Desarrollo General del Trabajo

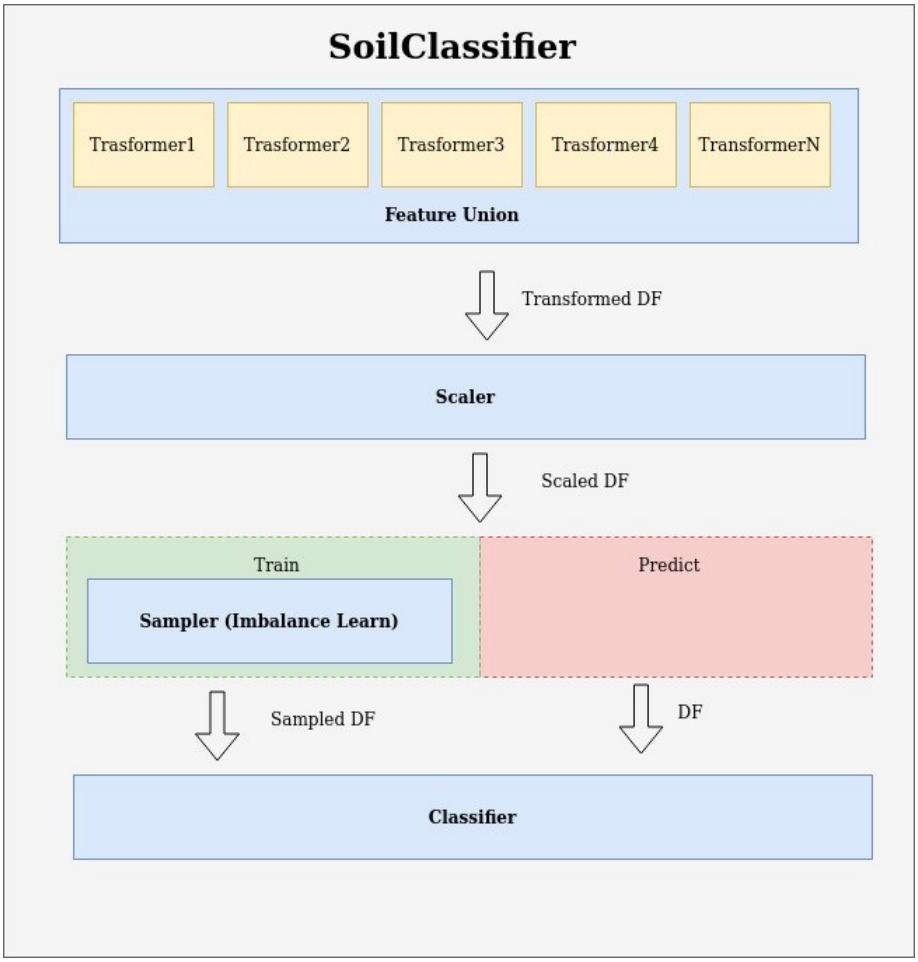
Para el desarrollo se utilizó el template Cookiecutter Data Science que brinda una estructura de archivos y carpetas con buenas prácticas para un proyecto de Data Science.

Todo el proyecto se encuentra disponible en Github

El proyecto incluye un fichero README.md que explica la estructura interna del mismo y los comandos *make* que se deben ejecutar para entrenar y generar las nuevas predicciones.

Modelo

En el caso del modelo, se utilizo un Pipeline de Scikit-Learn que tiene la siguiente estructura



Features

Para la manipulación de las variables se optó por un enfoque basado en la librería Scikit-Learn. Dicho enfoque consta de diversos transformadores que reciben un dataset y solo devuelven la feature deseada.

Luego mediante el transformador FeatureUnion se obtiene un dataframe uniendo la salida de todos los transformadores.

Scaler

El scaler permite escalar las variables antes de entrenar y/o predecir el modelo. En este caso se utilizó un MinMaxScaler, ya que este se puede aplicar a todas las variables, incluso a aquellas que se encuentran codificadas mediante OneHotEncoding.

Imbalance Learning

En este problema en particular las clases a predecir se encuentran muy desbalanceadas, por lo que a la hora de entrenar se utiliza la librería Imbalanced-Learn que permite utilizar solo algunos ejemplos de la clase mayoritaria y generar nuevos registros para las clases minoritarias. Este transformador solo se aplica a la hora de entrenar el modelo y no así cuando éste predice.

Classifier

El paso final del Pipeline es el clasificador. Para este modelo en particular se utilizo Gradient Boosting Classifier, es un modelo muy potente de clasificación que permite encontrar relaciones entre las variable que otros modelos no. A su vez se probó con otros clasificadores de diferentes naturalezas y este fue que mejor métricas dio en los datos de validación.

Análisis Exploratorio

El análisis exploratorio se llevo acabo en el fichero / notebooks/01_First_Exploratory.ipynb .

En dicho análisis, se analizaron a nivel general todas las variables, pero solo se profundizó en aquellas que eran interpretables, es decir, se excluyeron las variables de imagen y las geométricas. Las principales conclusiones obtenidas a partir de el fueron:

- Las clases están muy desbalanceadas. Se debe aplicar alguna técnica de *sampling* para evitar que el modelo sobre ajuste sobre los casos de la clase mayoritaria
- Existen valores nulos para algunas variables, los que serán tratados en los transformadores de dichas variables
- La mayoría de las variables pueden aportar información a la hora de predecir
- Latitud y Longitud son MUY importantes ya que la clase minoritaria se encuentra en las afueras del conglomerado.
- La variable de Área poseía muchos outliers, pero se vio que una transformación logarítmica sobre la misma generaba una distrubición mas normal.

- Las variables se encuentran en diferentes escalas, por lo cual se deben escalar para evitar problemas de convergencia y de inalterabilidad en un modelo

Elección y Tratamiento de Variables

Correlación de Variables

Se ejecutó un análisis de correlación de las features originales en los archivos notebooks/FeatureCorrlation.ipynb y notebooks/FeatureCorrlation2.ipynb donde se detecto que muchas de las variables originales de las imágenes, se encontraban altamente correlacionadas, lo cual no es bueno para nuestro modelo. Es por eso que se decidió quedarse con los cuantiles 1, 5 y 10 de las cuatro bandas.

Variables Ordinales

En el caso de la variable ordinal de información catastral, se aplico una transformación del tipo oneHotEncoding que tiene las siguientes características

Clases = ["A", "B", "C"]
Order = A > B > C

Traditional One Hot Encoding:
A = [1, 0, 0]
B = [0, 1, 0]
C = [0, 0, 1]

One Hot Ordinal Encoding:
A = [1, 1, 1]
B = [0, 1, 1]
C = [0, 0, 1]

Esto se vio que tuvo mejores resultados que el oneHotEncoding tradicional, ya que este no permite captar la información del orden de las categorías; y que haciendo de esta variable, una numérica.

Variables Cuadráticas

Se agregaron las variables cuadráticas para las variables de ubicación y geométricas. En el caso de las primeras daban alta correlación y su eliminación no afecto a la performance del modelo. Mientras que las ultimas si mejoraban el clasificador obtenido.

Productos EO Sentinel (Índices)

Investigando se encontró la existencia de algunos índices que permiten identificar zonas con vegetación y otras características de las superficies mediante la combinación de algunas de las bandas que brinda el satélite.

Eliminación Recursiva de Variables

También se hizo el análisis de eliminación recursiva de variables que permitió descartar algunas variables que no ayudaban al modelo.