CATS
Clasificador Automático Taxonómico de Suelos

Soil Taxonomic Classification Automation to drive Agriculture development.

Team 19: Juan Sebastián Serrano, Jaime Muñoz, Diana Vélez, Francisco Javier Lara, José Alejandro Montaña, Endhwyr Salas.

Highlights

•It is important to know and understand soils features to make better policies and take the best decisions to protect and improve soil's fertility, mitigate erosion and use them properly.

•This app was designed to be user friendly and to make the taxonomic classification process easier for the Agustin Codazzi Institute and the edaphologists.

Background

Paris agreement predictions have been exceeded by recent measures, indicating that extreme weather will impact not only people but also crops. To avoid famine, we must focus on soil, the third most important natural resource and it's important to optimize food production and fight hunger.

Objective

Our project's main goal is to support IGAC's mission by fitting a machine learning model that performs the soil classification, using the five most common taxonomic orders, making the process simpler and automatic.

Data

Data was provided from IGAC and it was gathered manually at the field. We found multiple null values and some variables are prone to subjectivity.

EDA allowed us to understand the data and find some important issues that had be taken into account when cleaning the data.

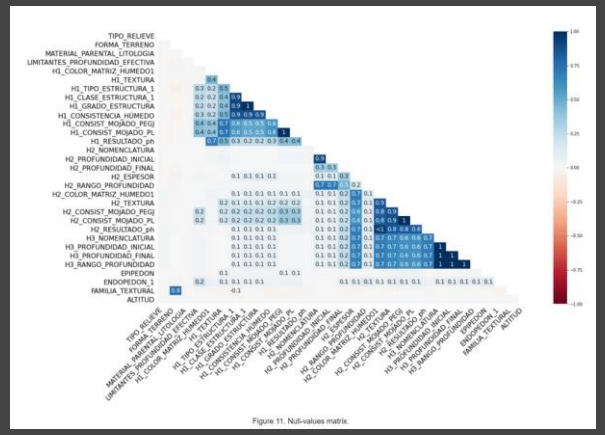


Figure 11. Null-values matrix.

Exploratory Data Analysis

API

It contains a method that allows to classify up to 4,000 elements in less that One second.

Model parameters

```
RandomForestClassifier(class_weight={ 0: 0.5020521353300056,
                                      1: 3.2270944741532976,
                                      2: 17.241904761904763,
                                      3: 0.4429655003670174,
                                      4: 2.6124098124098123 },
    criterion='entropy', max_depth=31, max_features='lo2',
    n_estimators=150, n_jobs= -1, oob_score=True,
    random_state=123)
```

Confusion Matrix

```
[[889   2   0  11   0]
 [  0  65   0  74   1]
 [  0   0  25   1   0]
 [  2  22   0 995   4]
 [  0   0   0  14 159]]
```

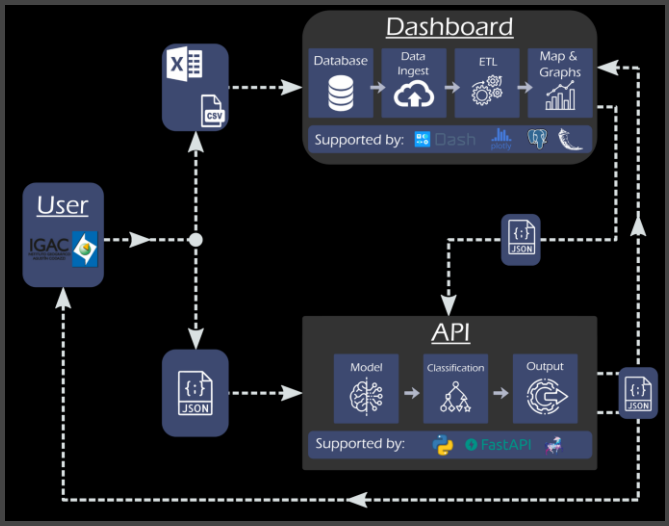
El accuracy de test es: 94.21378091872792 %

Model

After performing several specialized classification models, including multinomial regression, stacked models and random forest, we chose the random forest, after testing more than 70 trees, with different parameters, considering out of bag classification error as a selection measure, which is a more computationally efficient method than cross-validation method. This model generated an accuracy of 94% which gives us the confidence of obtaining high accuracy in future classifications.

Regarding multinomial regression, Was not chosen since influential points were detected by Cook's distances.

Backend



Frontend