MACHINE LEARNING APROACH FOR COLOMBIAN SOIL ANALYSIS

Final Project Report

Colombia | Team #50

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

AGROSAVIA
Corporación colombiana de investigación agropecuaria

correlation one

CONTENT

PROJECT DESCRIPTION

Overview

Agricultural soil testing is a complex technique that combines several analytical methods with their respective extractions, basically removing the most important nutrients from the soil to measure their availability for the plant. Based on the information obtained by soil testing, it is possible to carry out important studies of different variables.

In Colombia there are around 22 million **arable lands (hectares)**, not all of which make good use of the soil according to its maximum agricultural potential. AGROSAVIA carried out a **soil fertility testing**, which we want to use to carry out multiple analyzes on the use of the soil, suitable crops for seasons and for all the farmable land of the country.

Business Problem

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**.

They generally choose the crop to establish based on tradition or experiences with previous crops, without considering height, precipitation, temperature, or physical, chemical, and biological characteristics of the soil, which could end up in inadequate use of the soil, erosion of the same and not taking advantage of its vocation.

Motivation

Currently, the price of agricultural fertilizers has reached historical highs due to the war between the two nations, Russia, and Ukraine. For this reason, it is crucial to provide a tool that helps farmers evaluate their soils and choose the best product to cultivate with a productivity analysis and future demand.

We consider this topic of great importance, since we are aware of the vulnerability of the rural population and farmers, in the sense that there is a possibility that their crops will not produce the expected harvests, and therefore it will not be possible to sell them. This population cannot afford to lose a harvest, since entire families and their economic stability depend on it.

According to the Law Department of the Externado University, Colombia has a percentage of 54.2% of food insecurity, this means that each one of two households has insufficiencies in relation to Food and Nutritional Security (SAN).

For all these reasons, and in order to support our farmers, we want to provide them with a tool that allows them to identify different soil elements and choose the best product to grow, with an analysis of productivity, future demand and better use of the soil.

Datasets

The Dataset under study is part of the Open Data portal of the Colombian State. Which refers to: "The Results of the Soil Testing Service that the Laboratory of Soil Chemistry and Physics of AGROSAVIA, provides to the agricultural sector, focuses on the EVALUATION of Soil Fertility, Salinity and parameters that work as a tool for carrying out fertilization plans, application of amendments and adaptation of the land to achieve profitable productions". The dataset variables and their respective description are

detailed below in the Data dictionary section.

DATA DICTIONARY

This data dictionary supports different tables from several data sources such as Open data. The Soil Analysis and Stations catalog are linked as both hold Department Columns.

Table: Soil_analysis

Column	Data Type	Description
Row number	integer	Number uniquely identifying the record. It is completely empty.
Department	text	Department the sample is taken from
Municipality	text	Name of the city
Crop	text	Crop established or to be established type on the land
Status	text (Cat)	Status of the crop: ● To be established ● Established
Establishment Time	text (Cat)	Elapsed time since the crop was established. ● 0 to 1 year ● 1 to 5 years ● 5 to 10 years ● More than 10 years
Topography	text (Cat)	Descriptive topographic characterization of the land. ● slightly wavy ● moderately wavy ● Curly ● Wavy and Sloping ● Slope ● steep slope ● slight slope ● moderate slope ● Flat ● Flat and wavy ● Flat and slope
Drainage	Text (Cat)	Drainage classification in the studied area. Reported on a subjective scale. ● "Good drainage", "Good", "Very good drainage" ● "Regular", "Regular drainage" ● "Bad drainage", "Bad"
Irrigation	Text (Cat)	Irrigation type used: ● Aspersion ● Canyon ● Drip ● Gravity ● Hose ● Flood ● Microsprinkler

Column	Data Type	Description
		• Without irrigation
Applied Fertilizers	text	Fertilizers used in the crop
Analysis Date	text	Date for soil analysis. It is completely empty.
pH water: soil 2.5:1.0	text	Soil pH level in relation to water pH level. It has inexact values: "<3.80" and "<10.0"
Organic Matter (OM) %	text	Organic material percentage present in the soil sample (string, numerical)
Phosphorus (P) Bray II mg/kg	text	Amount of Phosphorus found in the soil sample (mg/Kg). Reported according to the Bray II method. It has inexact values: "<3.87"
Sulfur (S) Monocalcium phosphate mg/kg	text	Amount of Sulfur found in the soil sample (mg/Kg). It has inexact values: "<0.01"
Acidity (Al+H) KCL cmol(+)/kg	text	Acidity level of the soil sample.
Aluminum (Al) exchangeable cmol(+)/kg	text	Amount of interchangeable aluminum in the soil sample (cmol(+)/Kg)
Calcium (Ca) exchangeable cmol(+)/kg	text	Amount of interchangeable Calcium in the soil sample (cmol(+)/Kg). It has has inexact values: "<0.59" and "<0.55"
Exchangeable Magnesium (Mg) cmol(+)/kg	text	Amount of interchangeable Magnesium in the soil sample (cmol(+)/Kg). It has inexact values: "<0.20"
Exchangeable Potassium (K) cmol(+)/kg	text	Amount of interchangeable Potassium in the soil sample (cmol(+)/Kg). It has has inexact values: "<0.09" and "<0.06"
Exchangeable Sodium cmol(+)/kg	text	Amount of interchangeable sodium in the soil sample (cmol(+)/Kg). It has inexact values: "<0.14"
Cation exchange capacity (CICE) sum of bases cmol(+)/kg	Text	Cation exchange capacity (CICE) Sum of bases (cmol(+)/kg)
Electrical conductivity (EC) ratio 2.5:1.0 dS/m	text	Electrical conductivity of the soil sample. It has inexact values: ">12.88" and "<0.01".
Iron (Fe) available olsen mg/kg	text	Quantity of available Iron. Olsen Method (mg/Kg). It has inexact values: "<1.00"
Copper (Cu) available mg/kg	text	Quantity of available Copper (mg/Kg). It has inexact values: "<1.00"

Column	Data Type	Description
Manganese (Mn) available Olsen mg/kg	text	Quantity of available Manganese. Olsen Method (mg/Kg). It has inexact values: "<1.00"
Zinc (Zn) available Olsen mg/kg	text	Quantity of available Zinc. Olsen Method (mg/Kg). It has inexact values: "<1.00"
Boron (B) available mg/kg	text	Quantity of available boron. Double acid extraction method (mg/Kg)
Iron (Fe) available double acid mg/kg	text	Quantity of Iron available. Double acid extraction method (mg/Kg). It has inexact values: "<0.40"
Copper (Cu) available double acid mg/kg	text	Quantity of available copper. Double acid extraction method (mg/Kg). It has inexact values: "<0.40"
Manganese (Mn) available double acid mg/kg	text	Quantity of available Manganese. Double acid extraction method (mg/Kg). It has inaccurate values: "<0.40"
Zinc (Zn) available double acid mg/kg	text	Zinc quantity available. Double acid extraction method (mg/Kg). Has inexact values: "<0.40"
Sequential	Integer	Soil Test Sequential ID. It is completely empty

Table: Town Analysis

Column	Data Type	Description
DEP_ID	Text	Department Id
Departamento	Text	Department Name
MUN_ID	Text	Town Id
Municipio	Text	Town Name
TYPE	Text	Characterization of the town

Table: population center

Column	Data Type	Description
COD_DPTO	Text	Department Id
NOM_DPTO	Text	Department Name
COD_MPIO	Text	Town Id
NOM_MPIO	Text	Town Name
COD_CPOB	Text	population center Id
NOM_CPOB	Text	population center Name
TIPO	Text	Characterization of the town
LATITUD	float	LATITUDE
LONGITUD	float	LENGTH

EXPLORATORY DATA ANALYSIS

Data cleaning.

Before starting with the exploratory analysis of the data, it was necessary to clean them, in order to optimize the process and get more accurate information. The processes carried out for this cleaning are detailed below:

- **Row number** : A consecutive ID is added to each row.
- **Department, Municipality, Crop, Applied Fertilizers**: Special characters are replaced using the slugify package. Finally, by using the LOWER function all characters are converted to lowercase.
- **Status, Establishment Time, Topography, Irrigation:** Data type is changed to categorical value, and by using the LOWER function all characters are converted to lowercase.
- **Drainage:** By using the LOWER function all characters are converted to lowercase. Values “error na” were replaced as “no indica”. A dictionary was created in order to have a Likert scale.
- **Analysis Date, Sequential**: As all the values of the column are null, the function of DROP is applied to the column.
- **All physical/chemical variables (**pH water:soil 2.5:1.0, Organic Matter (OM) %, Phosphorus (P) Bray II mg/kg, Sulfur (S) Monocalcium phosphate mg/kg, Acidity (Al+H) KCL cmol(+)/kg, Aluminum (Al) exchangeable cmol(+)/kg, Calcium (Ca) exchangeable cmol(+)/kg, Exchangeable Magnesium (Mg) cmol(+)/kg, Exchangeable Potassium (K) cmol(+)/kg, Exchangeable Sodium (Na) cmol(+)/kg, Cation exchange capacity (CICE) sum of bases cmol(+)/kg, Exchangeable Sodium (Na) cmol(+)/kg, Iron (Fe) available olsen mg/kg, Zinc (Zn) available double acid mg/kg, Copper (Cu) available double acid mg/kg): Data type is changed to float. By using the REPLACE function several characters were replaced and the name of each variable was updated, so that they were easier to use and understand.

After getting a more optimal dataset to work with, different visualization techniques are used, which facilitate the exploratory analysis of the data. These techniques are detailed below:

Analysis by departments.

By doing an analysis by department we can see that Cundinamarca, Valle del Cauca and Antioquia have the highest number of crops tested, and Choco, Amazonas and Guania have the lowest number of analyzes carried out. In addition to this, Vaupes and San Andres do not have any records in this analysis. This result is shown in the following image:

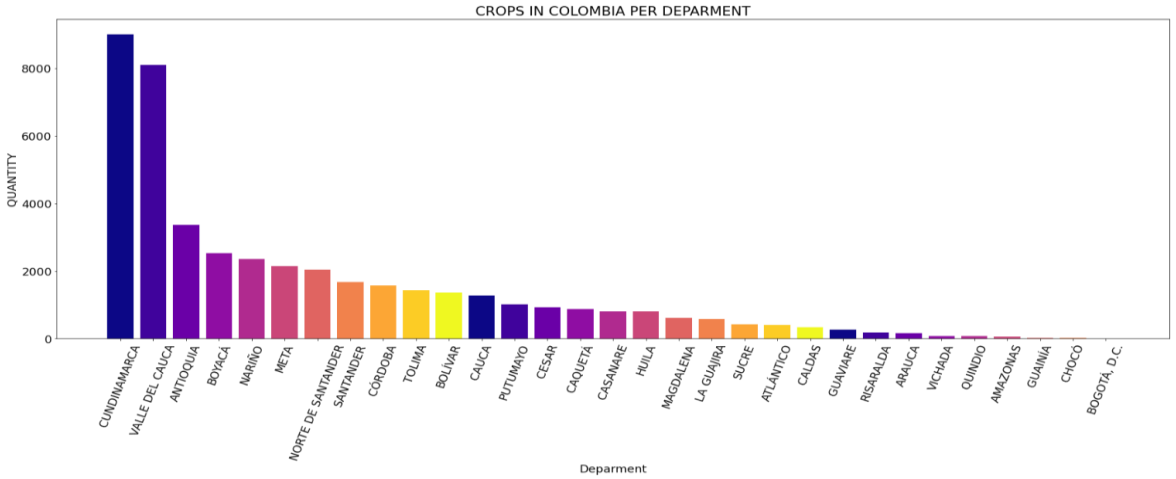


Image 1. number of crops per department.

Analysis by crops type

From *Image 2*, we can observe that “cacao” is the crop with the most entries in our database. However, it is important to recall that this distribution only applies for our project, and it may have a bias for the number of samples taken according to crop locations. The second one is “pastos” and that's something that we can imagine before the analysis, because Colombia is a cattle-raising country. The third one “aguacate” and that place has a relationship with the growth of exportations that Colombia have had among the last few years. The fourth one is “caña de azucar” which grows in practically all tropical and subtropical regions of the world. In Colombia, it is grown productively from sea level to altitudes above 2,000 meters in the most varied conditions of temperature, luminosity, rainfall, and soil quality. The fifth one is “café” that is our flagship crop, and we could guess that it must be at the top of the crops before the analysis. The remaining places are occupied by crops that are the basis of the Colombian diet and that were expected to be among the first, despite the possible bias of the data.

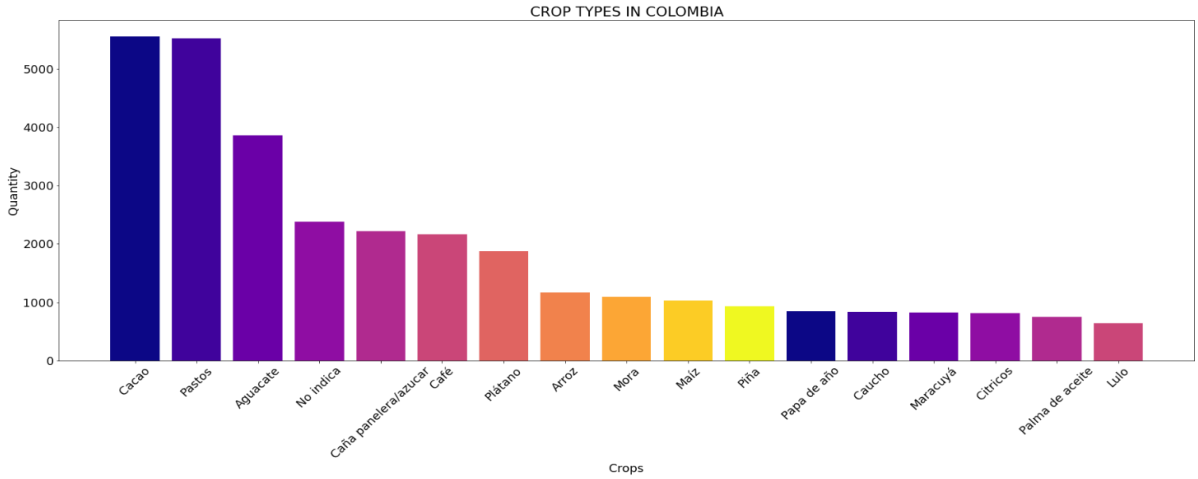


Image 2. Crop types in Colombia.

Analysis by Irrigation

When we see the types of irrigation in *Image 3*, most of the crops do not have irrigation systems and it is shown in the graph that the second position is “not shown”. There is a national plan of development in terms of agriculture to invest in irrigation systems in Colombia, as according to the government, in Colombia there are 18 millions of hectares with potential of irrigation, but only 1 million of hectares have irrigation system, so this data has a relation with the diagnosis of the development plan of the government, taking into account that the irrigation of crops is a factor of great importance in their productivity, for those that do not have an irrigation system, it is assumed that they depend entirely on the rain.

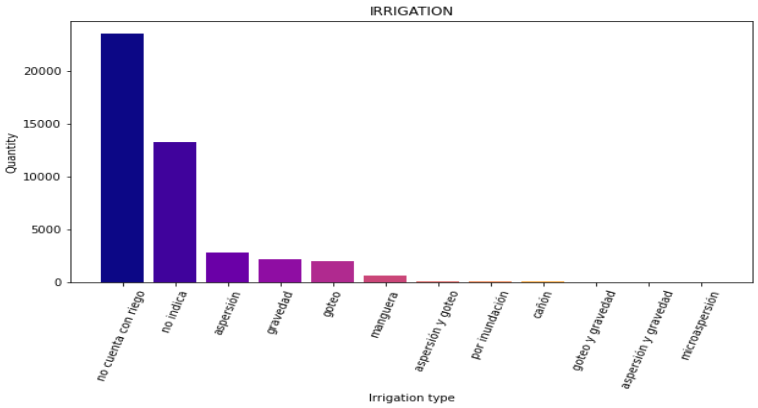


Image 3. Crop irrigation systems.

Analysis by drainage type

In the following graph (*Image 4*), it is possible to know that about 60% of the samples come from well-drained soils, which makes them suitable for the establishment of crops, while 28% of the samples correspond to soils with poor drainage. On the other hand, only 2% of the analyzed samples are considered to come from poorly drained soils. It is necessary to highlight that in 21% of the samples the type of soil drainage is not indicated.

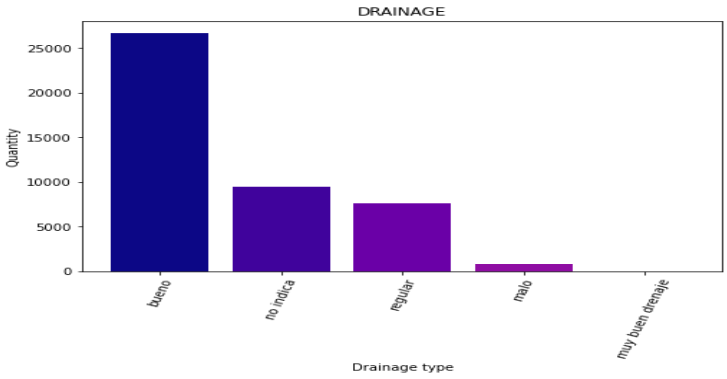


Image 4. Crop drainage.

Analysis by soil topography.

In the following graph it can be seen the top three are flat, undulating, and sloping topography, each followed by 34%, 23% and 21% of all samples. In total, these three topography types count for about 78% of all samples. This information is relevant as topography plays an important role in the distribution of vegetation, and when combined with aridity, it significantly distinguishes habitats and, therefore, influences the distribution, not only of species but also of plant communities. On the other hand, it can be noted that the combination of several topographic features is not very common, as they are the ones least representative on the graph.

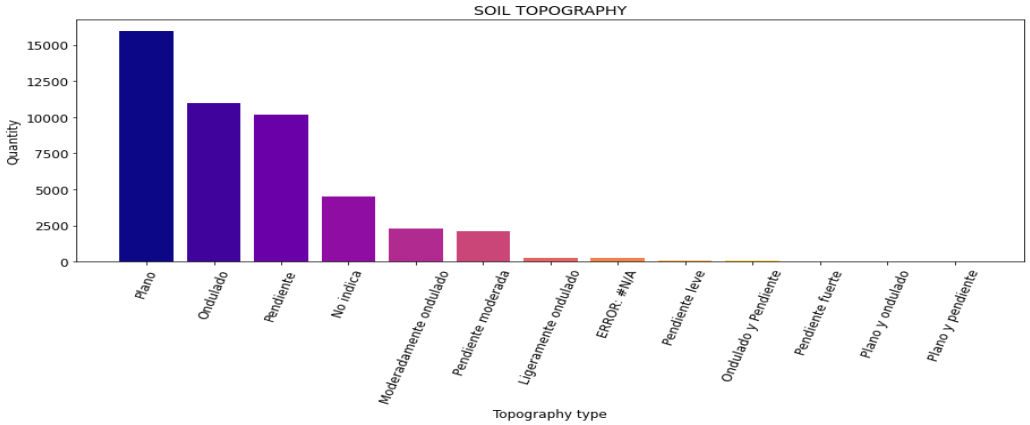


Image 5. Crop Topography.

Analysis by pH

The pH measures the degree of acidity of a soil, that is, the concentration of hydrogen ions (H+) that exist in the soil. On the scale of maximum value 14, the value of a neutral soil is 7, being acid all those that have values lower than 7, and basic all those that have values higher than this. Cultivated plants generally show their best development in values close to neutrality, since in these conditions the nutritional elements are more easily available and in a more adequate balance. If the soil is excessively

acidic, then hydrogen ions and aluminum abound in the soil exchange complex, preventing other necessary elements such as calcium, magnesium, sodium, or potassium from remaining in it, passing to the soluble fraction, and being easily eliminated with rainwater or irrigation. If the soil has less than 5.5 of the pH value, it would be convenient, in general, to raise it to a value close to 6/6.5 by adding a limestone amendment, so that the nutritive elements can be more easily available for the plants. If the soil is basic (for example in limestone soils) then the soil exchange complex is saturated and the excess of calcium in the medium prevents other elements, such as iron, from being absorbed by plants.

From the following table, a classification of the pH values presented in the tested crops was made, in order to facilitate our analysis.

pH values	Type	Observation
pH < 5.5	**very acidic**	**Difficulty in developing most crops, difficulty in retaining many nutrients**
5.5 <pH < 6.5	**acidic**	
6.5 <pH < 7.5	**neutral or close to neutrality**	**Optimal range for crops**
7.5 <pH < 8.5	**alkaline**	
pH > 8.5	**very alkaline**	**Difficulty in developing most crops, possible appearance of iron chlorosis**

In our analysis it is possible to see that most of the soil analyzed samples come from soils with acidic (32.5%) and very acidic (49 %) pH levels. On the other hand, the samples considered pH neutral represent only 13.4% of the analysis carried out. Finally, a small percentage corresponds to basic and very basic soils.

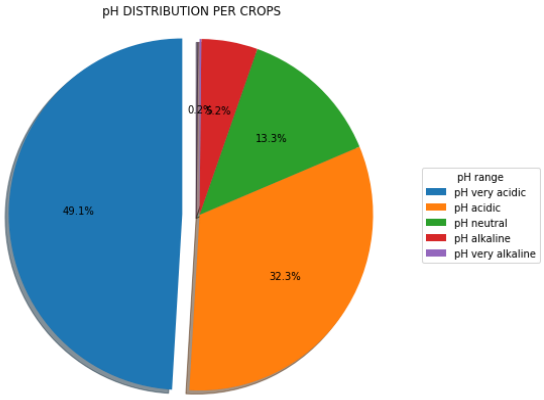


Image 6. Crop pH classification.

This graph is very important since that for example the Amazon region represents 41% of the territory in Colombia and the Amazon soil is very acidic; it has pH between 3.5 and 5.5. There are regions other than the Amazon that have similar soil acidity to the Amazon and are very rural states, for example Putumayo, Caquetá, Guaviare, Meta, and all southern and eastern Colombia, these regions definitely have many crops, that could explain that the distribution of the graph has 48% of the crops with very acidic PH

pH Analysis by type of crop

The *Image 7* shows the pH of the crops that are most present in Colombia, such as: the cultivation of avocado, cocoa, coffee, sugar cane, rice, etc. We can see that most of these crops have a pH in the range of 5.0 - 6.5. According to the table mentioned above, these crops have an acidic pH. It can also be noted that the corn crop tends to have a basic pH.

In addition, you can see that there are crops along the different pH types, most corn crops are around pH 5.5 and pH 8.0, which means that corn is very versatile in terms of pH of the soil.

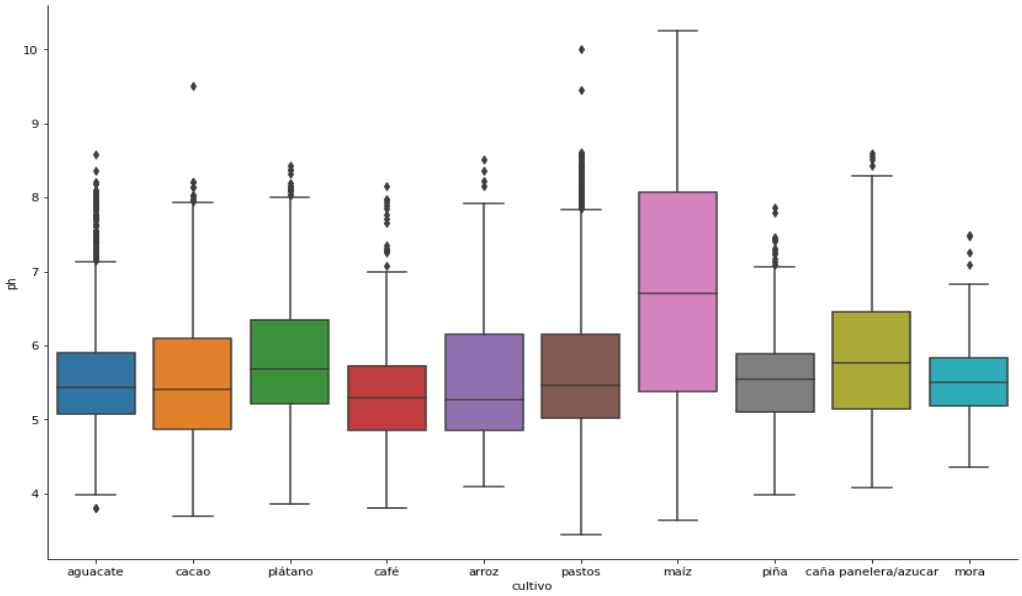


Image 7. Crop pH per crop type.

Analysis by numerical variables

In order to identify correlation between some numerical variables, such as nutrients, calcium, aluminum, phosphorus, etc., and properties such as pH, a pairplot graph was made, but due to the number of associated variables, this image is difficult to analyze, however, by means of it, it was possible to determine the relationship between some of these variables, which were later analyzed with the same type of graph, but in this case, it was done only for some types of crops of interest (types most analyzed crops), as shown in the *image 8*. The correlation between the aluminum and acidity, calcium and cice (cation exchange) variables, can be better observed in the graph below.

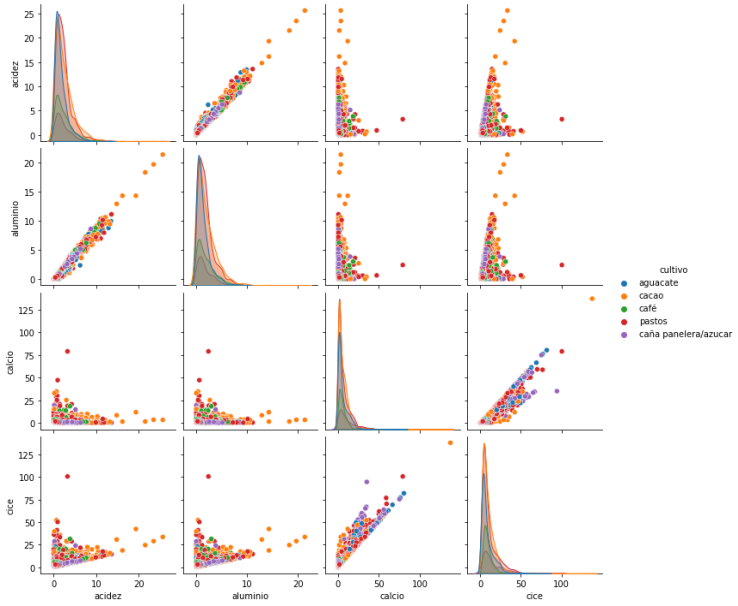


Image 8. Correlation between variables.

Image 9 shows the result of a Pearson correlation analysis conducted between numerical variables in the dataset where high values, either positive or negative, show a linear relationship, direct or inverse, respectively. The following table shows the interpretation of the values obtained.

Range of Correlation Coefficient Values	Level of Correlation	Range of Correlation Coefficient Values	Level of Correlation
0.80 to 1.00	Very Strong Positive	-1.00 to -0.80	Very Strong Negative
0.60 to 0.79	Strong Positive	-0.79 to -0.60	Strong Negative
0.40 to 0.59	Moderate Positive	-0.59 to -0.40	Moderate Negative
0.20 to 0.39	Weak Positive	-0.39 to -0.20	Weak Negative
0.00 to 0.19	Very Weak Positive	-0.19 to -0.01	Very Weak Negative

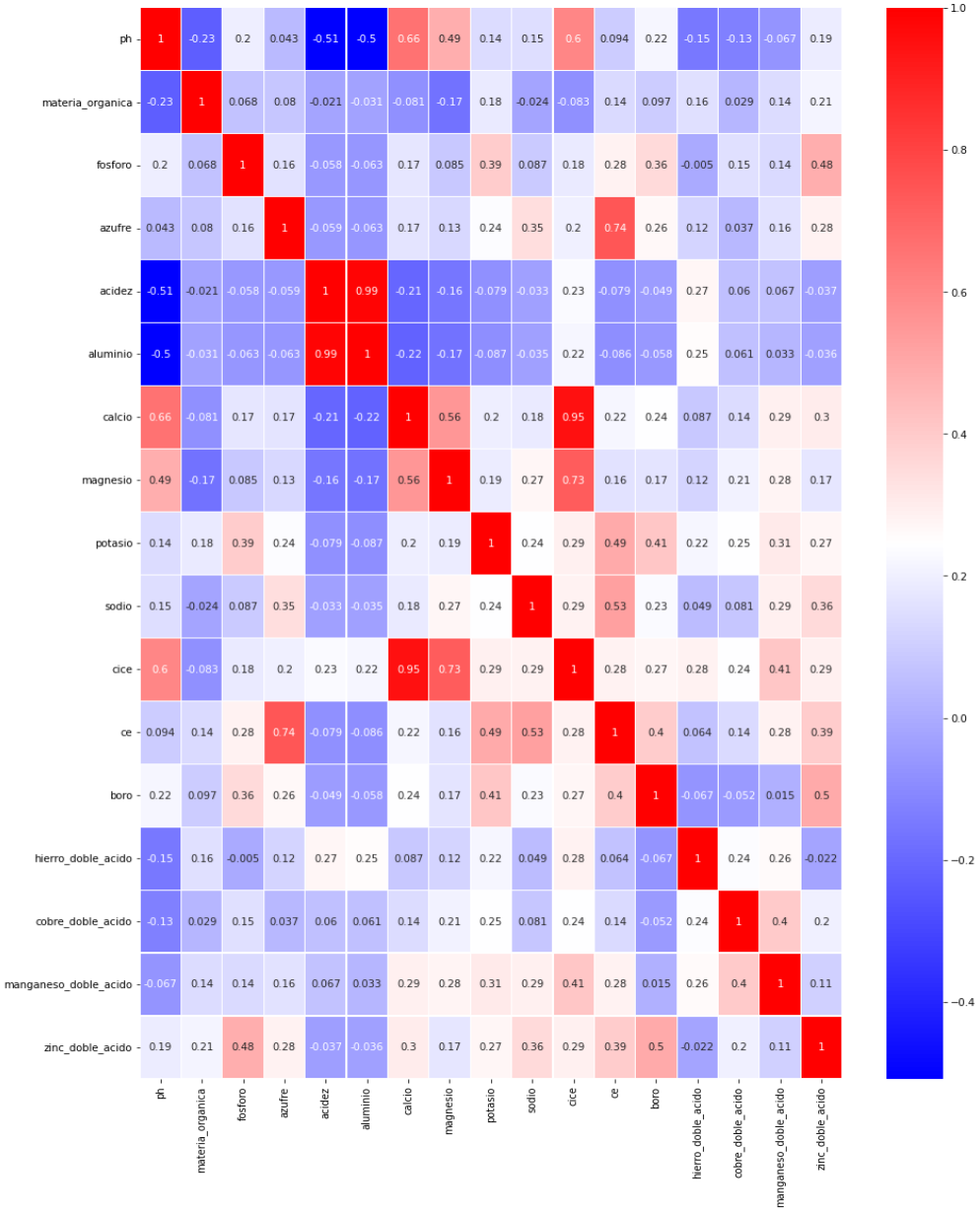


Image 9. Pearson correlation between variables.

DASHBOARD - COMPLETE FRONT END DESIGN

As a main objective, it is desired to predict the 5 top crops considering certain given characteristics of the soil; secondly, we want to visualize the statistical report made during the project and finally, estimate the number of outliers detected per region.

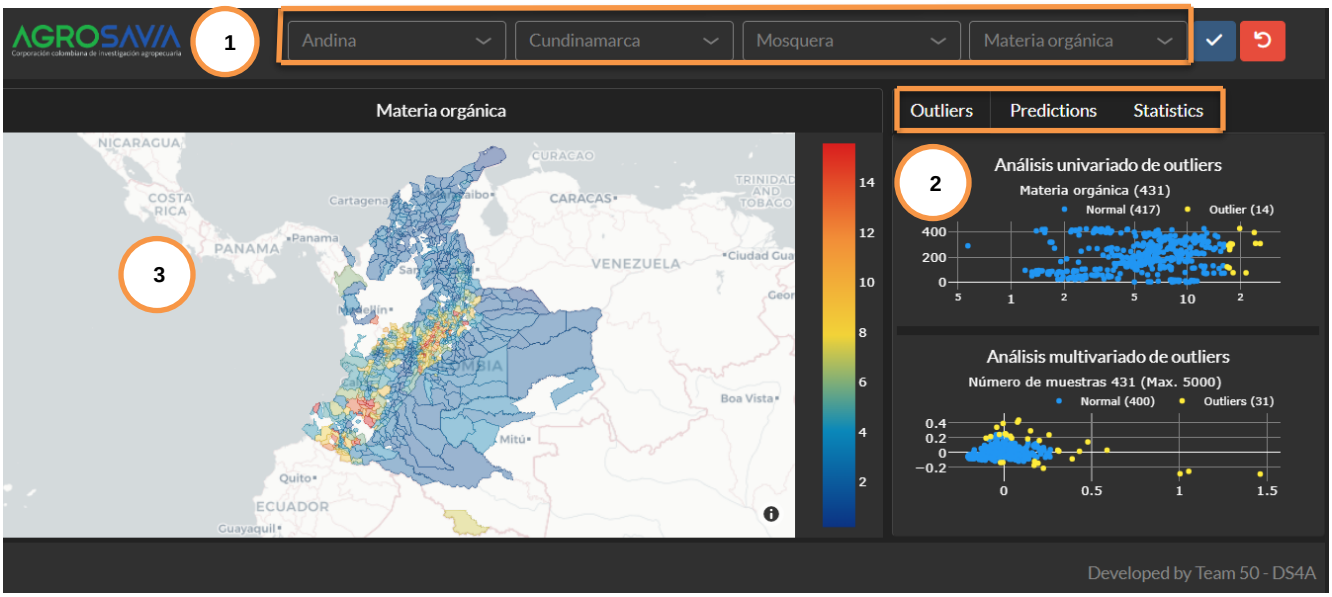


Image 10. Dashboard proposal.

1. Toolbar for selecting the region and the variable for which we want to analyze (statistical or outlier detection)

2. In the dashboard we can find three tabs: Outliers, Predictions and Statistics. Depending on the region selected in the toolbar, a detection of univariate and multivariate outliers, a statistical analysis and a crops recommendation model are made.

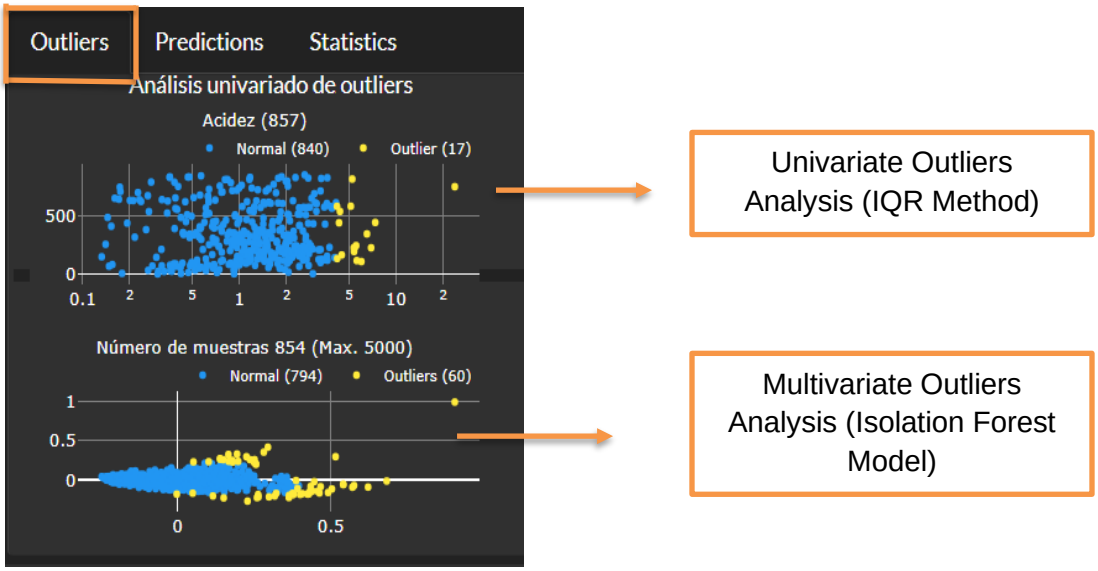
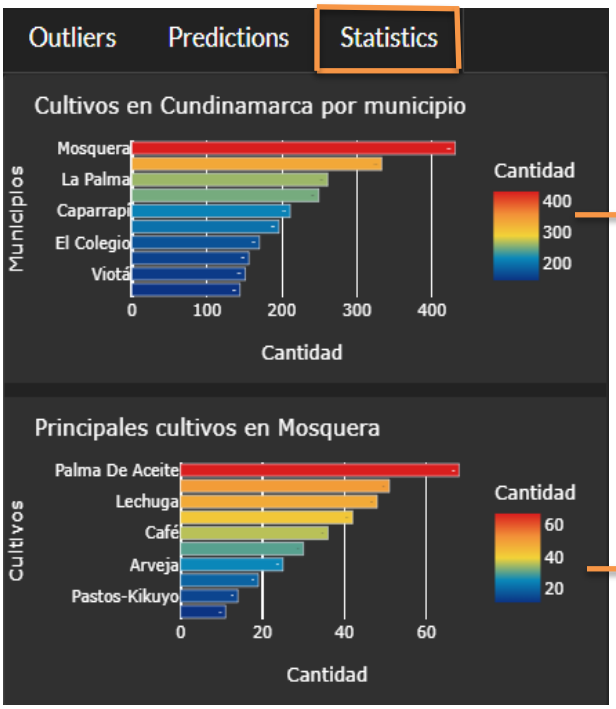


Image 11. Outliers' detection panel.



Number of crops in the selected department for the municipalities with more established crops in that department

Most frequent type of crops (according to our dataset) in the selected municipality

Image 12. Statistical analysis panel.

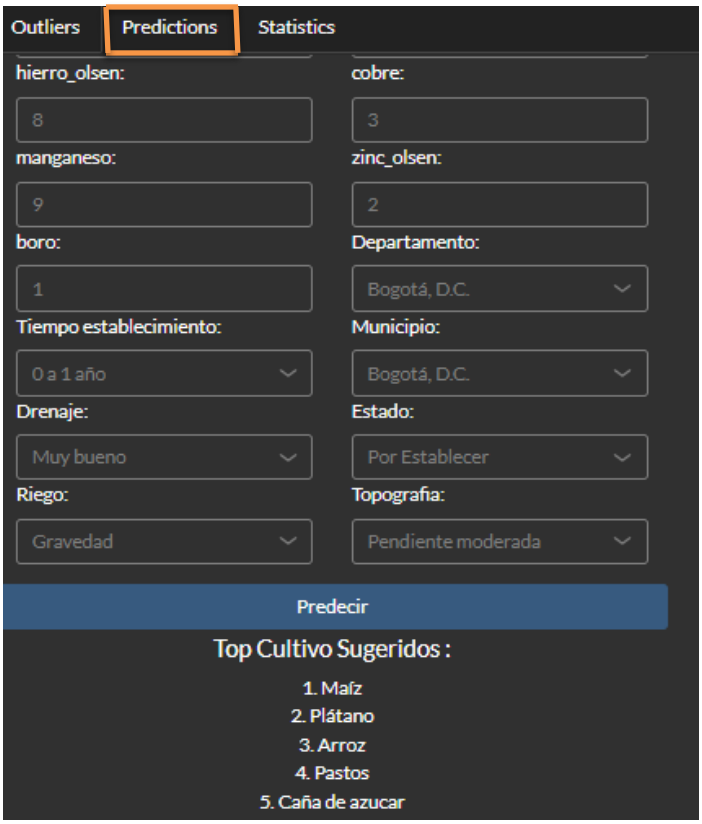


Image 13. Recommended crops prediction panel.

3. Analysis of the variable selected in the toolbar throughout Colombia: in the image below, we are analyzing the values for organic matter throughout Colombia and we can notice that most regions in Colombia have values between 2-4 for organic matter, other regions like el Tambo, San Agustin etc reach values up 16 for this variable.

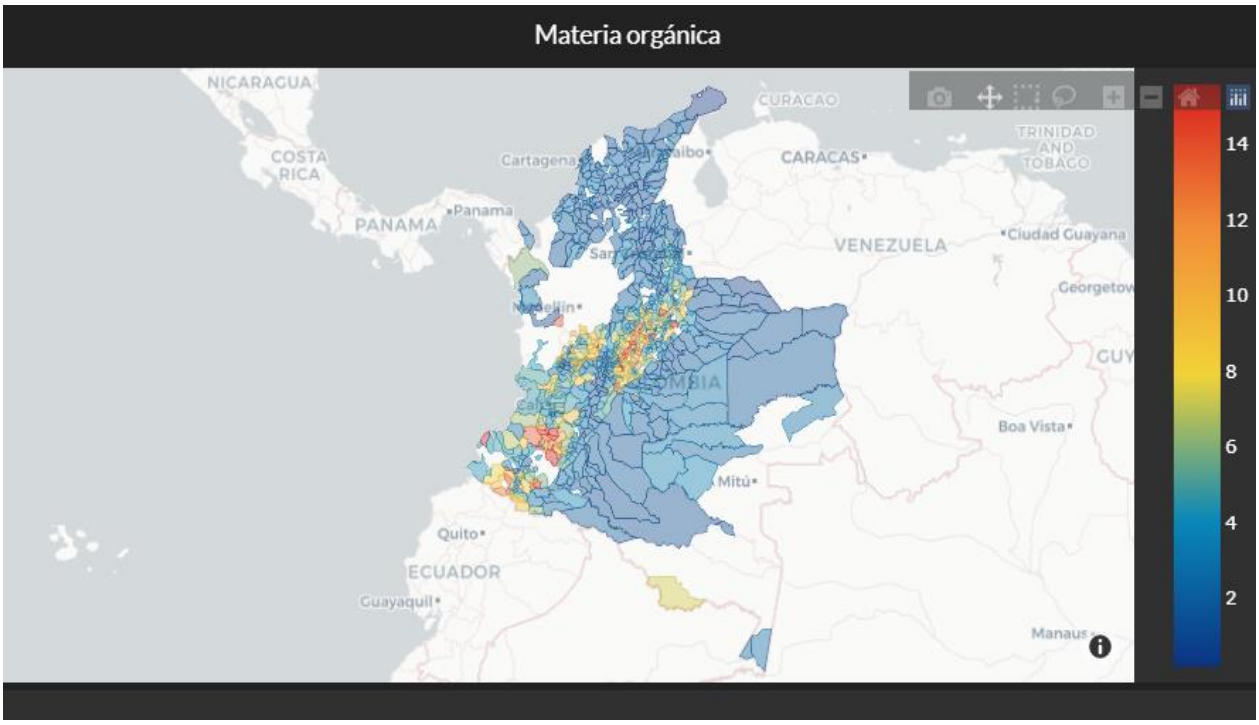


Image 14. Analysis of physical-chemical variable in Colombia

APLICATION MACHINE LEARNING MODELS

Outliers Detection Models

Outliers detection is a preprocessing technique which allows the identification of *strange* data points in a data set. After meeting with Agrosavia, a model for the detection of anomalies or atypical values in the soil analysis samples was determined to be of great importance to them.

At the suggestion of the Agrosavia company, the anomaly detection model by region was implemented, since the values for the physical variables can vary greatly depending on the region where the soil sample was taken. In this way, if we want to visualize the anomalies detected either by means of the univariate outliers detection model or by the models implemented for multivariate outliers detection, it is necessary to specify in which region (municipality or department) we want to detect outliers.

For this project, 3 techniques were implemented for the detection of anomalies:

- Univariate outliers detection- Inter Quartile Range: The outlier detection analysis is performed with each of the numerical variables using the Inter Quartile Range method. The possible variables to detect outliers are: pH, organic matter, phosphorus, calcium, magnesium, potassium, sodium, cice, etc.

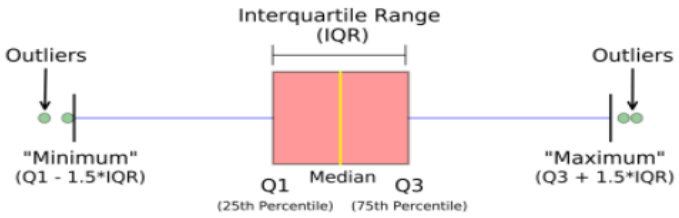
IQR (Inter Quartile Range) is a Probabilistic and Statistical Models which assumes specific distributions for data. Outliers are the points with low probability. This method is the most used and most trusted approach used in the research field to find outliers.

$$IQR = Quartile3 - Quartile1$$

To define the outlier base value is defined above and below datasets normal range namely Upper and Lower bounds, define the upper and the lower bound (1.5*IQR value is considered):

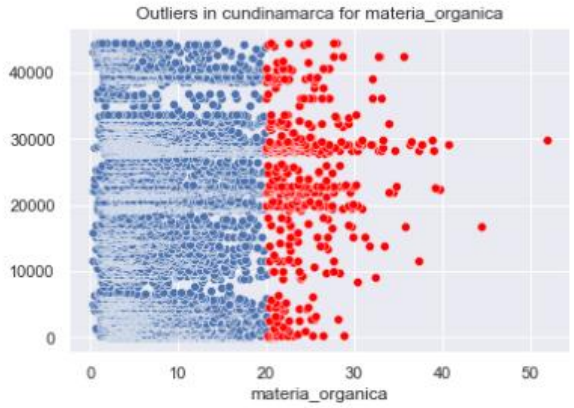
$$upper = Q3 + 1.5*IQR$$

$$lower = Q1 - 1.5*IQR$$



Having already implemented our outlier detection function for each variable, the result obtained for the organic matter variable is presented below:

```
De un total de 9011 de muestras en la region de cundinamarca, se detectaron 451 outliers
para la variable materia_organica
```




```
De un total de 820 de muestras en la region de cúcuta, se detectaron 11 outliers para la
variable materia_organica
```

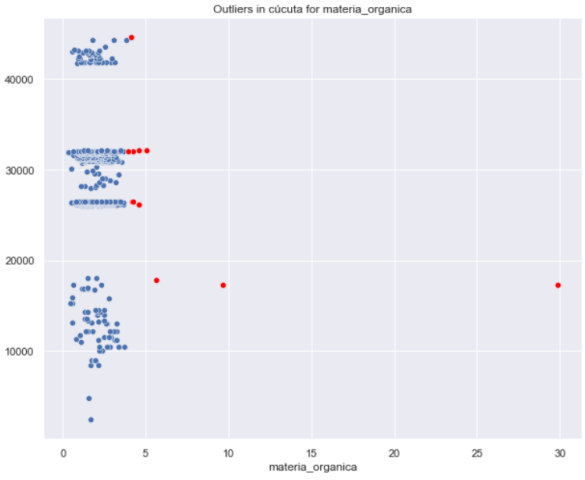


Image 16. Outliers / Organic_Matter

It should be noted that in the final dashboard the client (AGROSAVIA) will be able to choose not only the desired variable to display but also the region (per Department or Municipality). In order not to make the report very extensive, the analysis of outliers is presented only for the variable mentioned above.

- Multivariate Outliers Detection- Isolation Forest Model: this model explicitly isolates anomalies rather than profiling normal instances. Anomalies are usually 'few and different' instances in terms of the number and attribute value which make them more susceptible to isolation than normal points. So, we are going to take advantage of these two factors and will use a tree-based structure.

We ideally want to have an algorithm that would identify the outliers in a multidimensional space, so we are going to use an unsupervised anomaly detection algorithm that can detect outliers in a data set with incredible speed

The core of the 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

Hence, when a forest of random trees collectively produces shorter path lengths for some particular points, then they are highly likely to be outliers.

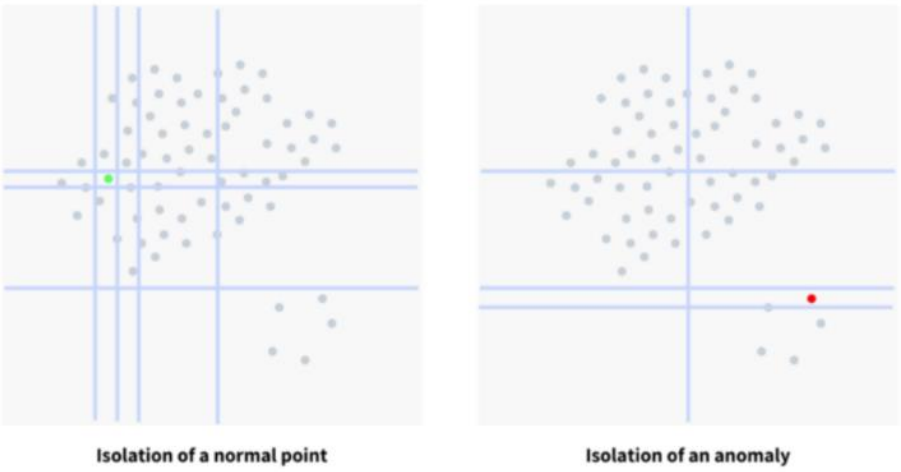


Image 17. Isolation forest technique for outliers detection

The diagram above shows the number of splits required to isolate a normal point and an outlier. Splits, represented through blue lines, happens at random on a random attribute and in the process building a decision tree. The number of splits determines the level at which the isolation happened and will be used to generate the outlier score.

The process is repeated multiple times and we note the isolation level for each point/instance. Once the iterations are over, we generate an outlier score for each point/instance, suggesting its likeliness to be an outlier. The score is a function of the average level at which the point was isolated. The top samples gathered based on the score are labeled as outliers.

clf.fit fits the base estimator using the max_samples count for the feature.

clf.predict returns -1 if observation is deemed an outlier, otherwise 1

clf.decision_function returns the measured outlier score based on fitted model

Having already implemented the multivariate outlier detection models, it was possible to notice that the visualization strategy that we were using (scatterplot with all involved variables) was not the best, since it does not allow us to clearly see the detected outliers due to the large number of variables involved; for this reason it was decided to implement a method of dimensionality reduction

Kernel Principal Component Analysis (KPCA): is a non-linear dimensionality reduction technique. It is an extension of Principal Component Analysis (PCA) - which is a linear dimensionality reduction technique - using kernel methods.

The idea of KPCA relies on the intuition that many datasets, which are not linearly separable in their space, can be made linearly separable by projecting them into a higher dimensional space. The added dimensions are just simple arithmetic operations performed on the original data dimensions.

So we project our dataset into a higher dimensional feature space, and because they become linearly separable, then we can apply PCA on this new dataset.Preforming this linear dimensionality reduction in that space will be equivalent to a non-linear dimensionality reduction in the original space.

As we said doing PCA in the transformed dataset will need a lot of computations. But using kernel methods we can perform these computations in the original state space. This is done by kernel function which gives us a way of computing dot product, between two vectors - in high dimensional space- in our original space.

Some examples of kernel functions are polynomial, Radial Basis Function (RBF) and Gaussian kernels. In our models we applied the polynomial kernel.

KPCA can also be considered as a visualization tool; by looking at the scatter plot of the projected data, we can distinguish the different clusters within the original data. We propose to use visualization given by KPCA in order to clearly identify the outliers. Thus, after implementing this technique we got the following results:

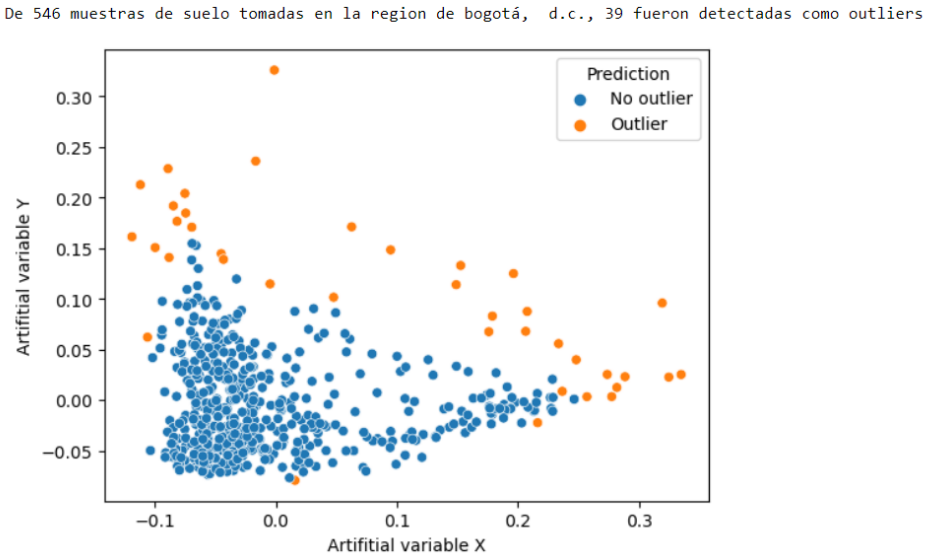


Image 18. Multivariate outliers detection-Isolation forest and KPCA

In this case we reduced the dimension to 2 in order to graph the detected outliers in two dimensions and clearly see the result obtained from our implemented machine learning models (Isolation Forest)

Crops type recommendation Model

One of the main objectives of our project is to create a machine learning model that allows us to predict the most suitable type of crop for a given soil sample: that is, given some physicochemical variables obtained from a soil analysis, our model will recommend which are the top 5 type of crop adequate for these characteristics, In this way, the farmer can take advantage of the current properties of the soil and harvest some of the recommended crops or he can also choose to modify these properties to harvest the crop that he wants. For this project we focus on the first 20 most frequent crops in the dataset, since we have more than 60 crop classes, not all of which have a significant total of samples.

Feature Engineering

Initially some classes of the response variable are grouped, since we have many classes of crops, some of them are repeated or belong to the same class, for this reason a single class is created for them; that is, in our dataset we have crops like citrus, citrus.citrus-orange, citrus-lemon etc; so all kinds of citrus were grouped into a single class called 'citrus', the same for caña de azuca, onions etc.

We also removed those variables that were highly correlated to avoid multicollinearity. In this case, the variables acidity and cice were eliminated from our model, since according to the analysis carried out in the EDA, they are highly correlated with aluminum and calcium.

The one hot encoding technique was implemented for the variables: topography, irrigation, established time, state, drainage, irrigation.

Finally, three new columns were created in our dataset: altitude, longitude and latitude. Since the region in which the soil analysis is done is of great importance and we have many municipalities for which it would be difficult to implement the dummies techniques for all, it was decided to use these 3 variables that represent the location of the analyzed sample.

After cleaning the data and a feature engineering process, we finally have that the **predictor variables** would be the following:

1 ph	27 topografia_no_indica
2 materia_organica	28 topografia_ondulado
3 fosforo	29 topografia_pendiente
4 azufre	30 topografia_pendiente_fuerte
5 aluminio	31 topografia_pendiente_leve
6 calcio	32 topografia_pendiente_moderada
7 magnesio	33 topografia_plano
8 potasio	34 drenaje_bueno
9 sodio	35 drenaje_malo
10 ce	36 drenaje_muy_buen_drenaje
11 hierro_olsen	37 drenaje_no_indica
12 cobre	38 drenaje_regular
13 manganeso	39 riego_aspersión
14 zinc_olsen	40 riego_cañón
15 boro	41 riego_goteo
16 estado_establecido	42 riego_gravedad
17 estado_no_indica	43 riego_manguera
18 estado_por_establecer	44 riego_microaspersión
19 tiempo_establecimiento_de_0_a_1_año	45 riego_no_cuenta_con_riego
20 tiempo_establecimiento_de_1_a_5_años	46 riego_no_indica
21 tiempo_establecimiento_de_5_a_10_años	47 riego_por_inundación
22 tiempo_establecimiento_mas_de_10_años	48 cultivos_agrupados
23 tiempo_establecimiento_no_aplica	49 altitud
24 tiempo_establecimiento_no_indica	50 latitud
25 topografia_ligeramente_ondulado	51 longitud
26 topografia_moderadamente_ondulado	

For our **response variable** we have:

0	cacao	5550
1	pastos	5491
2	aguacate	3846
3	caña de azucar	2457
4	café	2150
5	plátano	1874
6	citricos	1602
7	arroz	1157
8	mora	1088
9	maíz	1020
10	piña	916
11	caucho	834
12	papa de año	831
13	maracuyá	818
14	palma de aceite	743
15	pastos-kikuyo	735
16	lulo	638
17	uva	592
18	millo	530
19	frijol	489

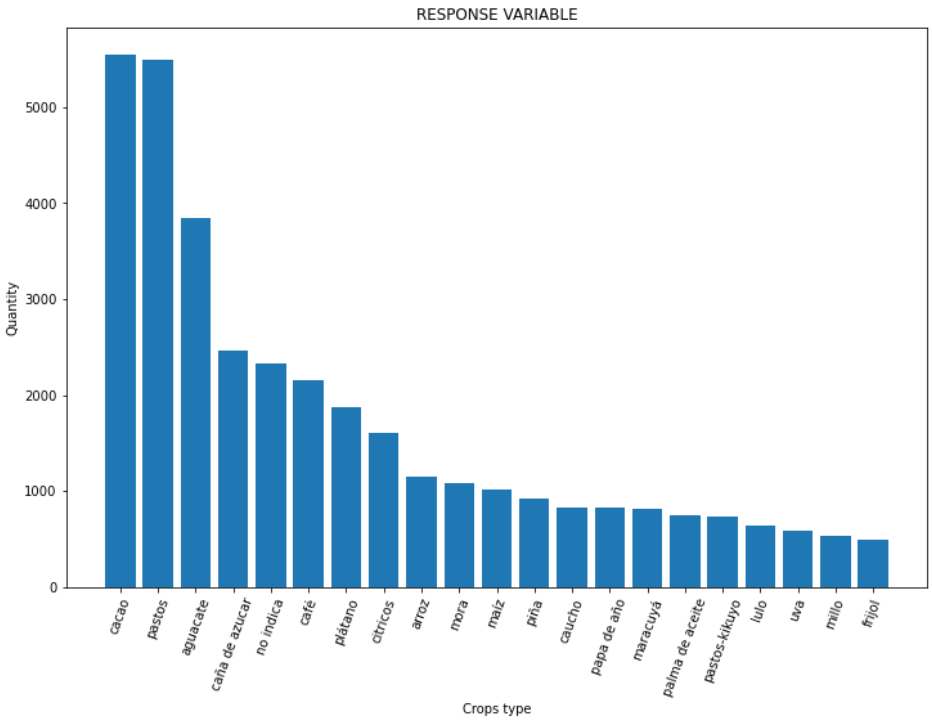


Image 19. Response variable

As we can see in the previous image, we are dealing with a multiclass classification model, where the classes are imbalanced. Imbalanced data typically refers to a problem with classification problems where the classes are not represented equally.

Preparing Data for Model Building

We use one-hot encoding to convert our categorical response variable into numerical variable. In this way, now our crops are numbered from 0 to 19.

After training our model and seeing that although we had an accuracy of 70% some of our classes had a very low recall metric score (those classes that had a small number of samples compared to the classes with a higher number of samples), this means that the model failed to “learn” the minorities classes well, thus failed to correctly predict the minorities classes labels, for this reason it was decided to implement some technique to balance our classes a bit.

There are various techniques involved in improving the performance of imbalanced datasets:

Under-sampling: Remove samples from over-represented classes

Over-sampling: Add more samples from under-represented classes

For our model we used SMOTETomek tool from the Imbalanced learn python library. This method combines the SMOTE ability to generate synthetic data for minority class and Tomek Links ability to remove the data that are identified as Tomek links from the majority class (that is, samples of data from the majority class that is closest with the minority class data).

After implementing this technique we did not get better results, so we decided to continue with our imbalanced classes.

Model implementation-XGBoost (Extreme Gradient Boosting)

XGBoost is a decision-tree-based ensemble Machine Learning algorithm that uses a gradient boosting framework. In prediction problems involving unstructured data (images, text, etc.) artificial neural networks tend to outperform all other algorithms or frameworks. However, when it comes to small-to-medium structured/tabular data, decision tree based algorithms are considered best-in-class right now.

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

After implementing this model, we got the following results:

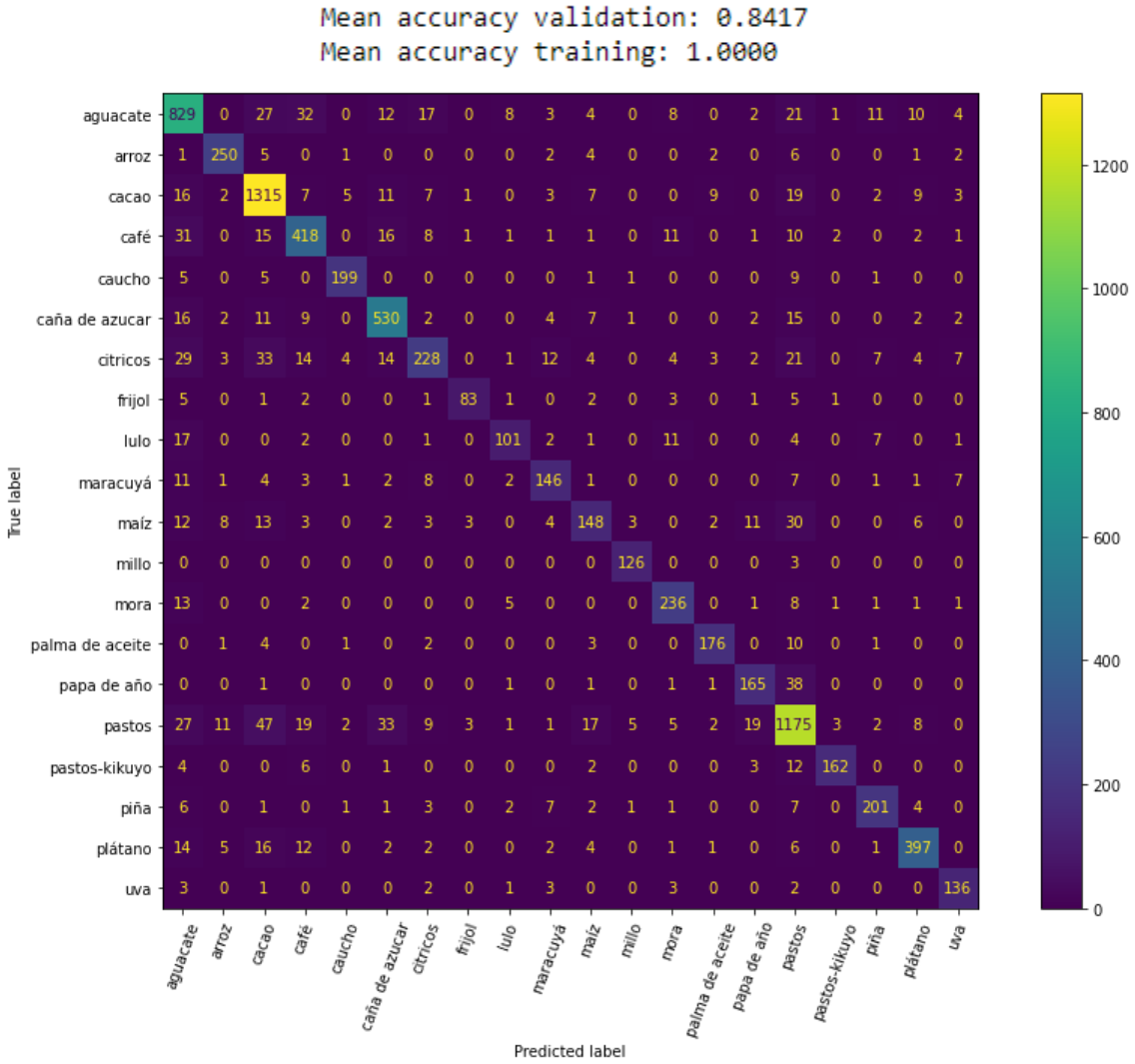


Image 19. Confusion matrix

	precision	recall	f1-score	support
0	0.80	0.84	0.82	989
1	0.88	0.91	0.90	274
2	0.88	0.93	0.90	1416
3	0.79	0.81	0.80	519
4	0.93	0.90	0.91	221
5	0.85	0.88	0.86	603
6	0.78	0.58	0.67	390
7	0.91	0.79	0.85	105
8	0.81	0.69	0.75	147
9	0.77	0.75	0.76	195
10	0.71	0.60	0.65	248
11	0.92	0.98	0.95	129
12	0.83	0.88	0.85	269
13	0.90	0.89	0.89	198
14	0.80	0.79	0.80	208
15	0.83	0.85	0.84	1389
16	0.95	0.85	0.90	190
17	0.86	0.85	0.85	237
18	0.89	0.86	0.87	463
19	0.83	0.90	0.86	151
accuracy			0.84	8341
macro avg	0.85	0.83	0.83	8341
weighted avg	0.84	0.84	0.84	8341

Image 20. Classification report

After implementing the model, we can notice that we have overfitting, since our model trains very well, however when validating the model the accuracy of this validation is much lower than that of the training model.

To find the 5 recommended crops, *model.predict_proba()* was used, which gives us a matrix of probabilities, and from there the 5 highest probabilities are extracted.

CONCLUSIONS

- Most machine learning algorithms work best when the number of samples in each class are about equal. This is because most algorithms are designed to maximize accuracy and reduce error, and Machine Learning algorithms tend to produce unacceptable predictions when faced with imbalanced datasets. In our model we had imbalanced class, so we tried to choose a machine learning technique that performs well with imbalanced classes. Under sampling and oversampling techniques are also implemented to balance the classes, however better results were not obtained.

- Outlier detection algorithms are so useful for any organization, in our case, the main goal for this model was Quality control – the detection of samples defects or sample characteristics that do not fit the same standards as the other samples.

- For Agrosavia laboratory operation, the detection of anomalies is key due to the number of samples that arrive at the institution. Being able to quickly identify potential sample problems allows them to review nutrient measurement procedures and speak with customers in time to determine the source of the anomaly. 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.

- Statistical outlier detection algorithms form one of the earliest, and still one of the most widely used,

techniques that are used for outlier detection. In most cases, statistically probability (or better formulated: statistical improbability) form the basis for detecting any outlier. However, it has the limitation of detecting just for one variable.

- Proximity based techniques are popular techniques for bi-variate and multi-variate data. They are relatively simple to implement, and work by detecting distance between data points. Within the proximity based techniques, k-Nearest Neighbour (k-NN) is by far the most widely used because of the simplicity of the underlying calculation. There is however an important drawback in using proximity based techniques for outlier detection. The runtime complexity is proportional to the size of the data; in our case, we solved this by better implementing the isolation forest technique to detect multivariate outliers.

- With different machine learning techniques, it is possible to predict the types of crops suitable for a certain soil, however, a soil can be modified in order to achieve the appropriate properties for a certain desired crop, for this reason, it would be good to implement a model that instead of predicting the 5 most suitable types of crops, predict the soil properties that must be modified for a certain desired crop.

BIBLIOGRAPHY

Cepal, Patrimonio Natural. (s.f.). Amazonia. Obtenido de Amazonia posible y sostenible: https://www.cepal.org/sites/default/files/news/files/folleto_amazonia_posible_y_sostenible.pdf

Gobierno de Colombia. (abril de 2020). Plan Nacional para la Reforma Rural Integral. Obtenido de Plan Nacional de Riego y Drenaje para la Economía Campesina, Familiar y Comunitaria: https://www.minagricultura.gov.co/Normatividad/Resoluciones/RESOLUCI%C3%93N%20NO.%20000091%20DE%202020.pdf

Maiz para Colombia Vision 2030. (Julio de 2019). Maiz para Colombia Vision 2030. Obtenido de Fenalce: https://www.fenalce.org/archivos/maiz2030.pdf

Ministerio de Agricultura, pesca y alimentación. (s.f.). Hojas divulgadoras. Obtenido de Interpretación y análisis de los suelos: https://www.mapa.gob.es/ministerio/pags/biblioteca/hojas/hd_1993_05.pdf

Peña-Venegas, C. P., & Vanegas Cardona, G. I. (Diciembre de 2010). Dinámica de los suelos Amazónicos. Obtenido de Procesos de degradación y alternativas para su recuperación: https://www.sinchi.org.co/files/publicaciones/publicaciones/pdf/librosuelosweb.pdf