, hay una predicción de hasta un 127 % más avistamientos hasta el término de la década	Histórico 283 Predicho 410 + 127 al alza Aumento de frecuencia en avistamientos
Colibrí Terciopelo	Preocupación menor	La curva muestra una tendencia menor a la histórica, se tendrá menor frecuencia de avistamientos de estos ejemplares hasta finales de la década	Histórico 458 Predicho 345.88 -112.12 a la baja Disminuye la frecuencia de observaciones
Gavilán Barreteado	Casi amenazada	La curva muestra una tendencia al alza. Una duplicidad en la frecuencia de avistamientos predice el modelo	Histórico 924 Predicho 1907 + 938 Aumento de frecuencia en avistamientos

Nota: Predicciones del modelo seleccionado frente a las especies objetivo de estudio.

Informe de Resultados

Este proyecto de investigación está centrado en descubrir tempranamente a través de algoritmos de predicción sobre las observaciones visuales en sitio de las diferentes especies de aves por investigadores para llevar un conteo que permita identificar cuando una especie dentro de su hábitat disminuye, aumenta su población, migra, etc. Esta tarea de campo tiene un importante uso para investigaciones y en especial con las nuevas tecnologías de minería de datos. 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.

Los objetivos definidos en este documento es crear un modelo que permita alertar de forma temprana si una especie está en peligro. Para responder a esta problemática se recopiló información de distintas fuentes para luego ser correlacionadas para formar un modelo en base a variables que permitan obtener una alerta anticipada.

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.

Tabla 20

Resumen de la predicción de alertas para las especies seleccionadas

Especie	Nivel Riesgo	Alerta temprana
Águila Andina	Críticamente amenazada	Se mantiene en iguales condiciones la cantidad de avistamientos anuales
Cóndor Andino	En peligro	Se observa un incremento en la cantidad de avistamientos anuales de ejemplares
Pájaro Paraguas	En peligro	Se observa un incremento en la cantidad de avistamientos anuales de ejemplares
Colibrí Terciopelo	Preocupación menor	Se observa un decremento en la cantidad de avistamientos anuales de ejemplares
Gavilán Barreteado	Casi amenazada	Se observa un importante incremento en la cantidad de avistamientos anuales de ejemplares

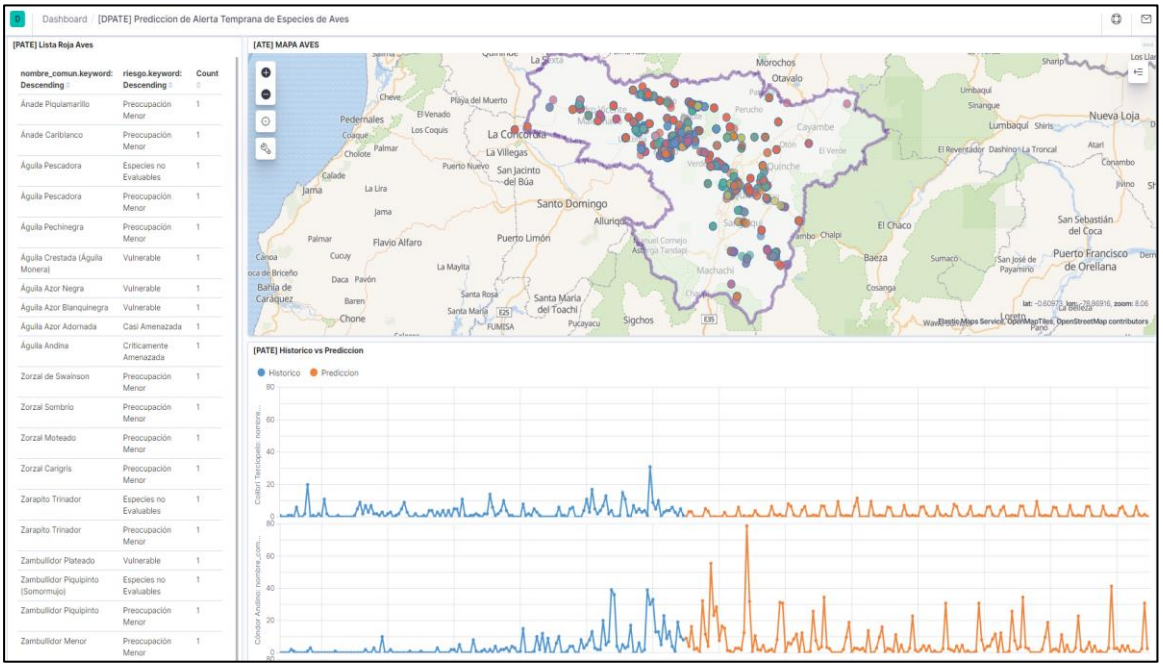
Nota: Alertas obtenidas del resultado predicho por el modelo GBT.

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.

Figura 46

Visualización de los resultados del modelo de predicción de alerta temprana



Nota: Dashboard de resultados de predicción de alerta temprana de especies, elaborado en la herramienta Kibana.

Capítulo VI

Conclusiones y Recomendaciones

Conclusiones

- 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.
- Se diseñó y construyó una arquitectura con alcance para Big Data, con el objetivo de un crecimiento escalable y orientado al volumen datos, cantidad de variables, procesamiento, etc., con el fin de abarcar y soportar todos los procesos establecidos en las metodologías.
- 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.

Recomendaciones

- Se recomienda con alta importancia la utilización de metodologías tanto para la parte general de la investigación como la parte de minería de datos, dado que establece una trazabilidad integral en los procesos para el desarrollo y cumplimiento de los objetivos propuestos.

- Es recomendable mantener un proceso de calidad de datos en todo el proceso desde la recolección hasta la salida de información representada en los tableros visuales para garantizar resultados efectivos, dado que tienen diferentes orígenes de generación de datos.
- Todo proyecto de ciencia de datos, deben tener una arquitectura que cubra todos los procesos de la corriente de datos que fluye desde la toma hasta la presentación de los mismos en tableros de visualización, por tanto, se recomienda que las arquitecturas diseñadas, tengan en cuenta los recursos u objetivos para lo que son utilizados.
- Se recomienda utilizar procesos de selección de variables, para realizar una reducción de variables y reducir el ruido de datos a los modelos.
- Es recomendable utilizar un modelo de Machine Learning que se adecué a los datos, el propósito de la investigación, como Gradient Boosted Tree que tiene la potencialidad de los árboles de decisión incluyendo formas escalonadas, optimizando funciones de coste, dando una mejor precisión del modelo.
- La utilización de esta investigación ayudará a mejorar la toma de decisiones a grupos de interés ambientales para proponer alternativas tempranas de protección de aves.
- La información resultante de las predicciones del modelo, ayudará a comprender cuales son las especies que necesitan una atención temprana frente a otras, por lo que se recomienda como siguiente paso, dialogar o comprometer a las organizaciones públicas o privadas en favor de la conservación de las especies de aves en la provincia de Pichincha.

Futuros estudios de investigación

Siguiendo la línea biodiversa, este estudio puede ser extendido hacia prácticamente cualquier especie, siendo esta animal o vegetal. Los objetivos pueden extenderse no solamente a riesgos, sino a estudios mucho más específicos o generales, dependiendo del objetivo propuesto.

Las variables independientes de este estudio o variables predictoras, pueden ser cambiadas por otras variables, o añadidas a otras, ampliando el rango de estudio o investigación.

Referencias Bibliográficas

BirdLife International. (2018). State of the world's birds: taking the pulse of the planet. In *BirdLife International*. https://doi.org/10.1007/BF01322725

CE. (2019). *Causas del cambio climático*. Comisión Eurpoea. https://ec.europa.eu/clima/change/causes_es

De Arruda, M. D. S., Spadon, G., Rodrigues, J. F., Goncalves, W. N., & Machado, B. B. (2018). Recognition of Endangered Pantanal Animal Species using Deep Learning Methods. *Proceedings of the International Joint Conference on Neural Networks*, *2018-July*. https://doi.org/10.1109/IJCNN.2018.8489369

Distler, T., Schuetz, J. G., Velásquez-Tibatá, J., & Langham, G. M. (2015). Stacked species distribution models and macroecological models provide congruent projections of avian species richness under climate change. *Journal of Biogeography*, *42*(5), 976–988. https://doi.org/10.1111/jbi.12479

FAO. (2016). *Alianza Mundial Por El Suelo*. http://www.fao.org/3/a-i5126s.pdf

Franklin, J., Serra-Diaz, J. M., Syphard, A. D., & Regan, H. M. (2017). Big data for forecasting the impacts of global change on plant communities. *Global Ecology and Biogeography*, *26*(1), 6–17. https://doi.org/10.1111/geb.12501

Freile, J. F., T. Santander G., G. Jiménez-Uzcátegui, L. Carrasco, D. F. C.-H., & E. A. Guevara, M. S.-N. y B. A. T. (2019). *Lista roja de las aves del Ecuador*. Ministerio del Ambiente, Aves y Conservación, Comité Ecuatoriano de Registros Ornitológicos, Fundación Charles Darwin, Universidad del Azuay, Red Aves Ecuador y Universidad San Francisco de Quito.

Freile, J. F., P. (2018). *Aves del Ecuador*. Pontificia Universidad Católica Del Ecuador. https://bioweb.bio/faunaweb/avesweb/home

Gironés, J., Casas, J., Mingillón, J., & Caihuelas, R. (2017). *Mineria_datos.pdf* (UOC (Ed.)). Oberta UOC.

Gobierno de Pichincha. (2019). https://www.pichincha.gob.ec/pichincha/datos-de-la-provincia/

HELEN SIMONS. (2009). *El estudio de caso: Teoría y práctica* (EDICIONES).

Hierro, L. (2018). *Deforestación*. El Pais. https://elpais.com/elpais/2018/06/26/planeta_futuro/1530040354_449192.html

INEC. (2019). *ESTADISTICAS*. https://www.ecuadorencifras.gob.ec/estadisticas/

Mayer-Schonbrger, V., & Cukier, K. (2013). *Big Data. La revolución de los datos masivos* (Turner (Ed.); 1st ed.). Turner.

Montenegro, C., Solitario, L. A., Manglar, S. F., & Danica Guinto, D. (2017). Niche modelling of endangered philippine birds using GARP and MAXENT. *Proceedings of the 7th International Conference Confluence 2017 on Cloud Computing, Data Science and Engineering*, 547–551. https://doi.org/10.1109/CONFLUENCE.2017.7943211

Muñoz, A. R., Márquez, A. L., & Real, R. (2013). Updating Known Distribution Models for Forecasting Climate Change Impact on Endangered Species. *PLoS ONE*, *8*(6), 1–9. https://doi.org/10.1371/journal.pone.0065462

NASA. (2020). *Giovanni*. https://giovanni.gsfc.nasa.gov/giovanni/

Omer, G., Mutanga, O., Abdel-Rahman, E. M., & Adam, E. (2015). Performance of Support Vector Machines and Artificial Neural Network for Mapping Endangered Tree Species Using WorldView-2 Data in Dukuduku Forest, South Africa. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *8*(10), 4825–4840. https://doi.org/10.1109/JSTARS.2015.2461136

OMS. (2019). *7 million deaths linked to air pollution annually*. OMS.

https://www.who.int/phe/eNews_63.pdf

Población total. (2019). Banco Mundial. https://datos.bancomundial.org/indicator/SP.POP.TOTL

RapimdMiner. (2021). *RapidMiner Tutorial*. https://academy.rapidminer.com/learning-paths/get-started-with-rapidminer-and-machine-learning

Wikipedia. (2021). *bubulcus ibis*. https://www.google.com/search?q=bubulcus+ibis&rlz=1C1CHBD_esEC912EC913&ei=_kNdYJaKAamFwbkP59essAM&gs_ssp=eJzj4tTP1TcwLEsyMzJg9OJNKk0qzUkuLVbITMosBgBkXgg0&oq=bulbucus+&gs_lcp=Cgdnd3Mtd2l6EAMYADIHCC4QChCTAjIECAAQCjIECAAQCjIECAAQCjIECAAQCjIGCAAQChAeMgYIABAKE