Soil Fertility Prediction Using Machine Learning

HARSHIT KUMAR SAHU

Dept. of Computer Science and Information Technology,
Guru Ghasidas Vishwavidyalaya, Bilaspur, Chhattisgarh, India

Abstract: Soil fertility plays a crucial role in determining crop production volume. However, if the composition of soil nutrients, such as fertilizers, is not properly controlled and maintained, it can result in lower crop yields. Therefore, the measurement of soil nutrients is essential for achieving better plant growth and effective fertilization. Calcium (Ca), phosphorous (P), and pH level are among the key parameters commonly measured to monitor soil fertility as they provide important information for determining the soil's fertility status.
In this paper, we propose the use of Machine Learning (ML) algorithms to build a predictive model that can assist farmers in estimating the quantity of soil properties, including Ca, P, and pH values. These parameters can guide farmers in determining the appropriate amount of fertilizers to be added to the soil based on the measured values of Ca, P, and pH, thereby maintaining consistent soil fertility. We evaluate the performance of various ML algorithms, such as Linear Regression, Random Forest, Gradient Boosting, Ridge, K-Nearest Neighbours, Decision Tree, and Artificial Neural Network, using Python programming packages to optimize the prediction accuracy.

Keywords: Machine learning, soil fertility, Linear Regression, Random Forest, Gradient Boosting, Ridge, K-Nearest Neighbours, Decision Tree, Artificial Neural Network.

INTRODUCTION

Indeed, agriculture is a significant sector in India that has a profound impact on the Indian economy. However, urbanization and industrialization have led to a reduction in cultivatable land and declining soil fertility, posing a challenge to increase agricultural production while preserving the environment. One approach to addressing this challenge is to enhance soil fertility by providing essential nutrients to plants in the right amount and at the right time.

Soil management practices are crucial for boosting crop production while maintaining soil nutrients. Farmers need to determine the soil fertility requirements to achieve better and more cost-effective crop production. Soil pH is a critical soil parameter as it provides valuable information about various aspects of soil fertility. Major soil nutrients that contribute to yield production include phosphorus, potassium, nitrogen, calcium, and pH. Insufficient nutrient levels or excessive fertilization can result in lower crop yields. Therefore, it is essential to apply the appropriate quantity of fertilizer for optimal plant growth.

Over the years, many researchers have focused on maximizing soil fertility to meet production requirements. One approach involves analysing the soil and its chemical composition, which plays a significant role in defining soil fertility. By understanding the soil's nutrient content and characteristics, farmers can make informed decisions regarding soil management practices,

fertilization, and other interventions to enhance agricultural productivity while minimizing environmental impacts.

PROBLEM STATEMENT

The success of crop production is heavily reliant on the availability and rate of soil nutrients. It is crucial for farmers to accurately determine the soil fertility requirements to maximize crop yields. Insufficient knowledge about the soil composition can pose a significant challenge as it may lead to planting without the necessary nutrients present in the soil. This, in turn, can result in poor agricultural production quality and reduced crop yields. Using inappropriate rates of soil nutrients can lead to crop degradation and lower productivity.

Improper usage of soil fertilizers can have detrimental effects on production quality. A notable example is seen in China, where the inappropriate usage of fertilizers has resulted in low product quality and even critical environmental problems. Hence, this current research focuses on evaluating various Machine Learning algorithms to develop a robust predictive model that allows farmers to estimate the quantity of soil nutrients present in the soil.

By utilizing Machine Learning algorithms, this research aims to create a reliable model that can assist farmers in predicting the levels of soil nutrients accurately. This predictive model will enable farmers to make informed decisions regarding the appropriate application rates of fertilizers, thus optimizing soil fertility and ultimately improving crop production. The evaluation of multiple Machine Learning algorithms ensures that the chosen model provides accurate predictions and practical utility for farmers.Overall, this research aims to address the challenges associated with soil nutrient management by leveraging Machine Learning techniques, offering a valuable tool for farmers to enhance crop production quality and yield by effectively predicting and managing soil fertility.

OBJECTIVE OF STUDY

The objective of this research study is to implement and compare various Machine Learning algorithms using Python libraries to develop an effective predictive model. The model will assist farmers in determining the quantity of specific soil compositions, specifically Calcium (Ca), Phosphorus (P), and pH levels. The study aims to analyse different types of soil samples provided and generate accurate predictions of these soil parameters. To achieve this objective, We will employ machine learning algorithms such as Linear Regression, Random Forest, Gradient Boosting, Ridge, K-Nearest Neighbours, Decision Tree, and Artificial Neural Network. These algorithms will be trained and tested using the available soil sample data, allowing for the comparison of their performance in predicting the quantities of Calcium, Phosphorus, and pH levels.

By evaluating and comparing the results obtained from the different algorithms, We aim to identify the best predictive model. This model will enable farmers to make informed decisions regarding the appropriate application rates of Calcium, Phosphorus, and pH adjustments based on the specific soil composition of their fields. Ultimately, the study seeks to provide farmers with a valuable tool for optimizing soil fertility management, thereby enhancing crop production and yield.

METHODOLOGY

DATASET

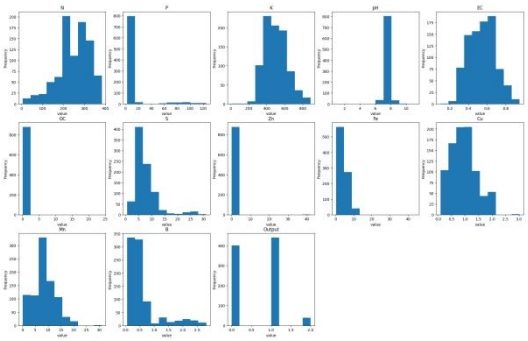
Dataset for this research is provided by https://www.kaggle.com/datasets/rahuljaiswalonkaggle/soil-fertility-dataset as the data set contains various attributes that are contributing in soil fertility , and the record of low moderate and highly fertile soil is mentioned in the dataset as output. The dataset contains 880 rows and 13 columns which have all unique conditions of soil. and can help us to train a good prediction model. The attributes present in the dataset are,

ATTRIBUTE	DESCRIPTION
N	Nitrogen, ppm
P	Phosphorous, ppm
K	Potassium, ppm
PH	PH of Soil
EC	Electrical Conductivity, deciSiemens per meter (dS/m)
OC	Organic Carbon %
S	Sulphur, ppm
Zn	Zinc, ppm
Fe	Iron, ppm
Cu	Copper, ppm
Mn	Magnesium, ppm
B	Boron, ppm
Output	(Low(0), Moderate(1), High(2))

Table1 : Attribute Description

Data Visualization

Dataset Visualization helps us to understand data more and eliminate any anomalies from dataset which can cause the unexpected effects on outcome . to view data we use bar graphs as they can help in visualization of data and help us to understand if data is clean or need any processing.

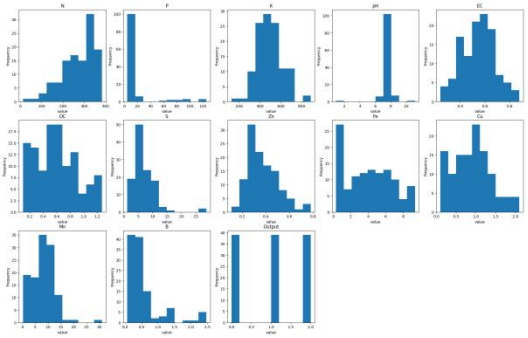


As we have visualize the dataset we get to know that the output is a multi-class data '0,1,2' which is unbalanced as the lower amount of data is present for the '2'. we need to balance the data before training our model for better output.

There are two ways of making the dataset balanced either we decrease the data elements to a minimum or create more sample data on the basis of previous data. Both methods have pros and cons. Lets see both the different methods.

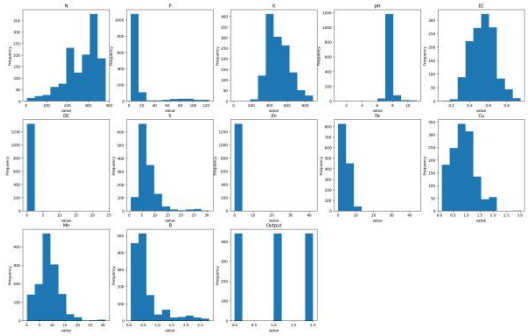
1) UnderSampeling

In under Sampling we reduce the data in dataset to a minimum number which balances the dataset and creates a better training model but in this data set we only had 880 columns and after undersampeling we were left with only 117 unique data which is very low and can lead to less reliable outcome with the trained model predictions.



2) OverSampeling (SMOTE)

In OverSampeling we create sample data on the basis of previous data for balancing the dataset. After the Oversampeling the dataset we get the 1320 unique data which can help to get a better output of trained models.



Now as we have visualized the data and stabilized the data we can train various models to get the result with accuracy.

TRAINING A MODEL

Analysing soil fertility using machine learning models can be a valuable approach for determining suitable cropping patterns in a specific area. By leveraging the power of data and algorithms, these models can provide insights into soil conditions and

help optimize agricultural practices.the models we are using are :

Linear Regression (Logistic Regression):

Linear regression helps in understanding the relationship between independent variables (such as soil chemical content, pH, organic carbon) and the target variable (soil fertility). By establishing a linear relationship, it can provide insights into which soil attributes are most influential in determining soil fertility. This information can guide farmers in optimizing nutrient management practices, adjusting fertilizer application rates, and maintaining the ideal soil chemical balance.

RandomForest

Random Forest is an ensemble model that combines multiple decision trees to make predictions. It can handle both regression and classification tasks related to soil fertility. Random Forest excels at capturing complex interactions and non-linear relationships between soil attributes and fertility. It can identify the most important features contributing to soil fertility and provide guidance on which factors to focus on for improving fertility. This information can aid in the selection of suitable cropping patterns, soil amendments, and management practices.

K-Nearest Neighbours (KNN)

KNN is a classification algorithm that assigns a class label to a data point based on the majority class labels of its nearest neighbours. In the context of soil fertility, KNN can help identify similar soil samples in terms of chemical content, pH, and other attributes. Farmers can leverage KNN to find areas with similar soil fertility characteristics and study the corresponding cropping patterns and management practices in those areas. This information can guide decision-making regarding crop selection and the implementation of appropriate practices for soil improvement.

Decision Tree

Decision trees are versatile models that can handle both classification and regression tasks. In the context of soil fertility, decision trees can be used to understand the hierarchy of factors influencing soil fertility and their interactions. By splitting the dataset based on different features, decision trees can identify the most important soil attributes for predicting soil fertility. This knowledge can help farmers prioritize their efforts in managing soil health and fertility.

Artificial Neural Network (ANN)

ANN is a powerful model inspired by the human brain's neural networks. It can learn complex patterns and relationships in the data, making it suitable for soil fertility prediction. ANNs can uncover non-linear relationships between soil attributes, such as chemical content, pH, and organic carbon, and soil fertility. By analysing large and diverse datasets, ANNs can provide accurate predictions and identify intricate soil-plant interactions, aiding in the optimization of agricultural practices.

Support Vector Machine (SVM)

SVM is a robust machine learning algorithm used for classification and regression tasks. In the context of soil fertility, SVM can assist in identifying decision boundaries that separate different soil fertility classes based on soil attributes. It can handle high-dimensional data and non-linear relationships effectively. SVM can be valuable in predicting soil fertility levels and assisting in the selection of appropriate crop management strategies.

XGBoost

XGBoost is a powerful gradient boosting framework known for its performance on structured and tabular data. By combining weak prediction models, typically decision trees, XGBoost can provide accurate predictions of soil fertility. It can handle complex interactions between soil attributes and predict soil fertility levels more accurately. XGBoost is useful for feature

importance analysis, enabling farmers to identify critical soil attributes affecting fertility and make informed decisions on soil management practices.

Result

As we focus on maintaining Soil fertility we need to get the highly positive results from the machine learning model to create a result.the scores of these following models performance are listed in table.

ML MODELS	ACCURACY % NORMAL DATA	ACCURACY % UNDERSAMPELING	ACCURACY % OVERSAMPELING (SMOTE)
Logistic Regression	87.72%	63.33%	72.12%
Random Forest	45.90%	36.66%	30.30%
K-Neatest Neighbors	80.45%	63.33%	77.57%
Decision Tree	90.45%	96.66%	88.78%
Artificial Neural Network	85.45%	30%	52.12%
Support Vector Machine	83.63%	63.33%	88.18%
XG Boost	89.09%	86.66%	94.54%

As we see the most accuracy is obtained by Decision Tree in undersampeling of dataset but it possibilities of under-fitting is keep occurring in undersampeling dataset so we need to consider the oversampled data where XG Boost model provides 94.54% highest in all of other models so can consider it .

Conclusion

The vast amount of data that is now being collected alongside agricultural crops holds immense potential and must be thoroughly analysed to extract its full value. In my research, I have discovered that XG Boost, with its accuracy of 94.54%, is an excellent choice as a foundational algorithm for predicting soil fertility. By incorporating additional meta-algorithms such as attribute selection and boosting, we can construct a powerful and reliable predictive model. This comprehensive approach harnesses the strength of various techniques to create a highly effective and accurate solution.

REFERENCES

1) Swapna, B., Manivannan, S., & Kamalahasan, M. (2022). Prognostic of soil nutrients and soil fertility index using machine learning classifier techniques. International Journal of e-Collaboration (IJeC), 18(2), 1-14.

2) Janvier, N. I. Y. I. T. E. G. E. K. A., Arcade, N., Eric, N. G. A. B. O. Y. E. R. A., & Jean, N. (2021). Machine learning based soil fertility prediction. International Journal of Innovative Science, Engineering & Technology, 8(7), 141-146.

3) Koley, S. (2014). Machine learning for soil fertility and plant nutrient management using back propagation neural networks. Shivnath Ghosh, Santanu Koley (2014)“Machine Learning for Soil Fertility and Plant Nutrient Management using Back Propagation Neural Networks” International Journal on Recent and Innovation Trends in Computing and Communication, 2(2), 292-297.

4) Keerthan Kumar, T. G., Shubha, C. A., & Sushma, S. A. (2019). Random forest algorithm for soil fertility prediction and grading using machine learning. Int J Innov Technol Explor Eng, 9(1), 1301-1304.

5) Janmejay Pant, Pushpa Pant, R. P. Pant, Ashutosh Bhatt, Durgesh Pant, Amit Juyal, Soil Quality Prediction for Determining Soil Fertility in Bhimtal Block of Uttarakhand (India) Using Machine Learning, Int. J. Anal. Appl., 19 (1) (2021), 91-109.

6) “KAGGLE,”https://www.kaggle.com/datasets/rahuljaiswalonkaggle/soil-fertility-dataset .

7) H. Zheng, J. Wu, and S. Zhang, “Study on the spatial variability of
farmland soil nutrient based on the kriging interpolation,” 2009 Int. Conf.

Artif. Intell. Comput. Intell. AICI 2009, vol. 4, pp. 550–555, 2009, doi: 10.1109/AICI.2009.137.

8) Prince Patel, “Why Python is the most popular language
used for Machine Learning,” 2018. [Online]. Available: https://medium.com/@UdacityINDIA/why-use-python-for-machine
learning-e4b0b4457a77. [Accessed: 20-Mar-2020].

9) N. Gupta, “Why is Python Used for Machine Learning?,” 2019. [Online].Available:https://hackernoon.com/why-python-used-for-machine learning-u13f922ug. [Accessed: 20-Mar-2020].

10) “Soil test”,Wikipedia,June 2012.

11) Vamanan R. & Ramar K. (2011) “CLASSIFICATION OF GRICULTURAL LAND SOILS A DATA MINING APPROACH”; International Journal on Computer Science and Engineering (IJCSE); ISSN: 0975-3397 Vol. 3

12) Witten I. and Eibe F. (2005), “Data Mining: Practical Machine Learning Tools and Techniques” 2nd Edition, San Francisco: Morgan Kaufmann