ML APPROACH FOR COLOMBIAN SOIL ANALYSIS

Santiago Rincón, Cristina Gomez Patiño, Daniella Bernal Delgado, Islandi Navarro, María Isabel Montoya González, Carlos Andres Cardenas Perez & Santiago Henao

Highlights

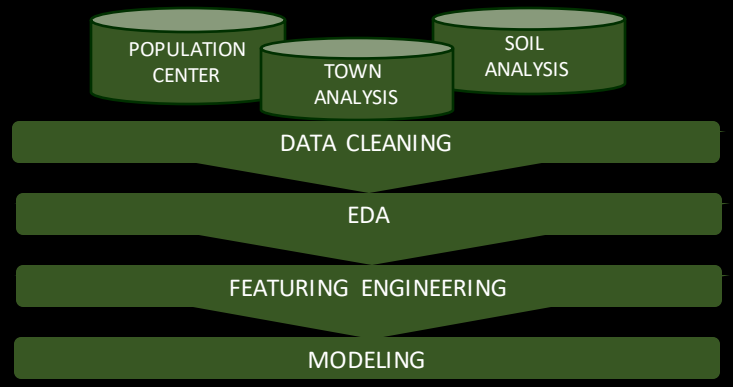
- The **KPCA** helps us to reduce the dimension of the table, and we can identify the outliers better for each region, by creating 2 artificial variables.
- The **tool** can predict with up to 84% accuracy which crops could be the best for a soil with specific conditions.
- The **model** could be overfitted, but the prediction results are not far away from the train results.

Background

Farmers in Colombia sometimes lose large harvests due to climatic factors, demand for their products, competition from them and poor soil management practices, which impacts their **family economies**. Our objective is to solve this social problem by providing a simple, practical and efficient tool.

Data

The Dataset under study is part of the Open Data portal of the Colombian State, developed by AGROSAVIA. It includes data of:

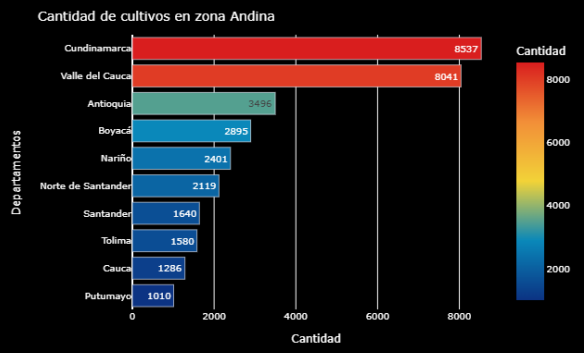


Recommendation Model

XGBoost offers several advanced features for model tuning, computing environments and algorithm enhancement. It's capable of performing the three main forms of gradient boosting (Gradient Boosting (GB), Stochastic GB and Regularized GB) and it's robust enough to support fine tuning and addition of regularization parameters.

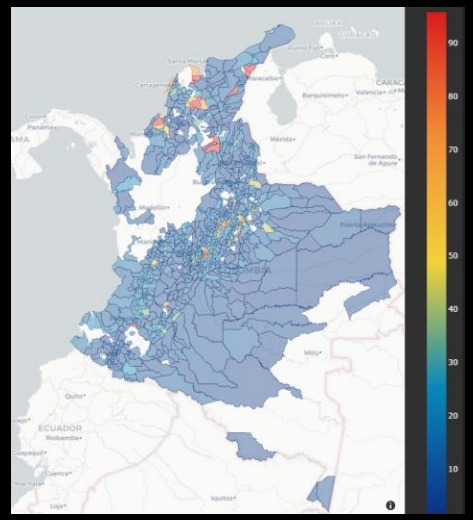
Outliers Detection

The core of the **Isolation Forest** algorithm is to "isolate" outliers by creating decision trees over random attributes. The random partitioning produces noticeable shorter paths for outliers since: - fewer instances (of outliers) result in smaller partitions - distinguishable attribute values are more likely to be separated in early partitioning



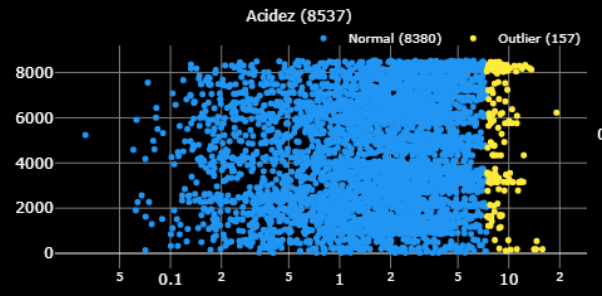
Statistics

We added part of the EDA findings in the DashBoard at Statistics section, like the number of crops in a selected Zone because people could find out which crops are in their surrounding whether for competitive or commercial reasons.



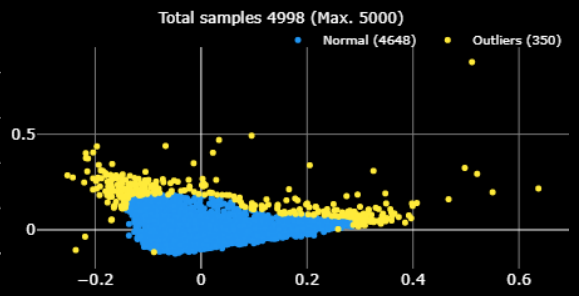
Analysis of the variable selected in the toolbar throughout Colombia.

For **Agrosavia laboratory operation**, the detection of anomalies is key due to the number of samples that arrive at the institution. Currently the laboratory does it manually, but if this tool goes into production, it could assist in the verification of hundreds of samples, which significantly reduces time.



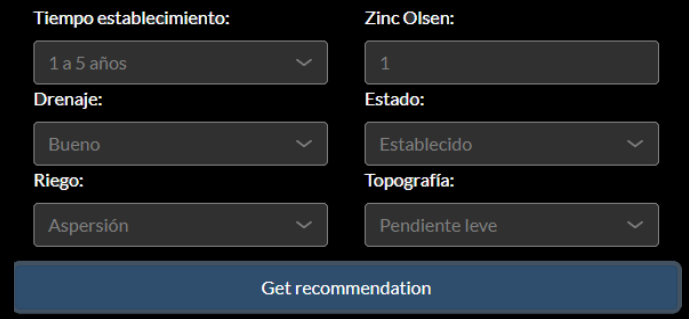
Univariate Outliers

The outlier detection analysis is performed with each of the numerical variables using the Inter Quartile Range method.



Multivariate Outliers

By looking at the scatter plot of the projected data, we can distinguish the different groups, normal and abnormal, within the original data. We propose to use visualization given by KPCA in order to clearly identify the outliers.



Given the physicochemical variables, the model will recommend which are the top 5 type of crop adequate for these characteristics

Top 5 recommended crops:

1. Pastos
2. Aguacate
3. Palma de aceite
4. Citricos
5. Caña de azucar