

MARINE LIFE FORESCASTING

David González Pérez



Fuentes de datos



GBIF



Open Street Map



Natural Earth



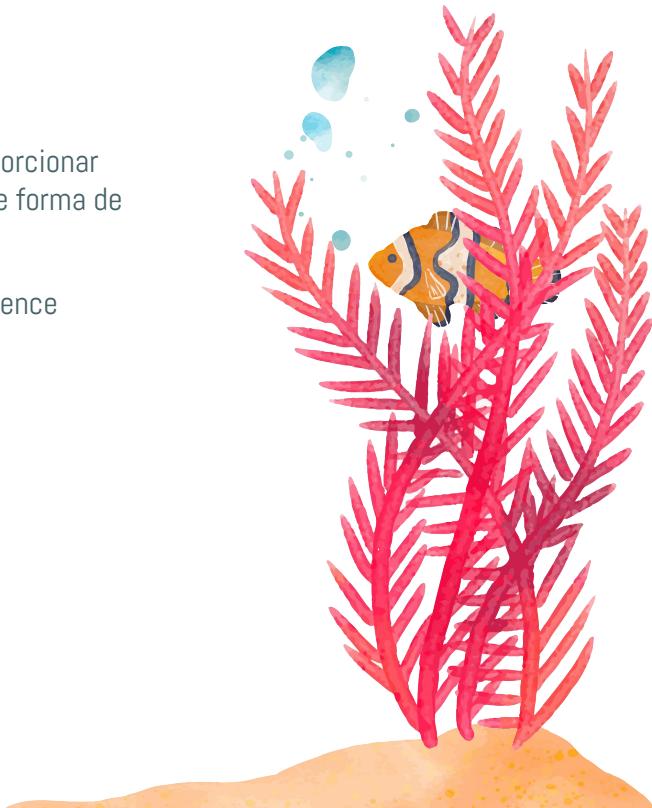
Wikipedia

Fuentes de datos



Organización internacional destinada a proporcionar acceso abierto a datos sobre cualquier tipo de forma de vida que hay en la Tierra.

- Diveboard - Scuba diving citizen science observations
- Species API



Fuentes de datos

Open Street Map

Conjunto de datasets de mapas de dominio público a diferentes escalas, descargado con Cartopy



Fuentes de datos

Natural Earth

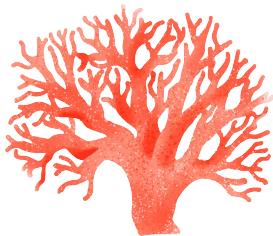
Es un mapa del mundo, de uso libre bajo una licencia abierta, descargado con Geopy



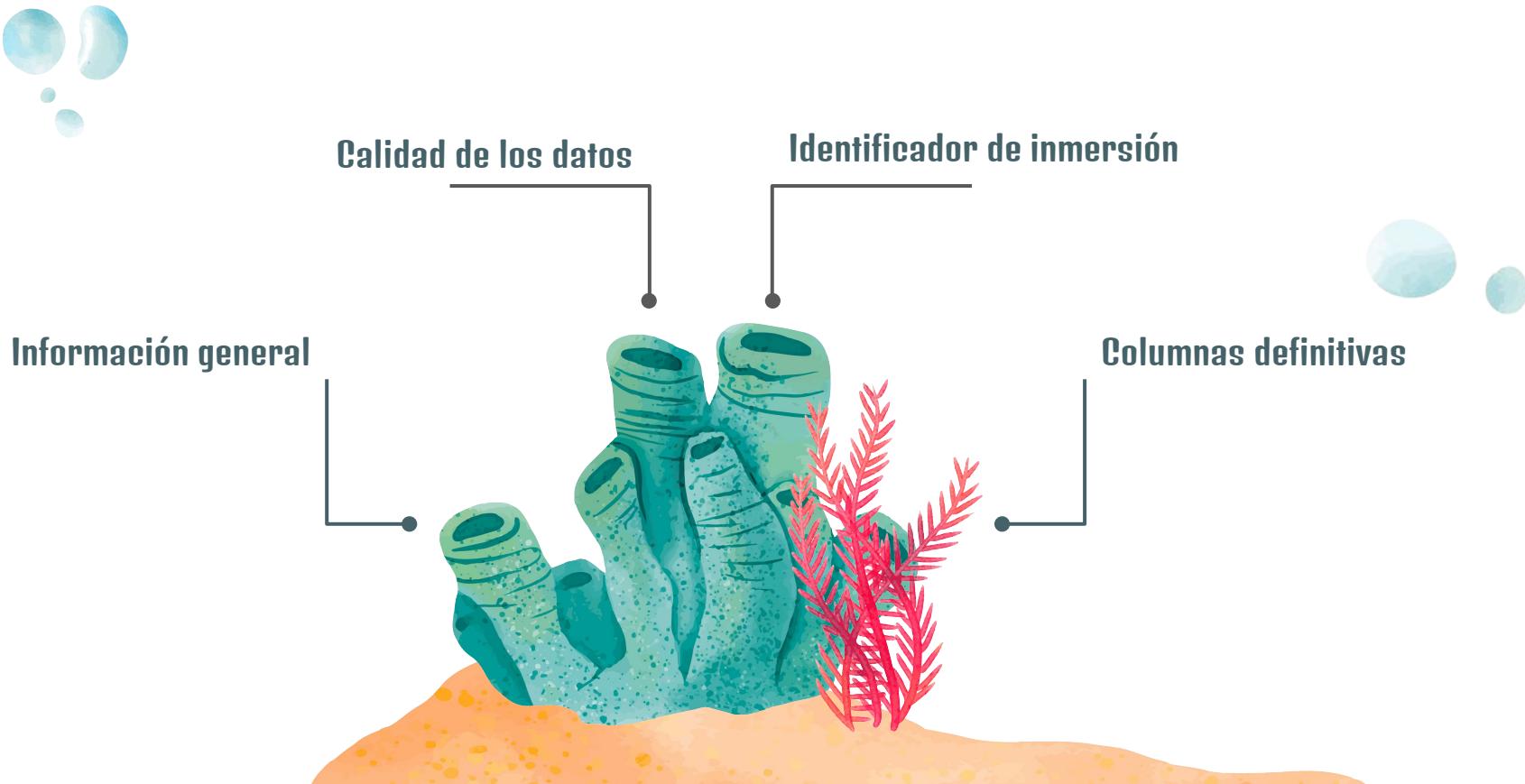
Fuentes de datos

Wikipedia

Es un proyecto enciclopedia web multilingüe de contenido libre basado en un modelo de edición abierta, descargado con la librería Wikipedia



Análisis de datos y preprocesado



Análisis de datos y preprocesado

Información general

Diveboard - Scuba diving citizen science
observations

- 38.324 filas
- 249 columnas

Las observaciones están categorizadas por diferentes taxonomías

Datos filtrados por el Reino Animalia:

- Animalia: 36.235
- Incertae sedis: 45
- Plantae: 29
- Chromista: 15

Análisis de datos y preprocesado

Calidad de los datos

Columnas vacías

Existen 167 columnas vacías.

Sigue el estándar Darwin Core

Columnas con pocos valores

Existen 27 columnas que no aportan información relevante

Columnas con información redundante

Las columnas depthAccuracy y datelidentified son redundantes



Análisis de datos y preprocesado

Calidad de los datos

Columnas con información incorrecta

Location, Water Body y
Country Code

Se ordena ascendenteamente y
se imputa el valor anterior

Elevation

Tiene valores superiores a 0m
en coordenadas sobre el mar.

Se elimina



Análisis de datos y preprocesado

Calidad de los datos

Columnas con información incorrecta

Coordenadas



Para las coordenadas 0,0 se han imputado los valores de otras coordenadas con la misma localidad y país

Los valores restantes a 0,0 se han eliminado

Con la cartografía Natural Earth se ha intentado establecer qué coordenadas están en tierra y cuáles en el mar, pero no es lo suficientemente preciso



Análisis de datos y preprocesado

Identificador de inmersión



El objetivo es clasificar los animales por inmersión, y está a nivel de observación de animal.

El campo references, que es la url de cada inmersión en Diveboard, sirve como identificador.

Se asigna un nº convirtiéndolo a categoría.

Para los registros nulos, se agregan por eventDate, decimalLatitude, decimalLongitude y depth y se continúa la serie en el último asignado anteriormente.

Análisis de datos y preprocesado

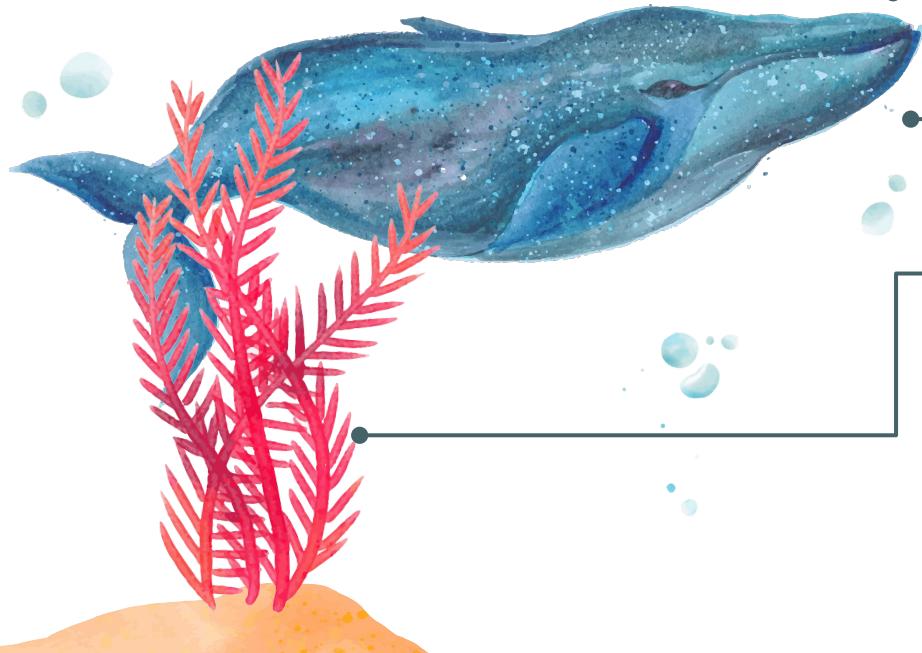
Columnas definitivas

- eventDate
- month
- day
- waterBody
- countryCode
- locality
- dive_id



- decimalLatitude
- decimalLongitude
- depth
- orderKey
- familyKey
- genusKey
- speciesKey

Enriquecimiento del dataset



Nombres comunes de los animales

Extracción de la hora

Selección de target

Agregación por inmersión

Enriquecimiento del dataset

Nombres comunes de los animales

Se aprovechan las claves de las taxonomías de los animales

Se usan esas claves para utilizarlas con la API de GBIF y obtener los nombre comunes de cada taxonomía

El orden de selección es:

- Familia
- Orden
- Género
- Especie



Enriquecimiento del dataset



Extracción de la hora

Se emplea el campo eventDate para extraer la hora de inmersión



Selección de target

1. Moray eels
2. Firefishes
3. Damselfishes
4. Sea turtles
5. Groupers



Agregación por inmersión

Se agrega el dataset por el identificador de inmersión. Si uno de los 5 animales es visto en una inmersión se informa con 1, en caso contrario, con 0



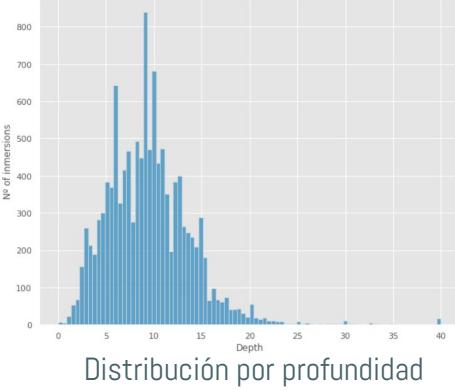
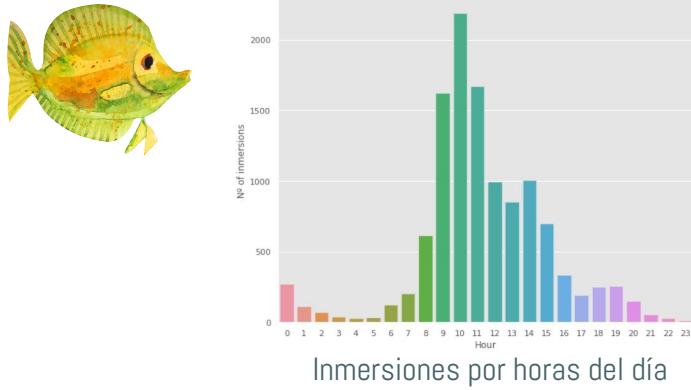
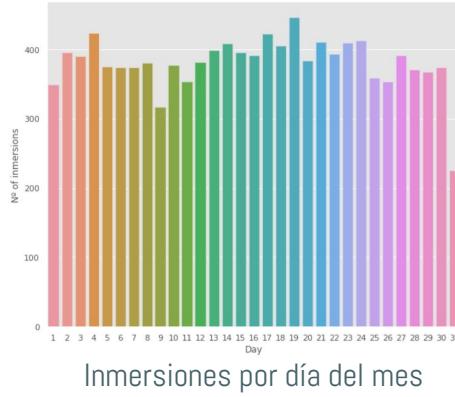
Insights

Immersiones

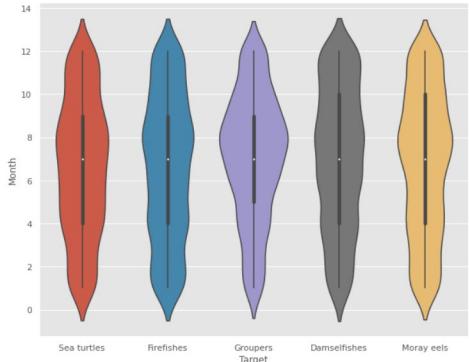
Target

Insights

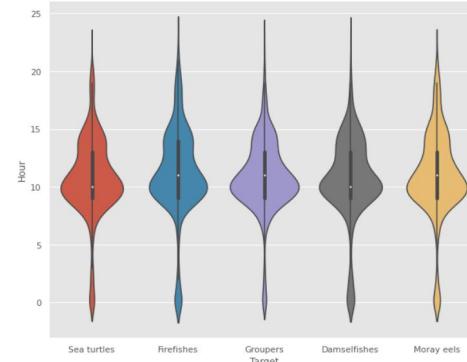
Inmersiones



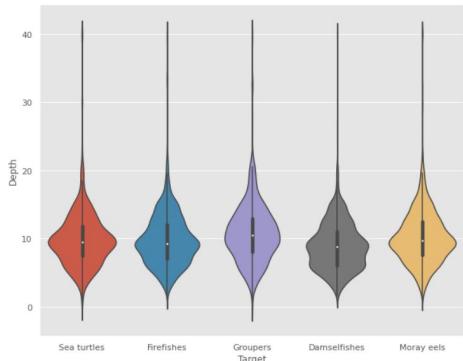
Insights Target



Meses



Horas



Profundidad



Feature Engineering



Variables
cíclicas



Variables
continuas



Variables
categóricas



Feature Engineering

Variables categóricas

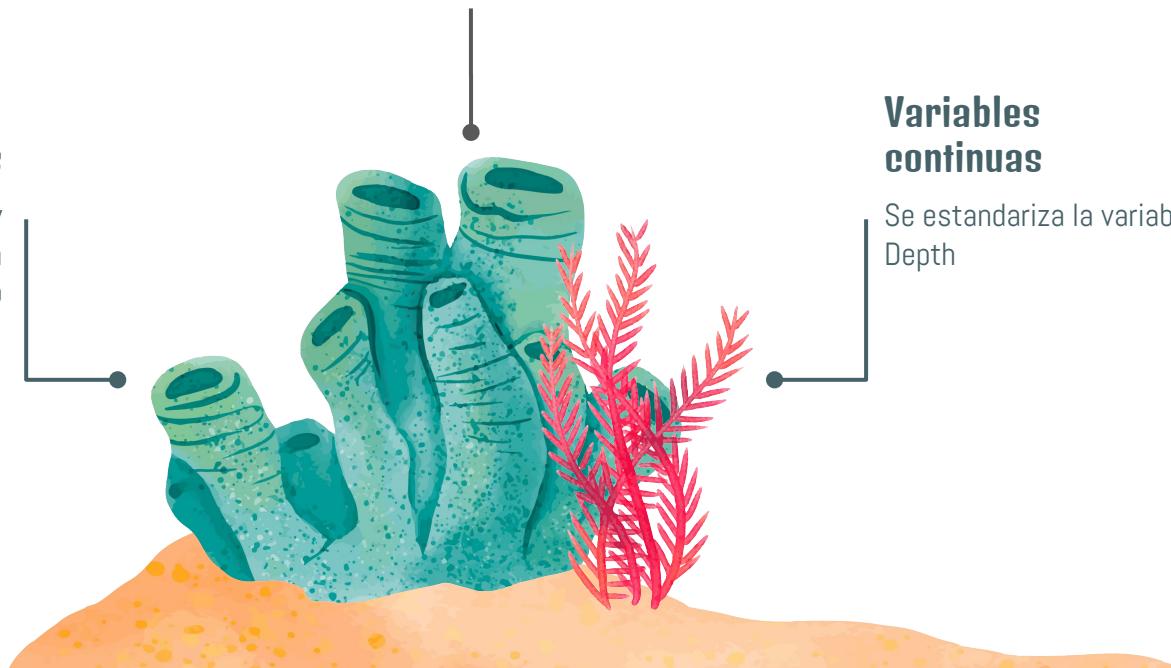
Se ha usado el mean target encoding para las variables Water Body, Locality y Country Code

Variables cíclicas

Las variables Hour, Day y Month se han convertido a seno y coseno

Variables continuas

Se estandariza la variable Depth



Importancia de las variables

Variables
categóricas

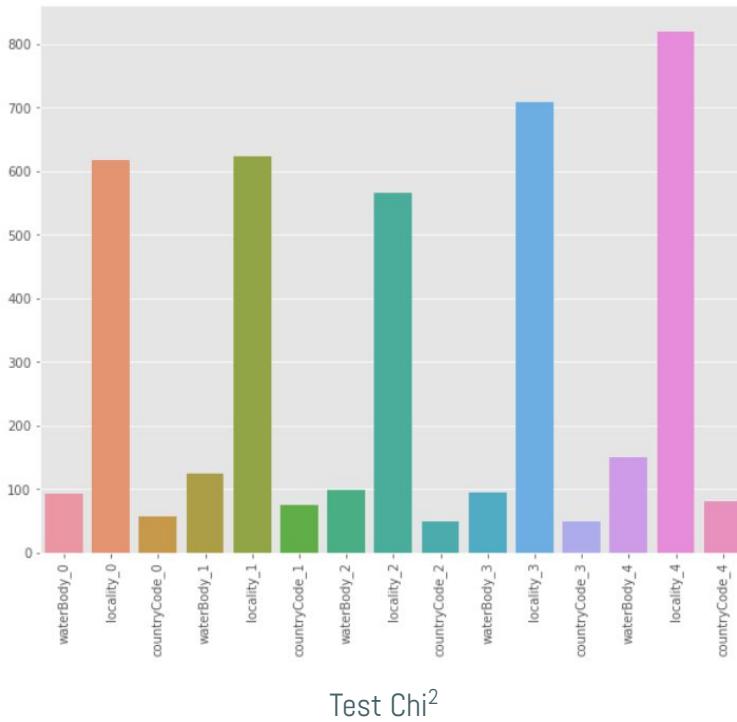


Variables
numéricas



Importancia de las variables

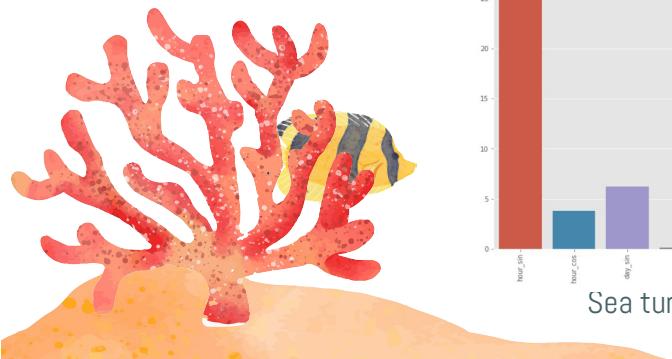
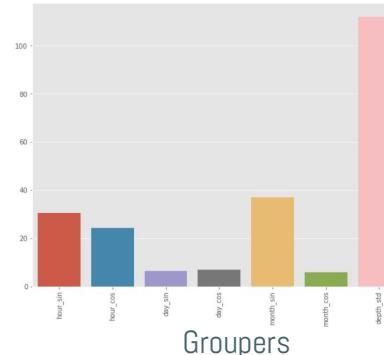
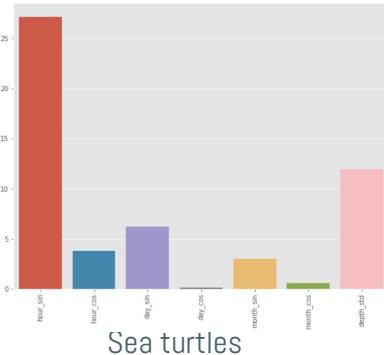
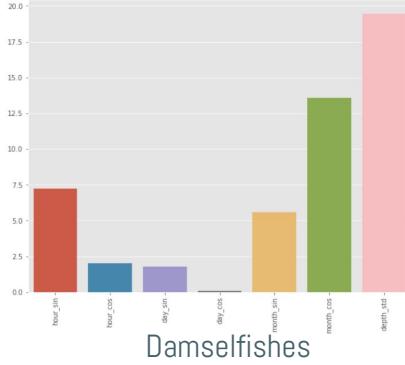
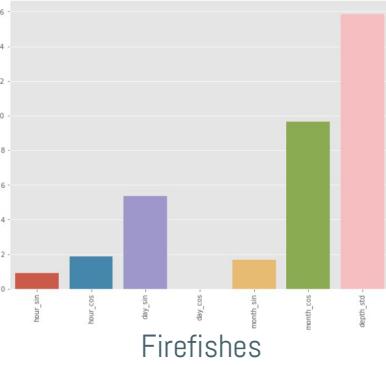
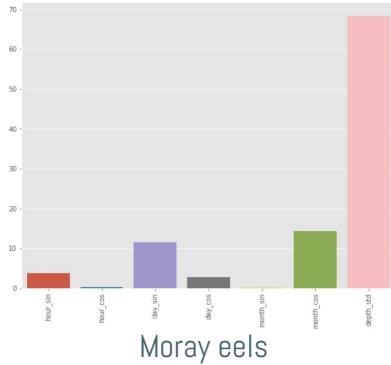
Variables categóricas



Importancia de las variables

Variables numéricas

Test Anova



Selección del modelo



Meta-estimadores



Estimadores



Métricas



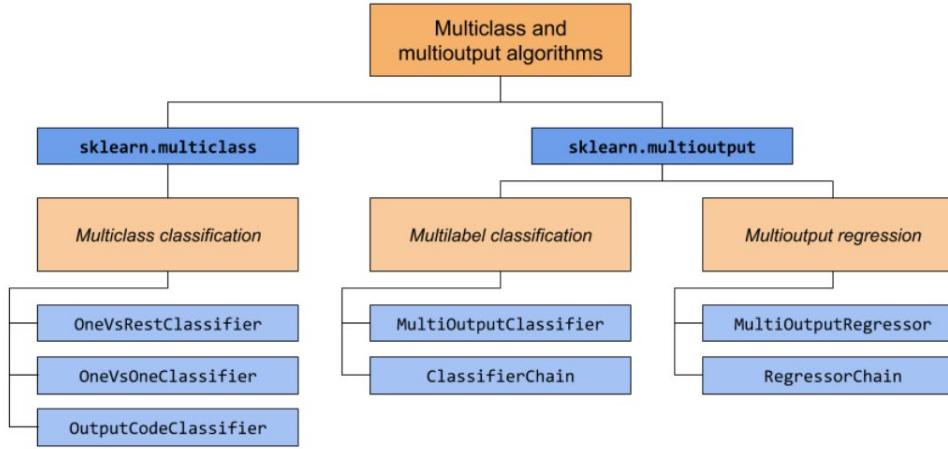
Selección de
hiperparámetros



Selección del
modelo final

Selección del modelo

Meta-estimadores



MultiOutputClassifier

Crea un estimador independiente por cada una de las columnas del target

ClassifierChain

Crea un estimador por cada columna del target teniendo en cuenta las etiquetas reales del resto de columnas

Selección del modelo

Estimadores



Logistic Regression

Va a ser el modelo de referencia. Permite ver fácilmente las relaciones entre variables y target



Support Vector Machine

Se ha utilizado en la predicción de migración de especies



XGBoost

Uno de los modelos más populares de los últimos años



Selección del modelo

Métricas

Hamming Loss

Calcula la media de la distancia de Hamming entre los valores reales y los predichos

Micro Average F1 Score

Calcula el F1 Score de la misma forma que si fuera una clasificación binaria

Micro Average Precision

Calcula Precision de la misma forma que si fuera una clasificación binaria

Micro Average Recall

Calcula Recall de la misma forma que si fuera una clasificación binaria

Accuracy score

Calcula el ratio entre etiquetas correctamente predichas y las etiquetas reales



Selección del modelo

Selección de hiperparámetros

ColumnTransformer

Sirve para asignar a cada columna la transformación necesaria

Pipeline

Contiene el ColumnTransformer, el meta-estimador y el estimador

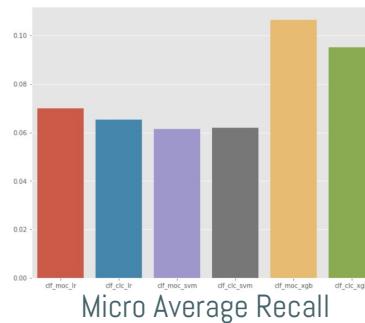
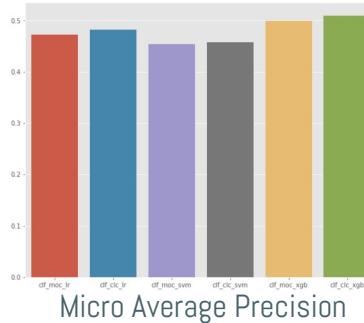
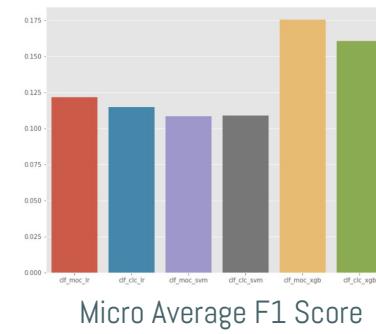
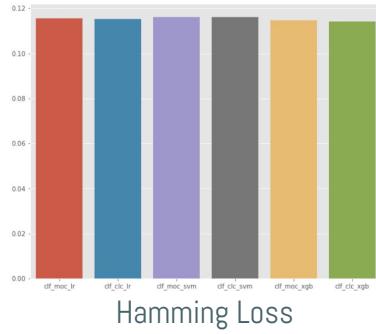
GridsearchCV

Contiene la búsqueda de los mejores hiperparámetros para entrenar el Pipeline



Selección del modelo

Selección del modelo final



Selección del modelo

Selección del modelo final

Classifier Chain XGBoost

```
{'clc_base_estimator_colsample_bytree': 0.3,  
 'clc_base_estimator_eta': 0.05,  
 'clc_base_estimator_gamma': 0.2,  
 'clc_base_estimator_max_depth': 10,  
 'clc_base_estimator_min_child_weight': 1,  
 'clc_order': None,  
 'clc_random_state': 17}
```

Demo

<http://localhost:8501>

