1ER ENTREGA PROYECTO

Machine Learning Operations Bootcamp

Fernando Alfredo Rojas Estrella

CDMX
19 de marzo, 2024

5545886730
rojasfery@gmail.com

Proyecto final individual

Machine Learning Operations Bootcamp

En el marco del curso Machine Learning Operations Bootcamp se ha definido la entrega de un proyecto final dividido en dos entregas que tiene como fin realizar la validación técnica de los conocimientos adquiridos hasta el momento con 4 módulos completados del curso. La validación técnica se abordará como un desafío que tiene como fin resolver un problema común de aprendizaje automático utilizando un dataset proporcionado “Anfibios”, permitiendo la evaluación de conocimientos teóricos en un contexto práctico, demostrando competencias en los temas vistos al momento y evaluando mis habilidades para resolver escenarios de la vida real.

1ER ENTREGABLE - 15 DE MARZO, 2024

En está primera entrega se mostrará la exploración de datos realizado sobre el dataset de “Anfibios”, en dónde se realizará un análisis exploratorio que permitirá entender a completitud la información de tal forma que se puedan cumplir las siguientes tareas:

1. Analizar y comprender el dataset proporcionado a través de un Análisis Exploratorio de Datos.
2. Determinar la pregunta que deseas abordar con un modelo de aprendizaje automático
3. Identificar por qué se necesita una estrategia de MLOps para este dataset. (Módulo 1)
4. Diseñar la arquitectura del pipeline para esta nueva iniciativa de aprendizaje automático.
5. Crear un modelo base para abordar tareas de predicción (clasificación, regresión, etc.) relacionadas con la pregunta. Este modelo no necesita una alta precisión, recall o puntuación F1; el objetivo es crear un modelo rápido para iteración. (Módulo 4)

A continuación, se hace referencia a la información del dataset asignado:

Dataset Name	Type of problem	Num of records	Num of Features Participant	Participant
Amphibians	Classification	189	23	Fernando Alfredo Rojas Estrella

Tabla con información del dataset asignado.

Como parte del desarrollo de la documentación se detallaran los siguientes puntos:

1. Comprensión del problema:
    a. Identificación y articulación del problema de clasificación.
    b. Explicación de la importancia del abordaje al problema dado.
    c. Descripción del impacto potencial obtenido al resolver el problema.

2. Entendimiento y exploración del dataset:
    a. Carga del dataset en el entorno de desarrollo.
    b. Obtención de estadísticas descriptivas.
    c. Identificación y manejo de los valores faltantes y/o atípicos.
    d. Visualización de los datos que muestren relaciones entre las variables.

3. Selección del modelo:
    a. Propuesta de posibles modelos a utilizar.
    b. Selección entre modelos.

4. Implementación del modelo base:
    a. Exploratorio de los datos.
    b. Pre procesamiento de datos (limpieza, normalización).
    c. Implementación del modelo base y ajuste de hiperparámetros.

5. MLOps y pipeline de datos:
    a. Diseño del pipeline de datos.

6. Conclusiones finales.

A continuación, se detallará cada punto propuesto de desarrollo del primer entregable.

COMPRENSIÓN DEL PROBLEMA

Con base en el dataset proporcionado se puede entender de forma general que el estudio al que se hace referencia fundamenta en una detallada compilación de datos geográficos y observacionales la interacción entre los hábitats de anfibios y las características ambientales de su entorno. La recolección de los datos está enfocada en evaluaciones de impacto ambiental para proyectos de infraestructura en Polonia, ofrece una base única para abordar cuestiones de conservación y biodiversidad. Además, presenta como objetivo el de emplear este conjunto de datos para desarrollar modelos predictivos que faciliten la identificación de áreas críticas para la conservación de anfibios, aprovechando tecnologías avanzadas de análisis espacial.

Identificación y articulación del problema de clasificación.

El problema de clasificación que abordamos se centra en la predicción de la presencia de especies de anfibios en embalses de agua. Este desafío está detalladamente descrito en la página del conjunto de datos disponible en el archivo de datos de la UCI (Universidad de California, Irvine) dedicado a los anfibios (https://archive.ics.uci.edu/ml/datasets/Amphibians). La esencia del problema radica en la utilización eficaz de datos provenientes de sistemas de información geográfica (SIG) e imágenes satelitales para analizar y predecir patrones de biodiversidad de anfibios en distintos hábitats acuáticos.

La necesidad de este análisis surge de la importancia crítica de utilizar la información de los anfibios como indicadores de la salud ambiental y su sensibilidad a las alteraciones de sus hábitats naturales. La presencia o ausencia de estas especies en áreas específicas proporciona información valiosa sobre el estado de conservación y la calidad del entorno acuático y terrestre.

El aprovechamiento de tecnologías avanzadas como el SIG y la teledetección satelital abre nuevas vías para el monitoreo ambiental ya que estas herramientas permitirán recoger, procesar y analizar grandes volúmenes de datos con precisión geográfica, facilitando la identificación de características

ambientales que influyen significativamente en la distribución y diversidad de los anfibios. El desafío, por tanto, no solo radica en la recolección y análisis de estos datos, sino también en la construcción de modelos predictivos capaces de interpretar correctamente estas influencias para así poder predecir la presencia de anfibios con una alta fiabilidad.

La construcción de modelos predictivos no solo mejora la capacidad para comprender y proteger la biodiversidad anfibia, sino que también contribuye a la conservación de los ecosistemas acuáticos y terrestres, asegurando que las intervenciones humanas que se espere realizar, se hagan de manera informada, aprovechando e impactando lo menos posible al medio ambiente.

En el caso del dataset asignado que contiene información de anfibios (https://archive.ics.uci.edu/dataset/528/amphibians), se habla de que estos datos fueron recopilados de los inventarios naturales que fueron preparados para la evaluación del impacto ambiental (EIA) para los proyectos viales planificados para 2 carreteras en Polonia, lo cuál sustenta la importancia de encontrar un modelo predictor de menor impacto medio ambiental para incidir en proyectos similares.

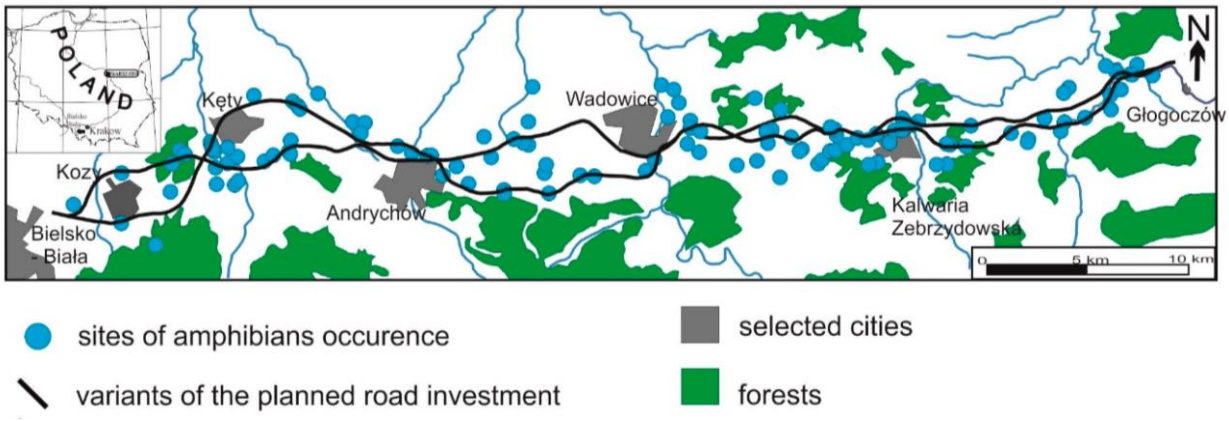

Imagen 1. Muestra de la distribución de la información.

A continuación, se presentan las variables contempladas en el dataset de anfibios:

Tabla 1. Detalle de las variables y sus catálogos.

Variable	Tipo de Atributo	Detalle	Id catálogo	Descripción de catálogo	Comentarios Adicionales
ID	Entero	Identificador único del registro (no usado en los cálculos).	-	-	-
Motorway	Categórico	Tipo de autopista (no usado en los cálculos).	-	-	-

SR	Numérico	Superficie del embalse de agua en metros cuadrados.	-	-	-
NR	Numérico	Número de embalses de agua en el hábitat. Cuanto mayor es el número, es más probable que algunos sean adecuados para la cría de anfibios.	-	-	-
TR	Categórico	Tipo de embalses de agua, que incluye características como embalses naturales o antropogénicos, estanques recién formados y de sedimentación, entre otros.	a.	Embalses con características naturales, como reservorios naturales o antropogénicos, no sometidos a naturalización.	-
			b.	Embalses recién formados, no sometidos a naturalización.	-
			c.	Estanques de sedimentación.	-
			d.	Embalses de agua ubicados cerca de viviendas.	-
			e.	Embalses de agua tecnológicos.	-
			f.	Embalses de agua en huertos familiares.	-
			g.	Zanjas.	-
			h.	Praderas húmedas, llanuras inundables, marismas.	-
			i.	Valles fluviales.	-
			j.	Arroyos y cursos de agua muy pequeños.	-
VR	Categórico	Presencia de vegetación dentro de los embalses, con categorías desde sin vegetación hasta embalses completamente cubiertos. La vegetación favorece a los anfibios pero el exceso puede ser perjudicial.	a.	Sin vegetación.	La vegetación en el embalse favorece a los anfibios, facilita la cría y proporciona alimento y refugio a las larvas. Sin embargo, un exceso de vegetación puede provocar el crecimiento excesivo del estanque y la escasez de agua.
			b.	Parches estrechos de vegetación en los bordes.	
			c.	Áreas densamente sobrecrecidas.	
			d.	Vegetación exuberante en el embalse con algunas partes sin vegetación.	
			e.	Embalses completamente sobrecrecidos con una tabla de agua que desaparece.	
SUR1	Categórico	Tipos dominantes de cobertura del suelo que rodean el embalse de agua.	a - i	Opciones mostradas en SUR3	El rasgo 'Alrededores' se asignó en tres etapas. Primero, se seleccionaron los alrededores dominantes. Luego, se escogieron dos tipos secundarios.
SUR2	Categórico	Segundo tipo más dominante de cobertura del suelo que rodea el embalse.	a - i	Opciones mostradas en SUR3	

SUR3	Categórico	Tercer tipo más dominante de cobertura del suelo que rodea el embalse, como áreas forestales, baldíos, jardines, etc.	a.	Áreas forestales (con prados) y densamente arboladas.	Los alrededores más valiosos para los anfibios son las áreas con la menor presión antropogénica y la humedad adecuada.
			b.	Áreas de baldíos y prados.	
			c.	Huertos.	
			d.	Parques y áreas verdes.	
			e.	Desarrollo denso de edificaciones, zonas industriales.	
			f.	Habitación dispersa, huertos, jardines.	
			g.	Valles fluviales.	
			h.	Carreteras, calles.	
			i.	Tierras agrícolas.	
UR	Categórico	Uso de los embalses de agua, que varía desde no utilizado hasta recreativo y económico, como la piscicultura.	a.	Sin uso por el hombre (muy atractivo para anfibios).	-
			b.	Recreativo y escénico (se realizan trabajos de cuidado).	-
			c.	Utilizado económicamente (a menudo para piscicultura).	-
			d.	Tecnológico.	-
FR	Categórico	Presencia de pesca, con categorías que abarcan desde la ausencia hasta la pesca intensiva. La pesca intensiva no es favorable para los anfibios.	a.	Falta de pesca o pesca ocasional.	La presencia de una gran cantidad de pesca, especialmente depredadora e intensiva, no favorece la presencia de anfibios.
			b.	Pesca intensa.	
			c.	Embalses de cría.	
OR	Numérico	Porcentaje de acceso desde los bordes del embalse a zonas no urbanizadas, con rangos que indican desde la falta de acceso hasta el acceso extenso.	a.	0 - 25% - falta de acceso o acceso deficiente.	-
			b.	25 - 50% - bajo acceso.	-
			c.	50 - 75% - acceso medio.	-
			d.	75 - 100% - amplio acceso a hábitats terrestres de la ribera en contacto con el hábitat terrestre de los anfibios.	-
RR	Ordinal	Distancia mínima desde el embalse de agua hasta las carreteras, con categorías que reflejan la proximidad y su efecto en la seguridad de los anfibios.	a.	Menos de 50 m.	Cuanto mayor es la distancia entre el embalse y la carretera, mayor es la seguridad para los anfibios.
			b.	50 - 100 m.	
			c.	100 - 200 m.	
			d.	200 - 500 m.	
			e.	500 - 1000 m.	
			f.	Más de 1000 m.	
BR	Ordinal	Desarrollo de edificaciones - Distancia mínima a los edificios. Cuanto mayor sea	a.	Menos de 50 m.	Cuanto más alejados estén los edificios, más favorables son las

		la distancia, más favorable será para los anfibios.			condiciones para la aparición de anfibios.
			b.	50 - 100 m.	-
			c.	100 - 200 m.	-
			d.	200 - 500 m.	-
			e.	500 - 1000 m.	-
			f.	Más de 1000 m.	-
MR	Categórico	Estado de mantenimiento del embalse, desde limpio hasta muy sucio. La basura puede devastar el ecosistema del embalse.	a.	Limpio.	La basura causa la devastación del ecosistema del embalse. También se debe considerar el relleno y nivelación de los embalses de agua con tierra y escombros.
			b.	Ligeramente basurado.	
			c.	Embalses muy o extremadamente basurados.	
CR	Categórico	Tipo de orilla del embalse, que puede ser natural o de hormigón. Una orilla de hormigón puede ser una barrera para los anfibios.	a.	Natural.	Una orilla de hormigón de un embalse no es atractiva para los anfibios. Una orilla vertical de hormigón suele ser una barrera para los anfibios cuando intentan salir del agua.
			b.	Hormigón.	-
Green frogs	Categórico; Etiqueta 1	Presencia de ranas verdes.	-	-	-
Brown frogs	Categórico; Etiqueta 2	Presencia de ranas marrones.	-	-	-
Common toad	Categórico; Etiqueta 3	Presencia de sapo común.	-	-	-
Fire-bellied toad	Categórico; Etiqueta 4	Presencia de sapo de vientre de fuego.	-	-	-
Tree frog	Categórico; Etiqueta 5	Presencia de rana arborícola.	-	-	-
Common newt	Categórico; Etiqueta 6	Presencia de tritón común.	-	-	-
Great crested newt	Categórico; Etiqueta 7	Presencia de tritón crestado.	-	-	-

Explicación de la importancia del abordaje al problema dado.

La predicción de la presencia de especies de anfibios cerca de embalses de agua cobra una importancia crítica cuando consideramos el impacto ambiental de las actividades humanas, como la construcción de carreteras. Un estudio realizado en Polonia ilustra cómo la combinación de tecnologías de información geográfica (SIG) y satelital, junto con inventarios naturales

detallados, pueden proveer datos esenciales para este propósito. En particular, se examinó la presencia de anfibios en las áreas propuestas para dos proyectos viales significativos: la Carretera A, afectando parte de la autopista A1 en Pyrzowice, y la Carretera B, en el tramo de Bielsko Biała a Wadowice-Głogoczów de la autopista S52.

Los informes de evaluación de impacto ambiental (EIA) para estos proyectos fueron cruciales para recopilar información sobre las poblaciones de anfibios en 189 sitios de ocurrencia. La metodología involucró análisis avanzados de datos SIG y satelitales, así como el escrutinio de información de inventarios naturales. Este enfoque multifacético permitió una comprensión detallada de la distribución de los anfibios y sus hábitats en las áreas afectadas por los proyectos viales.

Para el proyecto de la Carretera A, se identificaron 80 criaderos de anfibios en un área de estudio que abarcaba 500 metros a ambos lados del trazado propuesto. Este inventario fue complementado por observaciones entre 2014 y 2016, lo que ofreció una perspectiva longitudinal sobre las poblaciones de anfibios y sus dinámicas en respuesta a la alteración del hábitat.

De manera similar, el estudio de la Carretera B reveló 125 sitios reales y potenciales de presencia de anfibios. La metodología herpetológica aplicada incluyó el análisis de mapas y datos bibliográficos, seguido de observaciones de campo enfocadas en la época de primavera, cuando los anfibios están más activos. Este enfoque proporcionó una base sólida para identificar medidas de mitigación efectivas que pudieran incorporarse en el diseño y planificación de la carretera.

Los resultados de estos estudios son un testimonio del valor de integrar datos SIG y satelitales con inventarios naturales detallados. Estos hallazgos no solo subrayan la importancia de proteger a los anfibios como indicadores de la salud ecológica, sino que también ofrecen guías críticas para la planificación y desarrollo de proyectos con consideraciones ambientales. A través de este enfoque meticuloso, es posible minimizar el impacto negativo en los hábitats de anfibios, asegurando que las implicaciones ambientales sean consideradas cuidadosamente para evitar la degradación de hábitats críticos.

Este caso ilustra cómo la predicción y el estudio detallado de la presencia de anfibios cerca de embalses de agua y áreas de desarrollo humano son fundamentales para la conservación de la biodiversidad y la gestión sostenible

del impacto ambiental. La combinación de tecnología avanzada y métodos de campo tradicionales representa una estrategia prometedora para informar y guiar decisiones de conservación y desarrollo que respeten el equilibrio de nuestros ecosistemas.

Descripción del impacto potencial obtenido al resolver el problema.

El éxito en la predicción de la presencia de anfibios en embalses de agua tiene el potencial de generar impactos significativos y multidimensionales. Al abordar este desafío con una combinación de tecnologías avanzadas y análisis detallado, los resultados pueden traducirse en beneficios concretos en diversos ámbitos.

- **Conservación y biodiversidad:** El impacto más directo se vería en la conservación de especies y la protección de hábitats críticos. Al identificar áreas de importancia ecológica para los anfibios, los esfuerzos de conservación pueden ser más dirigidos y eficientes, contribuyendo significativamente a la salud y diversidad de los ecosistemas globales.

- **Desarrollo sostenible:** En el contexto del desarrollo y la planificación de obras humanas, la capacidad de predecir la presencia de anfibios permitirá incorporar consideraciones ecológicas en las etapas iniciales del diseño de proyectos. Esto no solo minimizará los impactos negativos en el medio ambiente, sino que también promoverá prácticas más sostenibles y responsables, alineadas con los objetivos de conservación a largo plazo del medio ambiente.

- **Avances científicos y metodológicos:** Desde la perspectiva científica, el refinamiento de metodologías para el monitoreo y gestión de la vida silvestre podría marcar un hito importante para tener enfoques adaptables a otras regiones del planeta. Por otro lado, los modelos predictivos avanzan en la comprensión de las interacciones entre especies y su entorno, ofreciendo herramientas valiosas para la investigación y conservación futuras.

- **Educación y conciencia ambiental:** Finalmente, los esfuerzos para resolver este problema también tienen un fuerte componente educativo. El aumento de la conciencia sobre la importancia de los anfibios y sus

roles ecológicos fomenta una mayor apreciación y responsabilidad hacia la protección del medio ambiente entre el público general, las comunidades locales y las autoridades que aprueban los proyectos de expansión de las manchas urbanas en el mundo.

En resumen, la resolución eficaz del problema de predicción de la presencia de anfibios abre caminos hacia la conservación más efectiva, el desarrollo sostenible, el avance científico, y la educación ambiental. Este enfoque integrador no solo aborda las necesidades inmediatas de protección de los anfibios, sino que también establece las bases para un futuro más sostenible y consciente del valor intrínseco de nuestra biodiversidad.

ENTENDIMIENTO Y EXPLORACIÓN DEL DATASET

Para el entendimiento del dataset, se realizó un ejercicio de integración de la documentación asociada al dataset, con lo que se integró la Tabla 1 que nos muestra el detalle preciso de cada variable disponible en el dataset, así como la información asociada a cada opción para las variables con múltiples opciones.

Carga del dataset en el entorno de desarrollo.

Primeramente, se realizó la carga del dataset en el entorno de desarrollo de Google Colab, realizándose los siguientes pasos:

- Importación de las librerías necesarias, incluyendo pandas para la manipulación de datos.
- Lectura del dataset utilizando la función pandas.read_csv().
- Explorando los datos a partir del método .head() para inspeccionar las primeras filas y obtener una visión preliminar de los datos.
- Además, se emplearon los métodos .info() y .describe() para revisar los tipos de datos presentes y obtener estadísticas descriptivas básicas, facilitando una comprensión inicial de la estructura y contenido del dataset.

Resultados: Se presentan algunas de las salidas de información asociadas al dataset cargado en el entorno de Google Colab.

```
-------- head()
    ID Motorway   SR  NR  TR  VR  SUR1  SUR2  SUR3  UR  ...  BR  MR  CR  \
0   1       A1  600   1   1   4     6     2    10   0  ...   0   0   1
1   2       A1  700   1   5   1    10     6    10   3  ...   1   0   1
2   3       A1  200   1   5   1    10     6    10   3  ...   1   0   1
3   4       A1  300   1   5   0     6    10     2   3  ...   0   0   1
4   5       A1  600   2   1   4    10     2     6   0  ...   5   0   1

   Green frogs  Brown frogs  Common toad  Fire-bellied toad  Tree frog  \
0            0            0            0                  0          0
1            0            1            1                  0          0
2            0            1            1                  0          0
3            0            0            1                  0          0
4            0            1            1                  1          0

   Common newt  Great crested newt
0            0                   0
1            1                   0
2            1                   0
3            0                   0
4            1                   1

[5 rows x 23 columns]
```

Imagen 2. Primeras líneas del dataset cargado en el entorno de ejecución.

```
--------describe()
                ID             SR          NR          TR          VR  \
count   189.000000     189.000000  189.000000  189.000000  189.000000
mean     95.000000    9633.227513    1.566138    4.952381    1.904762
std      54.703748   46256.078309    1.544419    5.590918    1.317407
min       1.000000      30.000000    1.000000    1.000000    0.000000
25%      48.000000     300.000000    1.000000    1.000000    1.000000
50%      95.000000     700.000000    1.000000    1.000000    2.000000
75%     142.000000    3300.000000    1.000000   12.000000    3.000000
max     189.000000  500000.000000   12.000000   15.000000    4.000000

             SUR1        SUR2        SUR3          UR          FR  ...  \
count  189.000000  189.000000  189.00000  189.000000  189.000000  ...
mean     4.232804    5.391534    5.84127    0.841270    0.846561  ...
std      3.434615    3.515185    3.29348    1.315291    1.349843  ...
min      1.000000    1.000000    1.00000    0.000000    0.000000  ...
25%      2.000000    2.000000    2.00000    0.000000    0.000000  ...
50%      2.000000    6.000000    6.00000    0.000000    0.000000  ...
75%      7.000000    9.000000    9.00000    3.000000    2.000000  ...
max     14.000000   11.000000   11.00000    3.000000    4.000000  ...
```

Imagen 3. Resultado de la ejecución del método describe() para algunas de las variables del dataset.

Como resultado general se encontró que el dataset contiene 189 entradas correspondientes a los 189 habitad reflejados en la información, en general no se encuentra información nula o faltante y se muestran 23 variables a considerar, de las cuáles el ID y Motorway no aportan valor alguno para el modelo que se desarrollará más adelante, esto mismo se menciona en la documentación. De cualquier forma, se realizará el exploratorio sobre todas las variables.

Obtención de estadísticas descriptivas.

Para la obtención de estadísticas descriptivas del dataset, se procedió con los siguientes pasos:

- Calcular medidas estadísticas básicas, incluyendo media, mediana, desviación estándar, valores mínimos y máximos de las variables numéricas. Fundamental para obtener una visión detallada de la distribución de los datos.
- Obtención de frecuencia de las categorías en las variables categóricas utilizando el método .value_counts(). Esto se volvió crucial para entender la diversidad y distribución de las especies de anfibios en el estudio.

Resultados: Antes de comenzar con está etapa fue necesario crear las listas de las variables por tipo, categóricas y numéricas que ayudarán en los procesos futuros para poder analizar de mejor manera cada variable.

```
-------- Listas de varibales por tipo

-------- Categóricas: ['Motorway', 'TR', 'VR', 'SUR1', 'SUR2', 'SUR3', 'UR', 'FR', 'MR', 'CR']

-------- Etiquetas: ['Green frogs', 'Brown frogs', 'Common toad', 'Fire-bellied toad', 'Tree frog', 'Common newt

-------- Numéricas: ['SR', 'NR', 'OR']

-------- Ordinales: ['RR', 'BR']

-------- Enteras: ['ID']
```

Imagen 4. Listados con variables por tipo, coincidente con la clasificación de la documentación del dataset de anfibios.

A partir de la creación de los listados de variables por tipo se determinó que las variables categóricas y de etiquetas serían consideradas como parte del tipo categóricas, mientras que el resto, numéricas, ordinales y enteras serían tratadas para efectos prácticos como variables numéricas.

Para el cálculo de las medidas estadísticas básicas se consideraron las variables numéricas obteniendo los resultados que se muestran en la siguiente imagen.

```
--------Estadísticas descriptivas
                SR          NR          OR          RR          BR
count    189.000000  189.000000  189.000000  189.000000  189.000000
mean    9633.227513    1.566138   89.962963    2.333333    2.502646
std    46256.078309    1.544419   19.904926    2.520132    2.640971
min       30.000000    1.000000   25.000000    0.000000    0.000000
25%      300.000000    1.000000   99.000000    1.000000    1.000000
50%      700.000000    1.000000  100.000000    1.000000    1.000000
75%     3300.000000    1.000000  100.000000    5.000000    5.000000
max   500000.000000   12.000000  100.000000   10.000000   10.000000
```

Imagen 5. Resultado de la obtención de estadísticas descriptivas sobre las variables consideradas numéricas del dataset objeto del estudio.

Para la obtención de frecuencias en las variables categóricas se realizaron conteos de las etiquetas en cada variable, información que fue almacenada para una futura exploración.

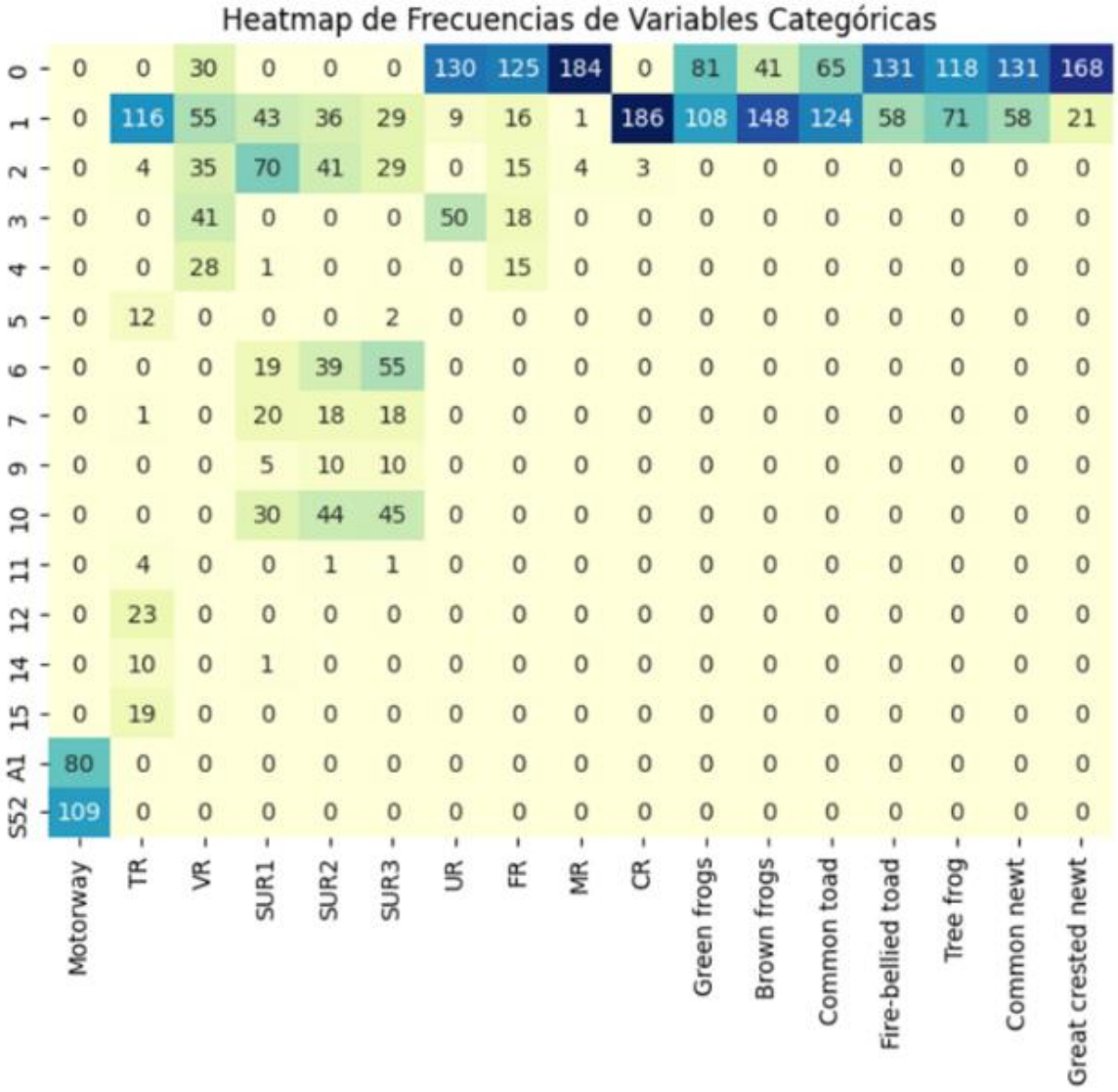

Imagen 6. Headmap del resultado del conteo de ocurrencia de cada etiqueta mostrada a la izquierda al realizar la iteración en los 189 habitad recopilados en el dataset. La visualización del heatmap solo tiene como fin el poder visualizar de forma más gráfica las etiquetas contadas para cada variable categórica.

Identificación y manejo de los valores faltantes y/o atípicos.

Para identificar y manejar los valores faltantes y/o atípicos, se realizaron los siguientes pasos:

- Empleando el método .isnull() junto con .sum() se cuantificaron los valores faltantes por variable, esto facilitó la identificación de columnas con una alta proporción de datos ausentes.

- Para el manejo de los valores ausentes se consideró el acercamiento bajo dos estrategias: imputación basada en la media o mediana para variables numéricas y moda para las variables categóricas.
- Para el manejo de los valores atípicos, se realizaron análisis de boxplot y se aplicaron métodos de filtrado basados en IQR (Rango Intercuartílico) para identificar estos valores, asegurando así la calidad y confiabilidad del dataset.

Resultados: Para la limpieza de datos se realizó la búsqueda de datos nulos en todas las variables, aunque desde el primer acercamiento utilizando info() ya se veía que no había nulos.

```
ID                0
Motorway          0
SR                0
NR                0
TR                0
VR                0
SUR1              0
SUR2              0
SUR3              0
UR                0
FR                0
OR                0
```

```
RR                0
BR                0
MR                0
CR                0
Green frogs       0
Brown frogs       0
Common toad       0
Fire-bellied toad 0
Tree frog         0
Common newt       0
Great crested newt 0
dtype: int64
```

Imagen 7. Resultado del proceso de búsqueda de nulos en la información.

Para los datos atípicos se realizó un acercamiento utilizando el método del rango intercuartílico (IQR), que generalmente es una medida de dispersión estadística que se utiliza para identificar valores atípicos. Recordemos que este método es preferido sobre otros porque no se ve tan afectado por valores extremadamente altos o bajos como lo sería, por ejemplo, utilizar la media. A continuación, se presentan los resultados para cada variable.

```
BR                0
Brown frogs      41
CR                3
Common newt       0
Common toad       0
FR                0
Fire-bellied toad 0
Great crested newt 21
Green frogs       0
ID                0
MR                5
Motorway          0
```

```
NR               41
OR               46
RR                0
SR               29
SUR1              0
SUR2              0
SUR3              0
TR                0
Tree frog         0
UR                0
VR                0
dtype: int64
```

Imagen 8. Resultado del proceso de búsqueda de “outliers” con el conteo de registros atípicos encontrados.

Después de revisar los resultados de datos atípicos consideré solo quedarme con los datos numéricos que mostraban “outliers”. Las variables SR (Superficie del embalse de agua en metros cuadrados), NR (Número de embalses de agua en el hábitat) y OR (Porcentaje de acceso desde los bordes del embalse a zonas no urbanizadas).

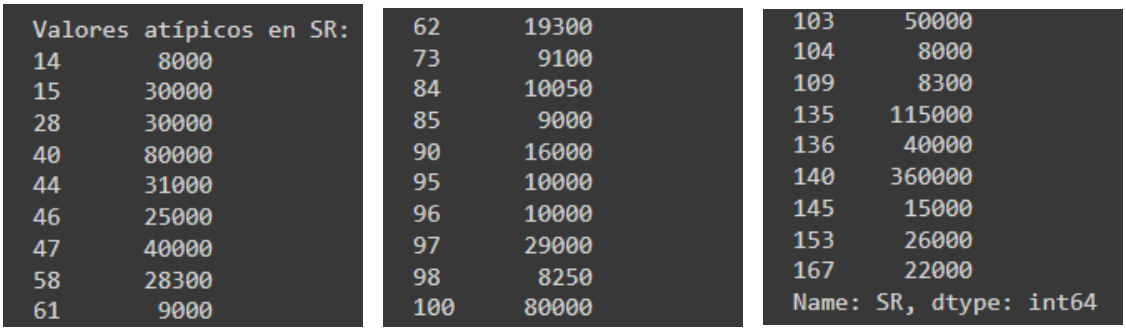

Imagen 9. Resultado de datos atípicos encontrados para la variable SR a lo largo de los 189 registros del dataset.

Después de analizar la información de las 3 variables y debido al origen de la información consideré que los datos atípicos encontrados podrían ser un reflejo de la realidad, por lo que no se realizó ninguna intervención en esos datos.

Visualización de los datos que muestren relaciones entre las variables.

La identificación de las relaciones entre las variables es un tema fundamental a llevar a cabo, esto se realizó a través de visualizaciones que permitieron observar de forma sencilla la identificación de que variables tienen mayor interdependencia con otras, para esto se siguieron los siguientes pasos.

- Utilización de gráficos de dispersión para examinar las relaciones entre variables numéricas.
- Utilización de gráficos de barras para mostrar la frecuencia de cada categoría en las variables categóricas, permitiendo identificar patrones de presencia de especies de anfibios en relación con diferentes características ambientales.
- Descubrimiento de insights importantes para guiar las siguientes etapas del análisis.

Resultados: Como parte del análisis exploratorio de datos se generó un pairplot para visualizar simultáneamente la distribución y las relaciones entre

las distintas variables numéricas del conjunto de datos. Esta herramienta gráfica como es sabido, revela patrones y correlaciones a través de múltiples gráficos de dispersión, mostrando cómo cada par de variables se relaciona entre sí. A su vez, las diagonales del pairplot proporcionan histogramas que resumen la distribución de cada variable individual.

La visualización mediante pairplot ha proporcionado un panorama general de la estructura de los datos. Dado el carácter categórico de muchas de las variables en nuestro dataset, hemos observado una tendencia a la agrupación de puntos, lo que sugiere un enfoque en el análisis de frecuencias y proporciones para estas categorías. Se realizarán gráficos de barras para ilustrar la prevalencia de cada categoría y boxplots para relacionar variables categóricas con numéricas. Esta aproximación nos prepara para sumergirnos en pruebas estadísticas más rigurosas, que discernirán las asociaciones significativas que serán fundamentales para el desarrollo de nuestro modelo predictivo.

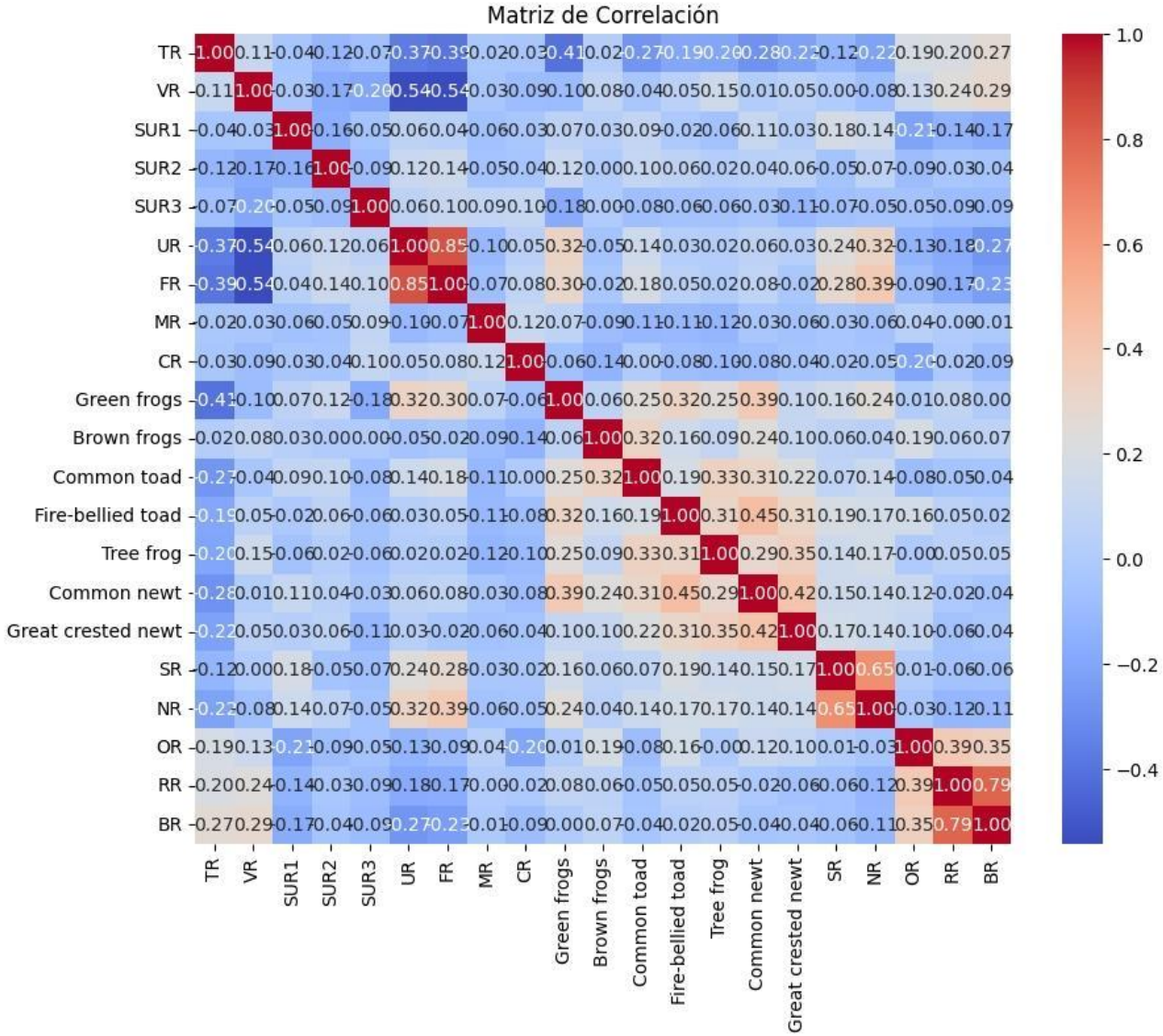

Imagen 10. Matriz de correlaciones entre variables del dataset.

Para analizar y entender la dinámica de los ecosistemas acuáticos y su impacto en la presencia de diversas especies de anfibios, se llevó a cabo un exhaustivo análisis de correlación entre distintas variables ambientales y características de los embalses de agua. Este análisis se centró en la identificación de las relaciones más significativas entre variables, tanto positivas como negativas, para lo que partiendo de la matriz de correlaciones mostrada en la imagen anterior se consideraron los valores máximos, y se decidió partir desde 0.5 para las relaciones positivas, para las relaciones negativas seleccioné -0.27 como punto de corte de umbral para cada relación. En la siguiente imagen se presentan las relaciones entre variables más significativa con el valor de su correlación.

```
-------- Correlaciones fuertes - valores únicos

  Variable1    Variable2  Correlation   Variable Pair
0        UR           FR     0.846052           FR-UR
2        SR           NR     0.652757           NR-SR
4        RR           BR     0.792273           BR-RR
0        TR           UR    -0.373549           TR-UR
1        TR           FR    -0.394966           FR-TR
2        TR  Green frogs    -0.406217  Green frogs-TR
3        TR  Common newt    -0.278239  Common newt-TR
4        VR           UR    -0.542904           UR-VR
5        VR           FR    -0.537696           FR-VR
```

Imagen 11. Par de variables con las relaciones más fuertes entre sí del dataset.

A continuación, se presentan los hallazgos más destacados de este estudio, los cuales revelan correlaciones fuertes y sugieren interacciones complejas entre el uso del suelo, características físicas de los embalses y la presencia de vida acuática. Estos resultados no solo arrojan luz sobre las condiciones óptimas para la conservación de los anfibios, sino que también proporcionan una dirección clara para continuar con los siguientes pasos del modelado.

- UR y FR (0.84): Correlación positiva muy fuerte que sugiere que el uso de los embalses (UR) está estrechamente relacionado con la presencia de pesca (FR). Esto podría indicar que los embalses utilizados para fines específicos, como recreación o pesca, tienen características comunes que afectan ambos factores.
- RR y BR (0.79): Correlación positiva entre la distancia mínima del embalse a las carreteras (RR) y la distancia mínima a los edificios (BR) sugiere que los embalses más alejados de las áreas desarrolladas tienden a estar más distantes tanto de carreteras como de construcciones, destacando embalses en áreas más naturales o menos desarrolladas.
- SR y NR (0.65): Correlación positiva fuerte entre la superficie del embalse (SR) y el número de embalses (NR) podría reflejar que áreas más grandes de agua tienden a contener más cuerpos de agua individuales, lo cual es intuitivo pero importante para entender el hábitat disponible para los anfibios.
- VR y UR (-0.54), VR y FR (-0.53): Correlaciones negativas fuertes indican que la presencia de vegetación en los embalses (VR) está inversamente relacionada con su uso (UR) y con la presencia de pesca (FR). Esto podría sugerir que la vegetación densa en los embalses hace

que sean menos aptos o atractivos para actividades recreativas o de pesca.

- TR y UR (-0.37), TR y FR (-0.39), TR y Green frogs (-0.40), TR y Common newt (-0.27): Estas correlaciones negativas fuertes con la variable tipo de embalse (TR) sugieren que diferentes tipos de embalses tienen influencias distintas en el uso del embalse, la pesca y la presencia de ciertas especies de anfibios. Esto podría indicar que los tipos específicos de embalses son menos favorables o accesibles para estos usos y especies.

Estas correlaciones fuertes proporcionan insights valiosos sobre cómo ciertas características de los embalses y su entorno están relacionadas entre sí y con la presencia de anfibios. Al construir modelos predictivos, estas relaciones pueden ser fundamentales para seleccionar las variables que serán especialmente relevantes para predecir la presencia de diferentes especies de anfibios. Además, la comprensión de estas relaciones podría indicar esfuerzos de conservación y gestión de hábitats de anfibios.

SELECCIÓN DEL MODELO

Avanzando hacia la fase de modelado en mi proyecto, es importante recordar la pregunta central que fue planteada desde el inicio con mayor detalle:

¿Cómo podemos predecir eficazmente la presencia de diversas especies de anfibios en embalses de agua utilizando datos ambientales y características específicas de estos hábitats?

La importancia de esta pregunta radica no solo en su relevancia ecológica, sino también en su potencial para informar estrategias de conservación y gestión de recursos naturales. Por lo tanto, con base en un análisis preliminar realizado en el punto anterior, se han revelado relaciones significativas entre variables ambientales y la presencia de anfibios, lo que permite iniciar la selección del modelo de machine learning más adecuado.

El modelo buscado pretende capturar la complejidad de estos ecosistemas y proporcionar predicciones precisas que puedan servir de guía para futuras investigaciones y acciones de conservación. En esta sección, se detallará el proceso de selección, implementación y evaluación del modelo, con el objetivo de desarrollar una herramienta predictiva robusta y confiable.

Propuesta de posibles modelos a utilizar.

La selección del modelo inicial es un paso crucial en el proceso de modelado de machine learning y para esta etapa inicial empezaremos con un acercamiento amplio debido a que queremos entender en lo general como funcionarían diversos modelos. El tema a abordar es un problema de clasificación complejo, es decir: predecir la presencia de diversas especies de anfibios en embalses de agua basándonos en características ambientales y datos específicos de los hábitats. Este problema se clasifica como un problema de clasificación múltiple, donde el objetivo es categorizar correctamente cada observación en una de varias clases (en este caso, especies de anfibios).

Dado este contexto, consideré varios modelos iniciales, ampliamente reconocidos por su eficacia en problemas de clasificación. A continuación, listo los modelos propuestos:

1. **Regresión logística:** A pesar de su simplicidad, la regresión logística es un punto de partida robusto para problemas de clasificación. Es particularmente útil para comprender la relación entre la variable dependiente y una o varias variables independientes.
2. **Regresión logística múltiple:** Aunque tradicionalmente se usa para clasificación binaria, puede adaptarse a problemas de múltiples etiquetas y ofrece un buen punto de partida debido a su simplicidad y facilidad de interpretación.
3. **Árboles de decisión y bosques aleatorios (Random Forest):** Los árboles de decisión son fáciles de interpretar y pueden manejar tanto datos numéricos como categóricos. Los bosques aleatorios, que construyen múltiples árboles de decisión y toman la decisión más votada, ofrecen mejor precisión y control sobre el sobreajuste.
4. **Máquinas de Vectores de Soporte (SVM):** Las SVM son efectivas en espacios de alta dimensión y pueden ser configuradas para manejar problemas de clasificación múltiple. Son particularmente útiles cuando la separación entre clases no es claramente definida.
5. **K-Vecinos más Cercanos (KNN):** Este modelo es intuitivo y sencillo, basando sus predicciones en la similitud entre ejemplos. Aunque su simplicidad puede ser una ventaja, el KNN puede volverse

computacionalmente costoso a medida que el tamaño del dataset aumenta.

6. **Redes Neuronales:** Las redes neuronales ofrecen flexibilidad y potencia, siendo capaces de capturar relaciones complejas entre las variables. Son especialmente útiles en problemas con grandes volúmenes de datos y donde las relaciones entre variables no son lineales o son de alta dimensión.
7. **Gradient Boosting Machines (GBM):** Algoritmos como XGBoost, LightGBM o CatBoost pueden manejar datos categóricos de manera eficiente y suelen proporcionar resultados de alta precisión, aunque a costa de una mayor complejidad computacional.

Cada uno de estos modelos tiene sus ventajas y desventajas, y la elección del modelo más adecuado dependerá de varios factores, incluidos la naturaleza y el tamaño de los datos, la interpretabilidad del modelo y los recursos computacionales disponibles. En la siguiente etapa del proyecto, implementaré una selección de estos modelos para determinar cuál ofrece el mejor rendimiento para nuestro conjunto de datos específico, con el fin de avanzar hacia un modelo predictivo óptimo.

Selección entre modelos.

Para abordar el desafío de clasificación múltiple en la predicción de la presencia de especies de anfibios en diversos embalses, he seleccionado inicialmente dos metodologías distintas y complementarias: Regresión logística múltiple con el enfoque de One-vs-Rest (OvR) y el algoritmo de K-Vecinos más Cercanos (KNN). Con estas técnicas se podrán explorar y modelar las complejas relaciones entre las características ambientales y la biodiversidad de anfibios desde diferentes perspectivas, simples pero potentes.

La regresión logística múltiple, aplicada mediante el enfoque One-vs-Rest, ofrece una estrategia probada para desglosar nuestro problema de clasificación múltiple en varios sub-problemas de clasificación binaria. En este contexto, se construirá un modelo de regresión logística para cada especie de anfibio, tratando de predecir su presencia contra la ausencia de todas las demás especies. Este enfoque no solo simplificará el problema complejo sino que también proporciona una visión clara sobre cómo cada variable contribuye

específicamente a la presencia de cada especie, ofreciendo insights valiosos que pueden informar esfuerzos de conservación y manejo de hábitat.

Por otro lado, el algoritmo de K-Vecinos más Cercanos nos permitirá implementar un método intuitivo basado en la similitud entre los casos de estudio. Este modelo clasificará cada embalse basándose en las etiquetas de los 'K' embalses más cercanos en el espacio de características, donde la cercanía se define por alguna medida de distancia, como la distancia euclidiana. A través de este enfoque, se capturarán los patrones complejos y no lineales en los datos que podrían ser difíciles de modelar con métodos lineales como la regresión logística.

IMPLEMENTACIÓN DEL MODELO BASE

Para la fase de implementación, se prepararán cuidadosamente los datos, ajustando los modelos para optimizar su rendimiento a través de la selección de hiperparámetros adecuados: el número 'K' de vecinos en el KNN y la regularización en la regresión logística. Posteriormente, se evaluarán los modelos utilizando métricas relevantes para problemas de clasificación, tales como la precisión, el recall y la puntuación F1, entre otros. La elección de estas dos técnicas proporcionarán una base sólida y diversa para explorar el conjunto de datos desde diferentes ángulos, maximizando las oportunidades de obtener un modelo predictivo robusto y efectivo para la conservación de los anfibios en sus hábitats naturales.

Exploratorio de los datos.

Se consideran los resultados obtenidos en la sección de exploración de datos y solo se dejan fuera las variables ID y Motorway.

Pre procesamiento de datos (limpieza, normalización).

La normalización de los datos fue un paso crítico en la preparación del dataset antes de aplicar técnicas de modelado de aprendizaje automático, sobre todo porque las características tienen escalas o rangos diferentes. Este proceso implicó ajustar la escala de los datos para que los valores de las características se transformarán dentro de un rango común, entre 0 y 1, es decir, para que tengan una media de 0 y una desviación estándar de 1. La normalización

aseguró que ninguna característica domine sobre las otras debido a su escala, permitiendo así que el modelo aprenda más efectivamente las relaciones subyacentes entre las características y las variables objetivo.

Por otro lado, algunos algoritmos que utilizan medidas de distancia como k-Vecinos más Cercanos (KNN) y que son sensibles a la magnitud de las características, como es el de Regresión logística con regularización, se benefician significativamente de la normalización, ya que esta ayuda a mejorar la convergencia del algoritmo, la precisión del modelo y su capacidad de generalización.

En mi análisis, apliqué técnicas de normalización para preparar los datos para el modelado, asegurando así que el entrenamiento y la evaluación de los modelos de clasificación fueran justos y eficientes.

Implementación del modelo base y ajuste de hiperparámetros.

Para la implementación del modelo base se realizaron los entrenamientos correspondientes con los siguiente valores:

- **Estandarización:** Las Variables categóricas fueron codificadas utilizando OneHotEncoder. Las variables numéricas fueron estandarizadas utilizando StandardScaler.
- **Variables utilizadas:** Se excluyen 'ID' y 'Motorway', ya que no aportan a la predicción de la presencia de especies de anfibios.
- **División del dataset:** Se dividió en un 80% para entrenamiento y un 20% para prueba, utilizando un random_state para reproducibilidad.
- **Modelo KNN:**
  - n_neighbors = 5
  - weights= uniform
  - metric= minkowski
  - Este modelo utiliza los 5 vecinos más cercanos con igual ponderación para clasificar cada etiqueta (especie de anfibio) basándose en la similitud de las características.
- **Modelo RLM:**
  - solver= lbfgs
  - penalty= l2
  - max_iter= 1000

- Se utilizó un enfoque One-vs-Rest para la clasificación multi-etiqueta, con regularización L2 para mejorar la generalización del modelo.

Cómo primer resultado de la ejecución de los modelos se creo una matriz con los resultados de las métricas de cada modelo:

Tabla 2. Resultados de métricas por modelo, KNN y RLM.

Etiqueta	KNN Precision	KNN Recall	KNN F1-Score	KNN Accuracy	RLM Precision	RLM Recall	RLM F1-Score	RLM Accuracy
Green frogs	0.59	0.58	0.58	0.08	0.71	0.71	0.71	0.71
Brown frogs	0.71	0.74	0.72	0.08	0.60	0.68	0.64	0.68
Common toad	0.49	0.50	0.50	0.08	0.62	0.63	0.62	0.63
Fire-bellied toad	0.72	0.74	0.72	0.08	0.72	0.76	0.73	0.76
Tree frog	0.47	0.50	0.48	0.08	0.63	0.63	0.60	0.63
Common newt	0.84	0.84	0.84	0.08	0.80	0.71	0.75	0.71
Great crested newt	0.95	0.97	0.96	0.08	0.95	0.92	0.93	0.92

De la tabla anterior se puede concluir que:

- Green frogs (Ranas verdes):RLM muestra una mejoría notable en todas las métricas comparado con KNN, especialmente en precisión y exactitud, lo cual indica una mejora en la capacidad del modelo para predecir correctamente la presencia de ranas verdes.
- Brown frogs (Ranas marrones):Aquí, KNN supera ligeramente a RLM en recall y F1-Score, pero RLM tiene una mejor precisión. Esto sugiere que RLM puede ser más conservador en sus predicciones, resultando en menos falsos positivos.
- Common toad (Sapo común):RLM supera a KNN en todas las métricas, mostrando una mejor capacidad general para predecir correctamente la presencia de sapos comunes.
- Fire-bellied toad (Sapo de vientre de fuego):RLM muestra mejoras en todas las métricas frente a KNN, lo que indica una mayor eficacia en la identificación de esta especie.
- Tree frog (Rana arborícola):Similar al sapo común, RLM supera a KNN en todas las métricas, destacándose en precisión y exactitud.
- Common newt (Tritón común):Aunque KNN tiene una mejor exactitud en esta categoría, RLM muestra una precisión superior. La exactitud superior de KNN sugiere una mejor capacidad general para clasificar

correctamente esta especie, aunque con una tasa más alta de falsos positivos que RLM.

- Great crested newt (Tritón crestado grande):Ambos modelos muestran un rendimiento excepcionalmente alto para esta especie, con RLM teniendo una ligera ventaja en precisión y exactitud.

Para acceder a la ejecución del código generado para los modelos, ir a la URL: https://github.com/rojasfery/MLOps_2024/blob/main/1er_entregable_proyecto_boot_camp_MLOps.ipynb en dónde podrán encontrar el archivo .ipynb trabajado desde Google Colab.

MLOPS Y PIPELINE DE DATOS

Antes de entrar de lleno con esta sección, es importante dar un repaso a lo que significa Machine Learning Operations (MLOps) y para eso empezaré diciendo que es un conjunto de prácticas que combina Machine Learning, DevOps, e ingeniería de datos con el fin de facilitar el ciclo de vida de los modelos de aprendizaje automático en entornos de producción. Su objetivo principal es automatizar y optimizar los procesos de construcción, prueba, despliegue y monitoreo de modelos, asegurando así la eficiencia y eficacia en el despliegue de soluciones de ML. MLOps se vuelve fundamental en el manejo con éxito de la complejidad y la dinámica de los sistemas de aprendizaje automático en producción, permitiendo un flujo de trabajo escalable, reproducible y mantenible.

Machine Learning Operations (MLOps) tiene como componentes clave del pipeline de datos los siguientes rubros:

- **Ingesta de datos:** La recolección y almacenamiento de datos constituyen el primer paso en el pipeline de datos. En el contexto de nuestro proyecto, los datos provienen de diversas fuentes, incluyendo imágenes satelitales y datos de sistemas de información geográfica (SIG), que son consolidados en un sistema de almacenamiento centralizado. La ingesta

implica asegurar la calidad y accesibilidad de los datos, así como su almacenamiento en formatos adecuados para su posterior análisis.

- **Limpieza y preparación de datos:** Este paso es crucial para garantizar la calidad de los datos que alimentarán los modelos de machine learning. Incluye la eliminación de duplicados, el manejo de valores faltantes y la transformación de los datos a un formato óptimo para el análisis. La normalización de características y la codificación de variables categóricas son ejemplos típicos de transformaciones realizadas en esta etapa.
- **Entrenamiento de modelos:** Aquí se seleccionan y configuran los modelos a entrenar, se dividen los datos en conjuntos de entrenamiento y prueba, y se ajustan los hiperparámetros. Esta fase también incluye la evaluación inicial de los modelos para determinar su eficacia en capturar las relaciones subyacentes en los datos.
- **Evaluación y validación de modelos:** Se utilizan métodos estadísticos y métricas de rendimiento para evaluar la precisión y efectividad de los modelos. Además, la validación del modelo en escenarios del mundo real es esencial para asegurar su aplicabilidad y fiabilidad en entornos de producción.
- **Despliegue y monitoreo:** Los modelos entrenados son desplegados en un entorno de producción, donde son integrados con aplicaciones y sistemas existentes. El monitoreo continuo es vital para detectar y corregir cualquier problema de rendimiento o precisión que pueda surgir, asegurando que el modelo se mantenga actualizado y relevante.

Unas de las características que se vuelven clave en el proceso de MLOps durante las canalizaciones de información son la posibilidad de la automatización, integración y despliegue continuo. La automatización del pipeline de datos y el ciclo de vida del modelo es clave para la eficiencia y rapidez en el desarrollo y mantenimiento de sistemas de machine learning. La Integración Continua (CI) y el Despliegue Continuo (CD) permiten el despliegue automático de modelos actualizados y la realización de pruebas automáticas para garantizar la calidad y el rendimiento de los modelos.

Finalmente, debemos considerar que implementar herramientas de seguimiento y registro es esencial para documentar los experimentos de entrenamiento, los parámetros de los modelos, las métricas de rendimiento y los cambios en el dataset. Esta práctica no solo facilita la reproducibilidad y la

auditoría de los modelos de machine learning, sino que también proporciona insights valiosos para la mejora continua de los propios modelos desplegados.

Diseño del pipeline de datos.

Para un diseño especifico de un pipeline de datos para el problema planteado, es importante aterrizar el uso que podría tener un pipeline de datos, considerando que existen múltiples retos ambientales actuales, como la biodiversidad y el estudio de especies como la de los anfibios que son indicadores clave de la salud de los ecosistemas. Ahora bien, pensar en tener un monitoreo de anfibios en diversos hábitats, demandaría herramientas capaces de manejar extensos datos geoespaciales para realizar predicciones acertadas sobre su presencia, por lo que enfoques de modelado de predicción se vuelven herramientas útiles.

Por tal motivo, propongo crear un pipeline de datos y un servicio web, aprovechando la tecnología de Sistemas de Información Geográfica (GIS) para predecir la presencia de anfibios de forma global. Este sistema facilitará a investigadores y planificadores ambientales el uso de análisis avanzados, integrando estas predicciones en sus trabajos.

Utilizando la infraestructura de Azure, mi enfoque se centrará en un modelo de aprendizaje automático que procesa datos GIS, prediciendo la presencia de anfibios basándose en características ambientales y geográficas. A través de una interfaz web intuitiva, los usuarios podrán subir datos GIS y obtener predicciones en tiempo real, promoviendo así la conciencia sobre la biodiversidad.

Este proyecto aplica tecnologías de punta en la conservación ambiental, ofreciendo una herramienta práctica para la evaluación de impacto, conservación, y manejo de hábitats. Integrando diversos datos para análisis predictivo, se podrán mejorar las capacidades de proteger la biodiversidad, contribuyendo a la sostenibilidad en un mundo cambiante.

A continuación, listo algunos puntos parte de la propuesta de realizar el pipeline de datos:

1. Infraestructura en la nube de Azure: La base de esta solución se sustenta en el aprovechamiento de los servicios de Azure para albergar y administrar tanto el pipeline de datos como los modelos de aprendizaje automático. Azure en la actualidad proporciona una gama de servicios especialmente diseñados para estas tareas.
    a. Azure Blob Storage: Esencial para el almacenamiento seguro y eficiente en costos de los datasets de GIS y otros datos relevantes, ofreciendo escalabilidad y accesibilidad.
    b. Azure Machine Learning Service: Clave para la construcción, entrenamiento y despliegue de modelos de aprendizaje automático en la nube. Este servicio optimiza la gestión del ciclo de vida del modelo, abarcando desde la experimentación hasta el despliegue, pasando por el entrenamiento y el ajuste de hiperparámetros.
    c. Azure Kubernetes Service (AKS): Proporciona las capacidades de orquestación y escalado automático de los contenedores que ejecutan tanto el servicio web como el modelo de predicción, asegurando su disponibilidad y eficiencia operativa.
    d. Azure DevOps: Facilita la implementación de prácticas de CI/CD, automatizando los procesos de prueba y despliegue para el servicio web y el modelo de aprendizaje automático, promoviendo un desarrollo ágil y seguro.
2. Página web y servicios web: El diseño del servicio web planteado, busca ofrecer una interfaz accesible y eficiente para los usuarios, permitiéndoles cargar datos de GIS y recibir predicciones sobre la presencia de anfibios en los espacios de interés.
    a. Front-end de la página web: Se propone el desarrollo de una interfaz de usuario intuitiva y moderna, empleando frameworks como React o Angular. Esta permitirá a los usuarios a subir archivos de datos GIS y visualizar las predicciones del modelo de manera clara y directa.
    b. API de servicio web: Se construiría una API REST, utilizando tecnologías como Flask o FastAPI en Python, y se alojaría en contenedores Docker gestionados por AKS. La API podría sería el vinculo entre los usuarios y el modelo alojado en Azure Machine Learning Service, procesando solicitudes y devolviendo predicciones de forma eficiente.
    c. Integración y despliegue continuo: A través de Azure DevOps, se propone la automatización de los flujos de trabajo de integración y despliegue para la aplicación web y la API. Esto garantizaría

actualizaciones fluidas y mantendría la alta disponibilidad del servicio.

3. Seguridad y escalabilidad: Existen aspectos fundamentales como la seguridad de los datos y la escalabilidad del sistema que se consideran prioritarios desde el inicio del proyecto.
    a. La implementación de medidas de autenticación y autorización, aprovechando Azure Active Directory, ahora Microsoft Entra ID, el cifrado de datos tanto en tránsito como en reposo, y el monitoreo constante de la infraestructura aseguran la protección de la información y la fiabilidad del sistema.

Esta propuesta pretende establecer no solo un servicio accesible y de gran valor para la predicción de la presencia de anfibios, sino que también propone una infraestructura robusta, segura y capaz de adaptarse a futuras demandas. La sinergia entre los servicios de Azure, las prácticas de MLOps y un servicio web cuidadosamente diseñado, proporcionarán una sólida base para un proyecto de largo alcance y éxito.

CONCLUSIONES FINALES.

Hasta este momento, para la 1era entrega del proyecto final, se han logrado avances notables en el análisis y modelado de datos para la predicción de la presencia de anfibios en distintos hábitats. A continuación, muestro los logros principales y las oportunidades de mejora identificadas:

- Análisis exploratorio de datos: Se ha realizado un análisis exploratorio profundo que ha permitido entender las dinámicas del dataset y las interacciones entre variables. Este paso inicial fue crucial para identificar patrones y tendencias relevantes para los modelos predictivos.
- Implementación de modelos básicos: Se han evaluado modelos básicos de aprendizaje automático, como K-Vecinos más Cercanos (KNN) y Regresión logística múltiple (RLM). Estos modelos proporcionaron insights valiosos sobre su rendimiento, estableciendo una base para enfrentar el desafío de modelado.
- Propuesta de pipeline de datos: Se realizó la propuesta de un pipeline de datos y una estrategia de MLOps aprovechando los servicios de

Azure, esenciales para una transición eficiente y sostenible de los modelos desde el desarrollo hasta la producción.

También se ha descubierto que existen áreas de oportunidad que pueden ser explotadas para la siguiente etapa de entrega del proyecto.

- Optimización de modelos: Existe espacio significativo para explorar algoritmos avanzados y técnicas de ensamble, potencialmente mejorando la precisión de las predicciones obtenidas hasta el momento.
- Ajuste de hiperparámetros: La optimización de hiperparámetros mediante técnicas como búsqueda en cuadrícula y validación cruzada es crucial para maximizar el rendimiento de los modelos.
- Implementación real del pipeline de datos: No me queda claro si se tendrá que realizar una implementación práctica del pipeline de datos y la estrategia de MLOps propuesta, sin embargo, el realizarlo como se ha indicado con la integración de servicios de Azure, permitiría probar de forma real la automatización de procesos y el establecimiento de prácticas de monitoreo continuo.
- Interfaz de usuario y accesibilidad: Se vuelve vital el desarrollar una interfaz intuitiva que facilite a los usuarios finales el acceso a las predicciones de los modelos.

Por toro esto, me parece que el proyecto ha creado una base robusta para el análisis predictivo de la presencia de anfibios, cubriendo tanto el aspecto técnico del modelado de datos como la implementación práctica. Aunque se han logrado importantes avances, aún quedan desafíos importantes para optimizar y aplicar efectivamente los modelos en contextos reales de conservación. Los próximos pasos del proyecto incluirán avanzar en la experimentación con técnicas de modelado más sofisticadas, la afinación de hiperparámetros, y la puesta en marcha efectiva del pipeline de datos. Estos esfuerzos me acercarán al objetivo del modelado de datos de contribuir a la conservación de la biodiversidad y promover la sostenibilidad ambiental.

BIBLIOGRAFÍA

Amphibians. (2020). UCI Machine Learning Repository. Recuperado de https://archive.ics.uci.edu/dataset/528/amphibians

Arquero, KJ; Kimes, RV Caracterización empírica de medidas de importancia de variables aleatorias de bosques. Computadora. Estadística. Análisis de datos. Recuperado de https://dx.doi.org/10.1016/j.csda.2007.08.015

Blachnik, M., Sołtysiak, M., & Dąbrowska, D. (2019). Predecir la presencia de especies de anfibios utilizando características obtenidas de SIG e imágenes de satélite. ISPRS International Journal of Geo-Information. Recuperado de https://www.mdpi.com/2220-9964/8/3/123#sec4-ijgi-08-00123

Buitrago, B. (2020, 14 de septiembre). Aprendizaje automático — Modelos de Regresión I. iWannaBeDataDriven. Recuperado de https://medium.com/iwannabedatadriven/machine-learning-modelos-de-regresi%C3%B3n-i-d293ae235e9a

Buitrago, B. (2020, 14 de septiembre). Aprendizaje automático — Modelos de Regresión II. iWannaBeDataDriven. Recuperado de https://medium.com/iwannabedatadriven/machine-learning-modelos-de-regresi%C3%B3n-ii-18abc01a9848