Soil health Prediction for agriculture using machine learning

Presented by : Pratik kage
Github link :

Abstract

Information about soil properties help the farmers to do effective and efficient farming, and yield more crops with less usage of resources. An attempt has been made in this paper to predict the soil properties using machine learning approaches. The main properties of soil prediction are Calcium, Phosphorus, pH, Soil Organic Carbon, and Sand. These properties greatly affect the production of crops. Four well-known machine learning models, namely, multiple linear regression, random forest regression, support vector machine, and gradient boosting, are used for prediction of these soil properties. The performance of these models is evaluated on Africa Soil Property Prediction dataset. Experimental results reveal that the gradient boosting outperforms the other models in terms of coefficient of determination. Gradient boosting is able to predict all the soil properties accurately except phosphorus. It will be helpful for the farmers to know the properties of the soil in their particular terrain.

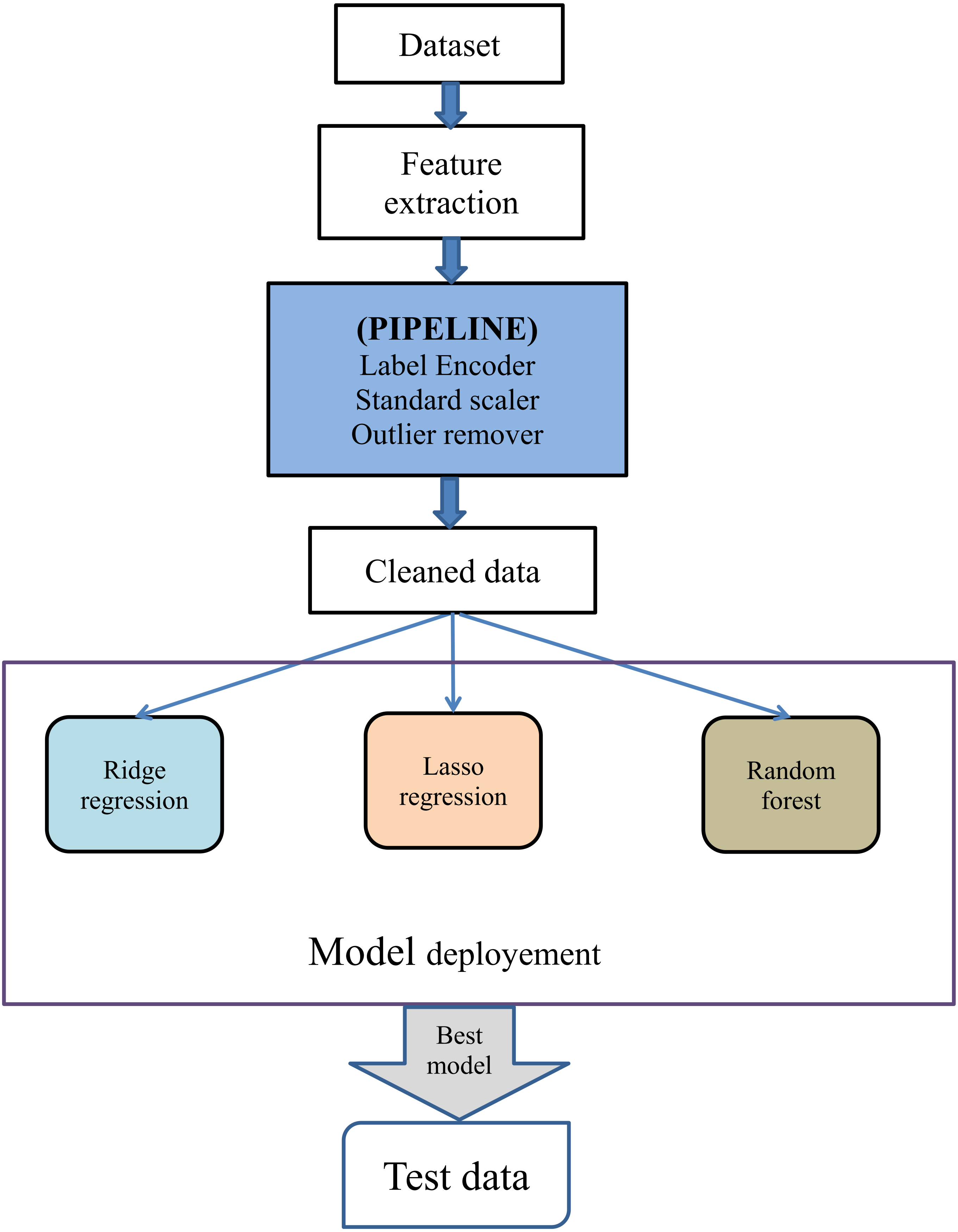
Introduction

India has a 1.27 billion population, which is second-most in the entire world. It is the seventh-largest country in the world with an area of 3.288 million sq km. Indians are very much dependent on agriculture. It is the largest source of livelihood in India. In rural households, 70% of people are primarily dependent on agriculture, with about 82% of farmers being small and marginal. In 2020-21, total food grain production was estimated at 308.65 million tonnes (MT). India is the largest producer (25% of global production), the consumer (27% of world consumption), and the importer (14%) of pulses in the world. India's annual milk production was 165 MT (2017-18), making India the largest producer of milk, jute, and pulses, with the world's second-largest cattle population of 190 million in 2012. With merely 2.4% arable land resources and 4% water resources , Indian agriculture is feeding nearly 1.3 billion people, which implicates huge pressure on land and other natural resources for continuous productivity.After the green revolution(which started in the 1960s), India made significant progress in agriculture production, which became possible due to modernization. With the development in technology, farmers have been provided with advanced farming techniques, better seeds (High Yielding Variety(HYV) seeds), mechanized farm tools, chemical fertilizers, facilities of irrigation, and electrical energy. Since the green revolution, there has been excessive use of chemical fertilizers which has increased the crop productivity manifold. However, it has turned into a problem as overuse of these chemical fertilizers has been detrimental for crop productivity and soil fertility. Fertilizer recommendations rarely match soil needs which has caused overuse of these chemical products.So, there is a need for accurate fertilizer recommendationsfor the farmer and accurately analyzing soil properties is the first step for that. Indian Agricultural Research institute(ICAR) recommends soil test-based, balanced and integrated nutrient management through conjunctive use of both inorganic and organic sources of plant nutrients to reduce the use of chemical fertilizers, preventing deterioration of soil health, environment and contamination of groundwater.

This project aims to study the ability of various machine learning techniques to accurately predict the soil properties relevant for agriculture using spectroscopy data. Over the last 20 years, soil spectroscopy has become a powerful technique for analyzing relative to the traditionally used chemical methods, particularly in the infrared range. Spectroscopy is known as a fast, economical, quantitative, and eco-friendly technique, which can be used in the fields as well as in the laboratory to provide hyperspectral data with narrow and numerous data. In this paper, the different properties of soil like Calcium, Phosphorus, pH, Soil Organic Carbon and Sand are predicted by using machine learning models. Issma et al. found consistently higher performance of machine learning methods over simpler approaches in spectroscopy. Yu et al. reported the decline in the use of some models such as Support Vector Machines (SVM) and multivariate adaptive regression spline, giving way to more advanced alternatives such as Random Forest (RF). In this project, machine learning algorithms such as Ridge Regression, Random Forest Regression,Lasso regression Sup-port Vector Machine, and Gradient boosting with a different degree of accuracy are used for comparative analysis.

The dataset is split into a training and testing dataset (80% training data and 20% testing data). The machine learning models are trained on the training data. After a model is trained, the testing data is used to check the accuracy of the trained model. Here, the coefficient of determination (COD) is calculated to check the working of the models after being trained. After training the models, the best working model is deployed to predict the properties of the soil (Calcium, Phosphorous, pH, Soil Organic Carbon, and Sand). These predicted values of the soil properties are going to be helpful in choosing the different suitable fertilizers.

Methodology



Materials and Methods

In this section, the dataset and techniques used for soil prediction are briefly described.

Data Set :

A collection of 1,886 soil sample measures is used for performance comparison of machine learning models. The soil was collected from a variety of locations in Africa. Each data point consists of 3,594 features :

1. PIDN: unique soil sample identifier

2. m7497.96 - m599.76:
   There are 3,578 mid-infrared absorbance measurements. For example, the ”m599.76” column is the absorbance at wavenumber 599.76 cm-1.

3. . Depth:
   Depth of the soil sample (2 categories: 1. ”Topsoil”, 2. ”Subsoil”)

4. BSA:
Average long-term Black Sky Albedo measurements from MODIS satellite images (BSAN = near-infrared, BSAS = shortwave, BSAV = visible)

5. CTI:
Compound topographic index calculated from Shuttle Radar Topography Mission elevation data

6 . ELEV:
Shuttle Radar Topography Mission elevation Data

7. EVI:
Average long-term Enhanced Vegetation Index

from MODIS satellite images

8. LST:
Average long-term Land Surface Temperatures from MODIS satellite images (LSTD=day time temperature, LSTN night time temperature)

9. Ref:
Average long-term Reflectance measurements from MODIS satellite images (Ref1 = blue, Ref2 = red, Ref3 = near-infrared, Ref7 = mid-infrared)

10. Reli:Topographic Relief calculated from Shuttle Radar Topography mission elevation data

11. TMAP TMFI:
Average long-term Tropical Rainfall Monitoring Mission data (TMAP = Mean Annual Precipitation, TMFI = Modified Fournier Index)
The five main target variables for predictions are: Soil Organic Carbon(SOC), pH, Calcium, Phosphorus, and Sand.

Target(labels):

12. SOC: Soil organic carbon

13. pH: pH values

14. Ca: Mehlich-3 extractable Calcium

15. P: Mehlich-3 extractable Phosphorus

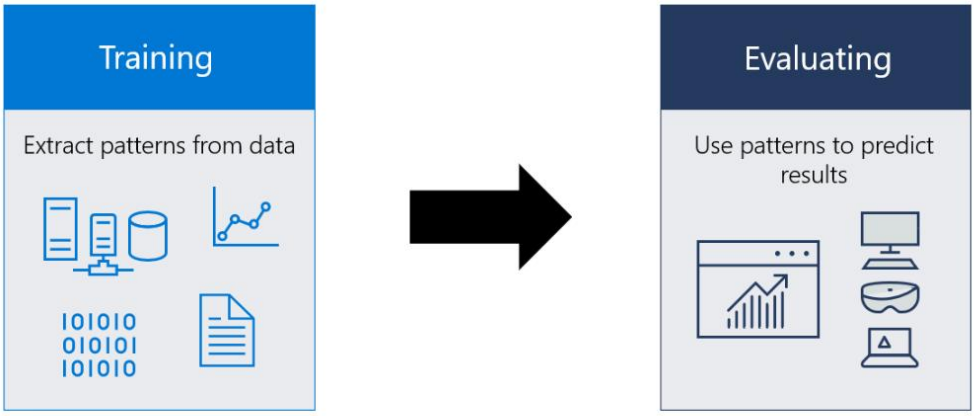
16. Sand: Sand content

Purpose

The purpose of this project is to use machine learning techniques to predict harmful soil properties that affect soil health. By analyzing soil data, the project aims to identify potential issues early on, helping farmers and agricultural stakeholders make informed decisions to improve soil quality and promote sustainable farming practices

Machine learning Modelling

The process of modeling means training a machine learning algorithm to predict the labels from the features, tuning it for the business need, and validating it on holdout data. The output from modeling is a trained model that can be used for inference, making predictions on new data points.

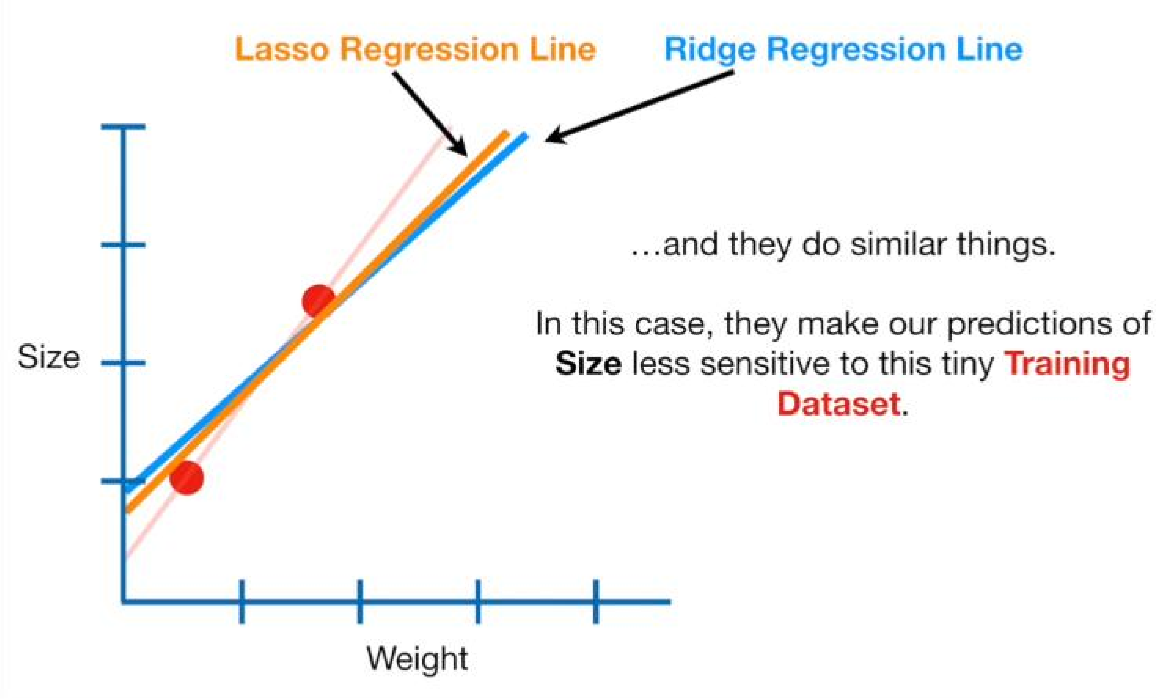


A machine learning model itself is a file that has been trained to recognize certain types of patterns. You train a model over a set of data, providing it an algorithm that it can use to reason over and learn from those data. Once you have trained the model, you can use it to reason over data that it hasn't seen before, and make predictions about those data. For example, let's say you want to build an application that can recognize a user's emotions based on their facial expressions. You can train a model by providing it with images of faces that are each tagged with a certain emotion, and then you can use that model in an application that can recognize any user's emotion

Ridge regression

Ridge regression is a method of estimating the coefficients of multiple-regression models in scenarios where the independent variables are highly correlated. It has been used in many fields including econometrics, chemistry, and engineering.Also known as **Tikhonov regularization**, named for Andrey Tikhonov, it is a method of regularization of ill-posed problems.It is particularly useful to mitigate the problem of multicollinearity in linear regression, which commonly occurs in models with large numbers of parameters. In general, the method provides improved efficiency in parameter estimation problems in exchange for a tolerable amount of bias (see bias–variance tradeoff).

The theory was first introduced by Hoerl and Kennard in 1970 in their *Technometrics* papers "Ridge regressions: biased estimation of nonorthogonal problems" and "Ridge regressions: applications in nonorthogonal problems". This was the result of ten years of research into the field of ridge analysis.



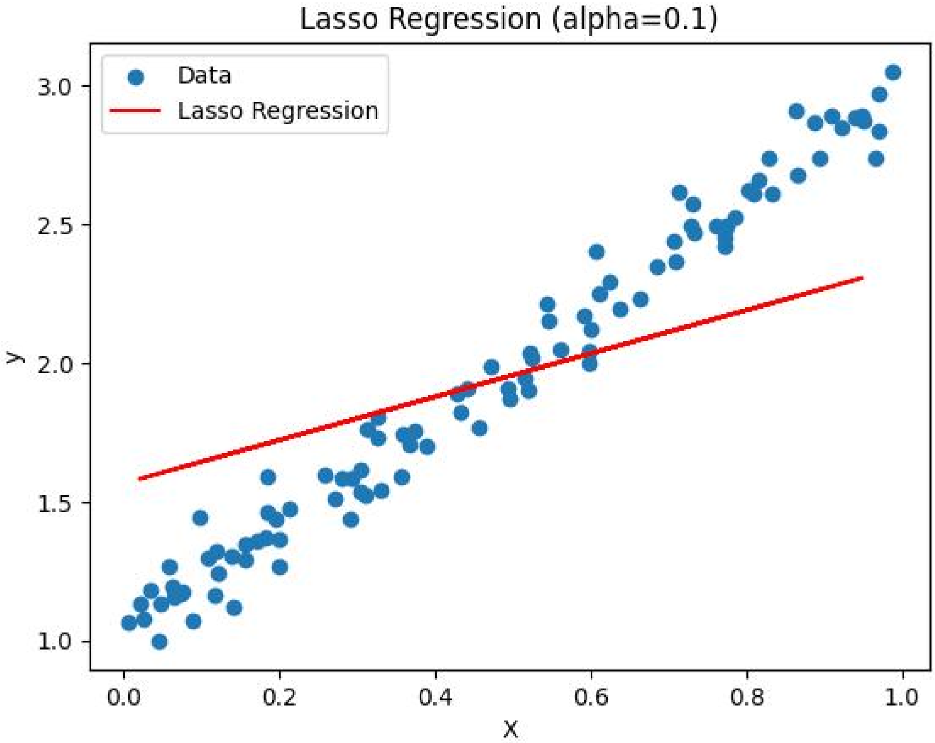
Ridge regression was developed as a possible solution to the imprecision of least square estimators when linear regression models have some multicollinear (highly correlated) independent variables—by creating a ridge regression estimator (RR). This provides a more precise ridge parameters estimate, as its variance and mean square estimator are often smaller than the least square estimators previously derived.

Lasso regression

In statistics and machine learning, **lasso** (**least absolute shrinkage and selection operator**; also **Lasso** or **LASSO**) is a regression analysis method that performs both variable selection and regularization in order to enhance the prediction accuracy and interpretability of the resulting statistical model. The lasso method assumes that the coefficients of the linear model are sparse, meaning that few of them are non-zero. It was originally introduced in geophysics, and later by Robert Tibshirani, who coined the term.

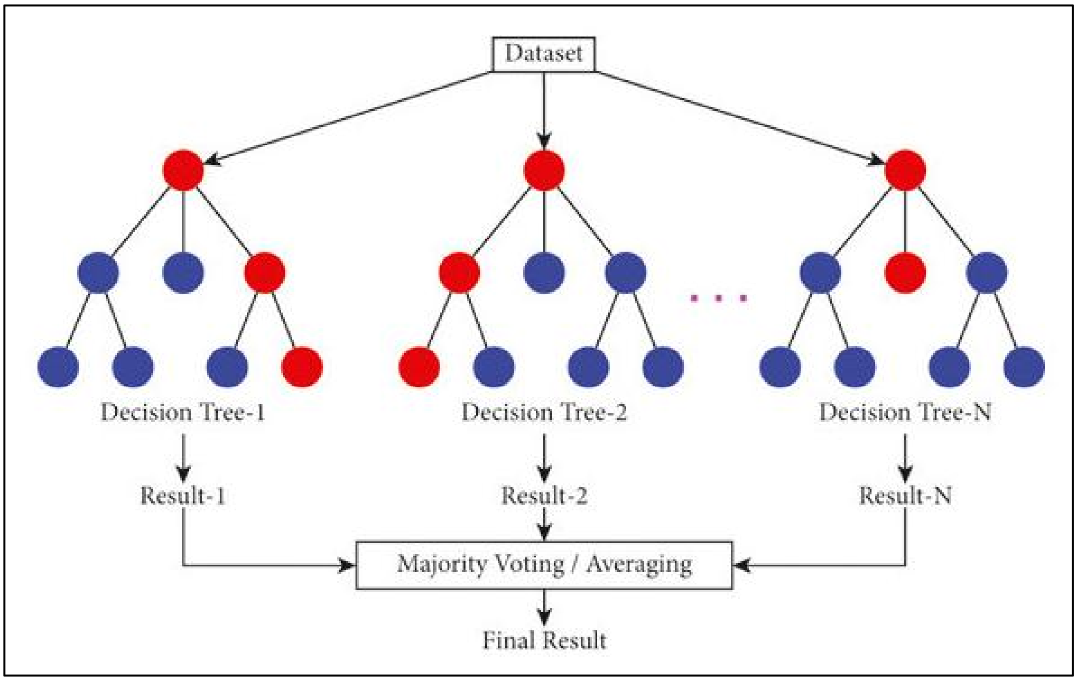
Lasso was originally formulated for linear regression models. This simple case reveals a substantial amount about the estimator. These include its relationship to ridge regression and best subset selection and the connections between lasso coefficient estimates and so-called soft thresholding. It also reveals that (like standard linear regression) the coefficient estimates do not need to be unique if covariates are collinear.

Though originally defined for linear regression, lasso regularization is easily extended to other statistical models including generalized linear models, generalized estimating equations, proportional hazards models, and M-estimators. Lasso's ability to perform subset selection relies on the form of the constraint and has a variety of interpretations including in terms of geometry, Bayesian statistics and convex analysis.



Random forest¶

Random forest is a supervised learning algorithm. The “forest” it builds is an ensemble of decision trees, usually trained with the “bagging” method. The general idea of the bagging method is that a combination of learning models increases the overall result.One big advantage of random forest is that it can be used for both classification and regression problems, which form the majority of current machine learning systems. and It also resists overfitting found in decision trees.



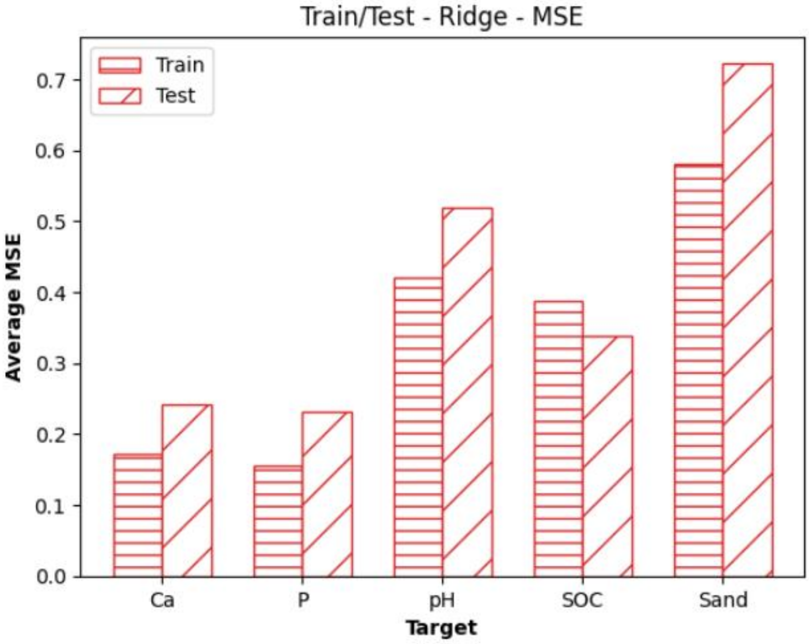
Conclusion :

This project concentrated on the AI/ML methods to anticipate soil properties for accuracy farming. 3 AI/ML methods were utilized to assess the dirt properties like Calcium, Phosphorus, pH, Soil Organic Carbon, and Sand. These strategies were prepared and tried on the Africa Soil Property Prediction dataset. It is seen from the outcomes that Random Forest algorithm performed better compared to different strategies. Random forest had the option to anticipate all properties better than Ridge regression and Lasso regression. It tends to be seen that there is a possibility to involve spectroscopy as an elective strategy for soil part examination. Profound learning and crossover models might be utilized for anticipating soil properties in a compelling and proficient way. The principal constraint of our review is the utilization few soil parts for expectation. This study can be stretched out by utilizing an enormous dataset and different models.

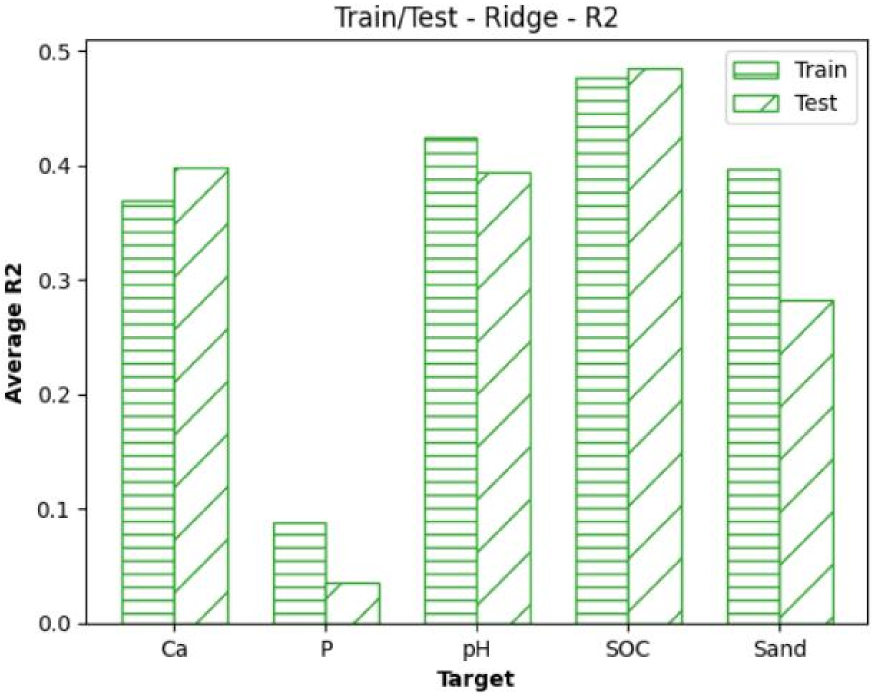
Results :

1. Ridge regression :

    a) MSE

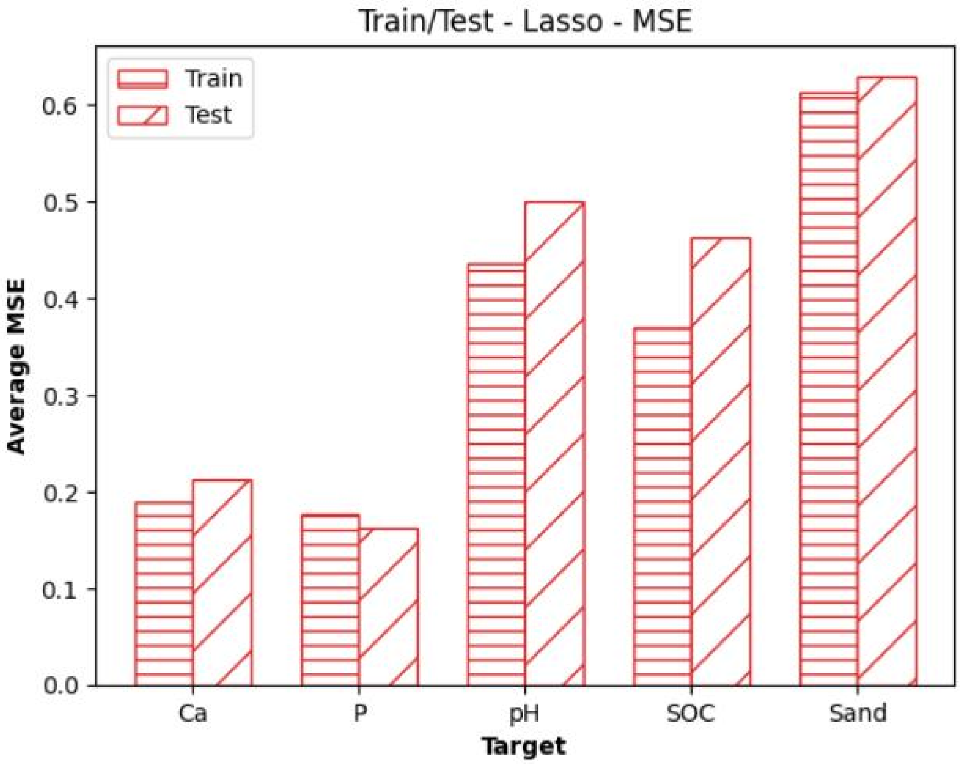


    b) R2 score

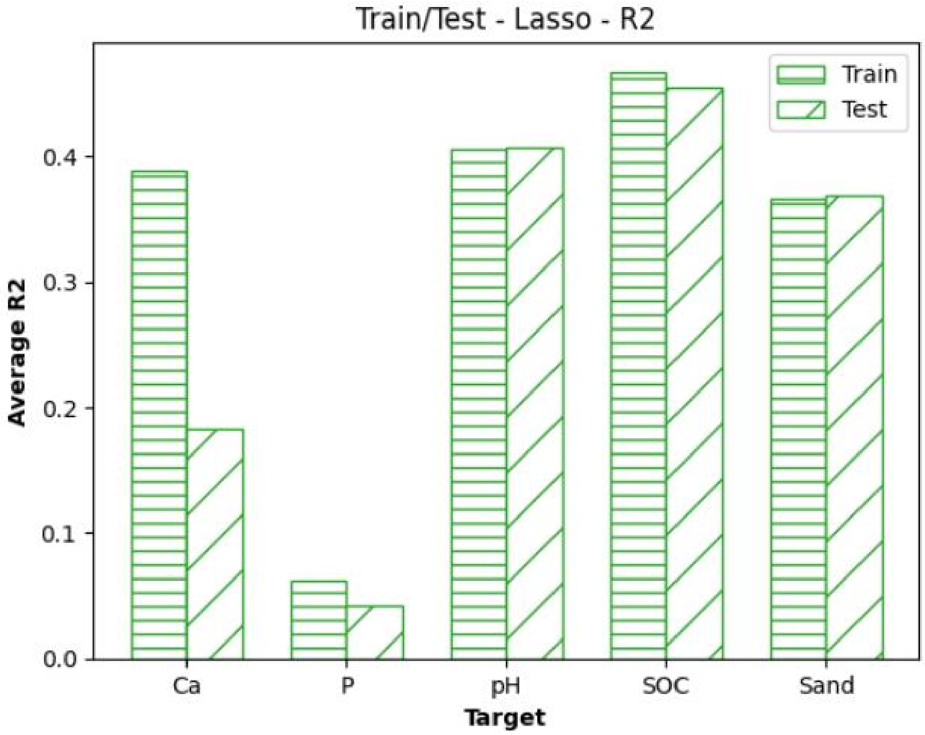


2. Lasso regression :

a) MSE

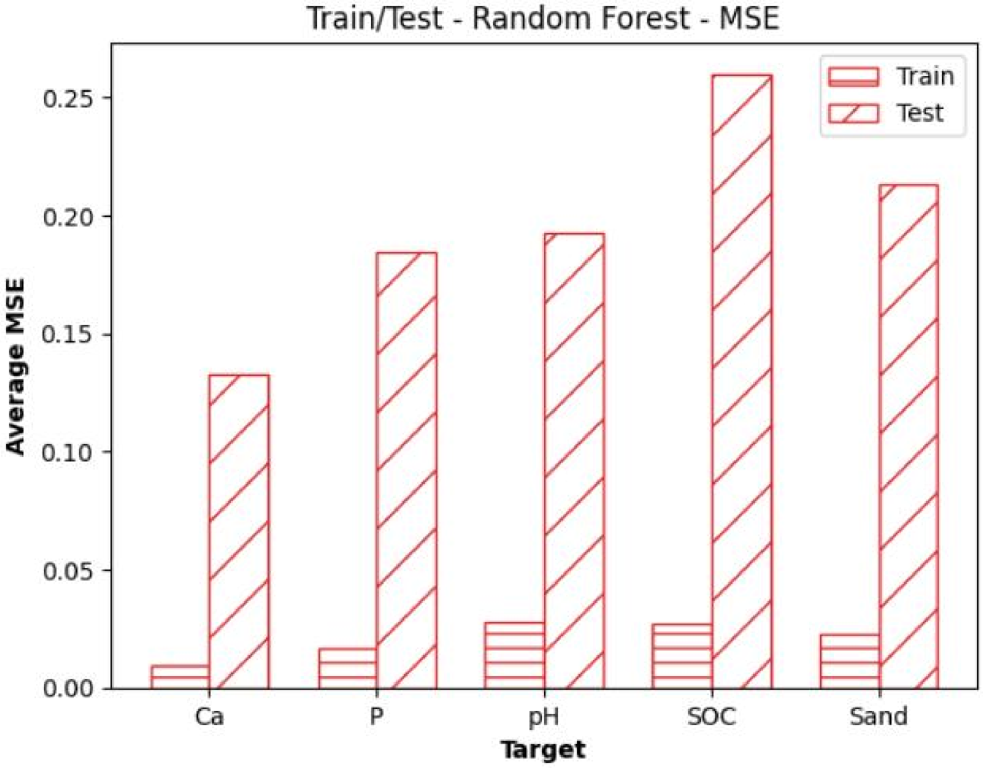


b) R2 score

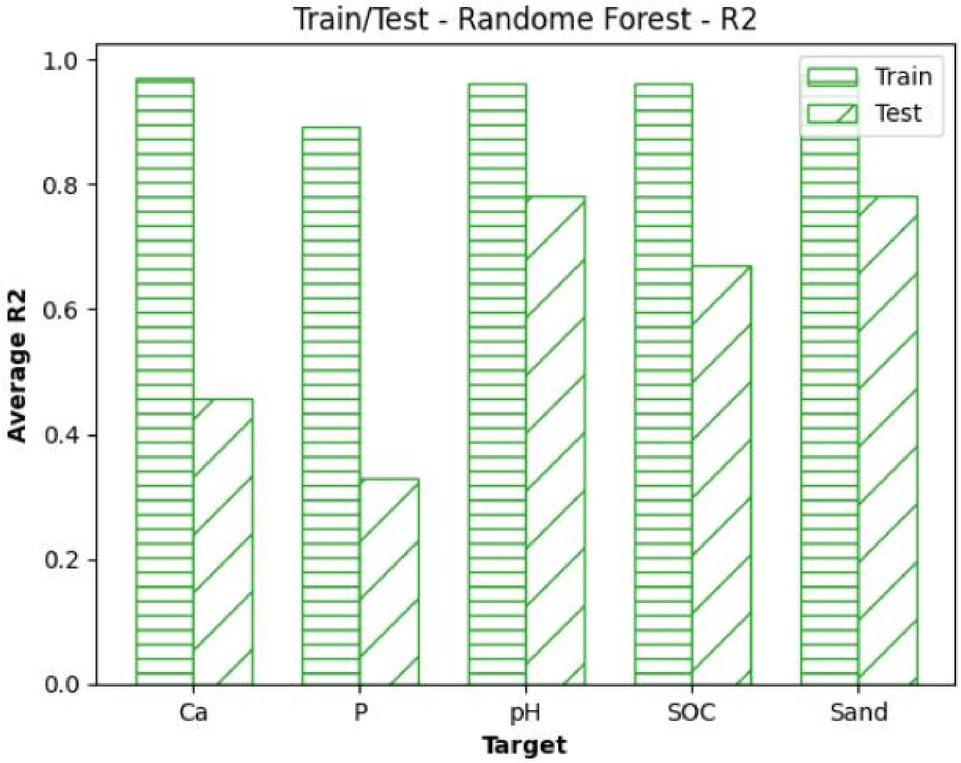


3. Random forest :

a) MSE



b) R2 score


Train/Test - Randome Forest - R2
Average R2
Train
Test
Ca
P
pH
SOC
Sand
Target