IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 5, June 2022
Impact Factor: 6.252

Soil Testing Prediction System

Yash Bhardwaj
Student, Department of Computer Science and Information Technology
Dronacharya College of Engineering, Gurugram, India

Abstract: *Soil Testing Prediction aims to forecast the functional qualities of a soil sample (calcium, phosphorus, pH, sand, and soil organic carbon). Soil Testing Prediction is used in agriculture, farming, and research. It can aid in cost-effective crop management and increased agricultural output. We investigated how old soil testing methods could be substituted with modern Machine Learning approaches, resulting in more cost-effective, time-efficient ways with little to no environmental impact. It tries to bring the labs to the user rather than the user going to the labs, and it trains to lower the technical expertise required at the user's end. Instead of taking the user to the lab, it tries to bring the lab to the user.*

Keywords: Linear Regression, Feature Selection, Soil Functional Properties, Extraction Methods for Mehlich-3

I. INTRODUCTION

Because agricultural area is reducing every day in modern times, it is critical to analyse the soil before undertaking any agricultural operation in order to produce the best potential output. The existing preventative measures have proven to be time consuming, costly, and ineffective. As a result, in this fast-paced environment, we require a more effective testing method that can assist us in overcoming the aforementioned challenges.

Due to the massive expansion in the country area, we will need to seek new areas for farming in the near future, and we will need to check the quality of our soil to produce the greatest products. As a result, soil testing prediction plays an important part in:

1. Determining whether a specific type of soil is suitable for use.
2. Forming fertiliser recommendations based on the greatest feasible estimate of the soil's fertility.
3. Diagnosis of plant problems and quality plant production, among other things
4. Instead of the slow and expensive preemptive methods currently in use, the measurement can often be completed in seconds.

Traditional procedures could be used to do this soil testing. These methods are referred to as WET Tests because they require the use of chemicals and are time intensive, making these tests highly costly. The high cost of the test kit, as well as the demand of professional skills, are the reasons behind the high cost of these tests. This is where machine learning comes in. We will use Africa soil Infrared Spectroscopy data with 3600 features and 1157 samples to train our machine learning models. This data will only be used to train and cross-validate the data; it will not be tested on a previously unknown dataset.

II. BACKGROUND

2.1 Wet Test (Traditional)

WET tests are time-consuming and expensive traditional testing methods that involve chemical-based methodologies. Furthermore, they necessitate a high level of technical competence, and even minor deviations from ideal conditions might result in significant mistake. This problem can be overcome through automation, which allows us to eliminate humans from the equation and achieve more effective results.

2.2 Kit Possibilities

The WET Test, which may be done with the use of a WET Testing kit, is the only currently available method for testing soil. The main issue with this kit is that it requires a high level of experience to utilise it, as well as the fact that it is fairly pricey, preventing it from being used by its direct users.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 5, June 2022
Impact Factor: 6.252

As a result, the idea of customers using the kit directly has to be replaced with our proposed Machine Learning Model, which would not require as much technical competence or as much money from direct consumers. They could readily predict required content using our user-friendly interface.
Another option is to deploy IoT-based infrastructure that can generate real-time data on-site. This information can be fed into real-time processing systems, yielding real-time forecasts.

2.3 Limitations

1. Most small-scale and large-scale end consumers, as well as some major customers, cannot employ current soil testing methods directly.
2. The existing soil testing prediction method is difficult for a novice to employ.
3. The existing soil testing procedure cannot exist without the use of toxic chemicals, which would in any case cause environmental devastation.
4. Conventional methods these days require a significant amount of time to provide test results, which is problematic given the size of the Indian consumer market.

III. PROPOSED METHODOLOGY

The goal of the project is to forecast the following:

1. SOC stands for soil organic carbon.
2. pH: pH levels
3. Ca: extractable Mehlich-3 Calcium
4. P: Extractable Mehlich-3 Phosphorus
5. Sand content: The amount of sand in a given amount of soil.

IV. DATA CONFUSION

Cleaning data, i.e. making it more normal and removing any outlier or unscaled features, is referred to as wrangling.

4.1 Removal of Outliers

Outliers are data points that do not follow the general trend in the data, i.e., they will have an incorrect impact on the result, fluctuate it from the actual one, and raise the error, or divergence from the projected behaviour.

- Scaling
- Normalization

$$x_{scaled} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

Where x is the sample point
x_{min} is the minimum value over the domain
x_{max} is the maximum value over the domain

4.2 Reduction of Dimensionality

Because the data set's dimensions were large, PCA was employed to decrease it to a manageable number. Principal component analysis is a technique for reducing a dataset's dimensionality (feature set). It creates a projection of a higher-dimensional feature set to a lower-dimensional feature set, resulting in a set of new features with higher variance and, as expected, better predictive power. However, it does cost us some knowledge when compared to the original dataset.

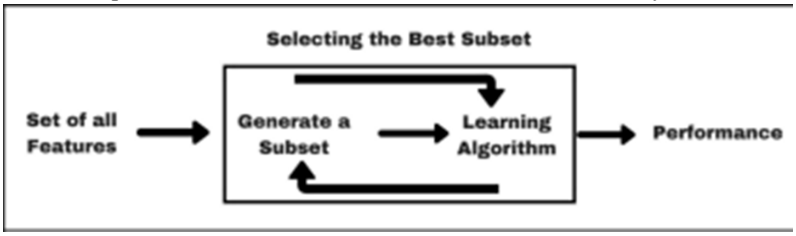
4.3 Selection of Features

We utilised recursive feature removal and the Random forest Regressor to extract the needed amount of features (50 in total).

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 5, June 2022
Impact Factor: 6.252

4.4 Feature Elimination Recursively

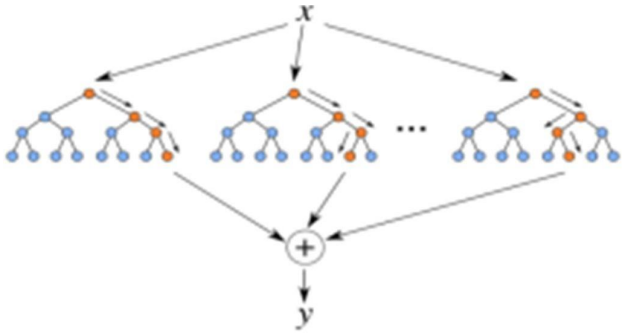
RFE (recursive feature elimination) is a feature selection strategy that allows us to extract the best subset of features from a large number of features. It does this by fitting the model to the given data and removing the least significant feature one by one. The procedure is repeated until the desired number of features or accuracy is obtained.



Forest Regressor

4.5 Random

A random forest regressor, as the name implies, uses a number of random trees to generate a random forest. Based on the random forest regression scores, it attempts to extract the best possible features for the model in order to predict the targets. It is one of the most popular feature selection techniques, and it is available for both classification and regression tasks. It trains each tree on a subset of a given dataset, sampling to improve accuracy.

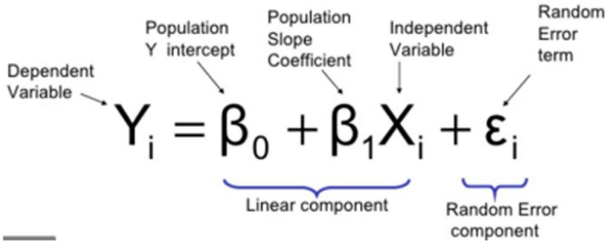


V. MODEL TRAINING

We utilised two techniques to create the model: ordinary least square regression and light gradient boosting machine.

5.1 Least Squares Regression (OLSR)

Ordinary Least Square Regression, commonly known as basic linear regression, is one of the numerous approaches used in machine learning that come from the study of statistics. It is based on the assumption that the target value and the attributes have a linear relationship. It calculates the optimum fit line for the feature-target relationship, which, despite its simplicity, has proven to be extremely accurate.



Machine that boosts the light gradient

LightGBM, short for Light Gradient Boosting Machine, is a distributed gradient boosting system for machine learning that was created by Microsoft. It is used for ranking, classification, and other machine learning applications and is based on decision tree algorithms.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 5, June 2022
Impact Factor: 6.252

VI. ACCURACY MEASURES

The Predictions are scored upon Mean Column wise root mean squared error (MCRMSE).

$$\text{MCRMSE} = \frac{1}{5}\sum_{j=1}^{5}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_{ij} - \hat{y}_{ij})^2},$$

VII. RESULT DISCUSSION

Accuracy for OLSR -0.83577
Accuracy for LGBM -0.46892

MCRMSE (mean columnwise root mean squared error)

```
[ ]  pans = 0.0
     for i in range(len(pres)):
       pans  = pans + (pres[i] - y_test_P.values[i])**2

     socans = 0.0
     for i in range(len(socres)):
       socans  = pans + (socres[i] - y_test_SOC.values[i])**2

     phans = 0.0
     for i in range(len(phres)):
       phans  = phans + (phres[i] - y_test_pH.values[i])**2

     sandans = 0.0
     for i in range(len(sandres)):
       sandans  = sandans + (sandres[i] - y_test_Sand.values[i])**2

     MCRMSE = (((ans+pans+socans+phans+sandans)/382)**0.5)/5

[ ]  MCRMSE

     0.835778088242345
```

VIII. CONCLUSION

As we all know, soil testing has become a need in today's environment. However, using standard methods would take around 5 to 6 business days, but our model proposes to get a nearly identical answer in seconds, with a Mean column wise root mean square error of 0.83577 and 0.46892 for OLSR and LGBM, respectively.

IX. ADVANTAGES

1. Cost Efficient
2. Fast processing and predictions in real time
3. Better Accuracy
4. Minimal or no use of Chemicals

REFERENCES

[1]. "Carge`le Nduwamungu , Noura Ziadi , Le´on-E´tienne , Gae¨tan F. Tremblay , and Laurent Thurie`s (2009) Opportunities for, and limitations of, near infrared reflectance spectroscopy applications in soil analysis: A review" sklearn.linear_model.LinearRegression

[2]. "Soil Analysis using Mehlich 3 Extractant Technique for Sample Preparation . ÚKZUZ (Ústřední kontrolní a zkušební ústav zemědělský) Central Institute for Supervising and Testing in Agriculture Hroznová 2, CZ-65606 Brno, Czech Republik L. Vlk, M. Horová, R. Krejča; R. Špejra Chromservis S.R.O., Jakobiho 327, CZ-10900 Praha-10, Petrovice, Czech Republik