Proceeding Paper

Enhancing Process Control in Agriculture: Leveraging Machine Learning for Soil Fertility Assessment [†]

Ashutosh Sarangi [ID], Sailesh Kumar Raula, Sohamdev Ghoshal, Swadhin Kumar, Chinta Sai Kumar and Neelamadhab Padhy *[ID]

School of Engineering and Technology, Department of Computer Science and Engineering, GIET University, Gunupur 765022, Odisha, India; ashutoshsarangi328@gmail.com (A.S.); 21cseds001.saileshkumarraula@giet.edu (S.K.R.); 21cseds.sohamdevghoshal@giet.edu (S.G.); 21cseds062.swadhinkumar@giet.edu (S.K.); 21cseds004.chintasaikumar@giet.edu (C.S.K.)
* Correspondence: dr.neelamadhab@giet.edu
† Presented at the 3rd International Electronic Conference on Processes—Green and Sustainable Process Engineering and Process Systems Engineering (ECP 2024), 29–31 May 2024; Available online: https://sciforum.net/event/ECP2024.

Abstract: Context: The primary factor in determining whether or not a crop can be produced on a certain type of soil is soil fertility. When faced with many options, farmers frequently find it difficult to decide which crop to plant. We created this project to address that particular issue. The provision of soil data is mandatory since it will significantly influence the determination of the soil's fertility. The output and accuracy of the model may suffer if the data are not supplied discretely. The nature of the dataset indicates that the result is a binary value, i.e., either "Fertile" or "Non-Fertile", along with the accuracy percentage of each algorithm. **Objective**: The main aim of this paper is to determine whether the soil is fertile based on soil properties like N, P, K, Ph, nutrient level, moisture levels, temp, rainfall, and topography. **Material/Method**: We used the dataset from Kaggle, where N, P, K, and pH values are input into the model, and the ML determines whether it is fertile or not. In this article, four machine learning classifiers are trained, and determine the best classifier based on the performance metrics. **Result**: The results demonstrated that the machine learning classifier significantly improves prediction accuracy. We used LR, KNN, NB, and DT classifiers to increase the accuracy, as well as to increase the efficiency of the soil fertility assessment. The DT classifier exhibited well in comparison to other classifiers. The DT classifier's accuracy was 89%, but the performance metrics precision, LR, and KNN, was 90%.

Keywords: soil data; soil analysis; soil constituents; machine learning; performance analysis

1. Introduction

India is largely a tropical country; hence, the organic carbon content (OC) of the soils is low. Most of the nitrogen (N) in the soil is present in organic forms; thus, the faster decomposition of this organic matter leads to a release of nitrogen, which is removed by the plants or gets wasted without any proper use. Therefore, the deficiency of nitrogen (N) is universal in India. Most Indian soils have low to medium phosphorus (P) content. Due to intensive cropping and no replenishment, potassium (K) deficiency has also become widespread in the country. Among the micronutrients, the deficiency of zinc is the most acute, followed by boron. Therefore, the deficiency of nitrogen (N) is universal in India.

In the field of agriculture and gardening, the use of soil fertility is for maximizing crop productivity and ensuring the sustainable use of land. There are various factors influencing soil fertility and among them are the three essential nutrients: nitrogen (N), phosphorus (P), and potassium (K), which hold a particular level of significance. These nutrients play a major and vital role in the growth and development of plants. Nitrogen plays a part in leafy growth and overall plant vigor, whereas phosphorus plays a part in root development and

flowering, and potassium helps with water regulation and disease resistance. NPK values are often used to indicate soil fertility, as by measuring the levels of these key nutrients, farmers and researchers can analyze the quality and health of the soil and make decisions regarding fertilizers and soil management practices. It is important to recognize that soil fertility is a concept influenced by numerous other factors beyond NPK, including soil pH, organic matter content, and microbial activity.

In the past years, machine learning algorithms played a vital role in various domains of study, which ranged from healthcare to finance and from marketing to cybersecurity. As such, the accuracy provided by machine learning algorithms built into the models significantly impacts the decision-making process of these domains. To overcome these challenges, techniques such as the Synthetic Minority Over-Sampling Technique (SMOTE) and ensemble methods came into the picture, which gained a huge amount of attention. SMOTE focuses on balancing the class by generating synthetic samples for minority classes to increase the accuracy of machine learning algorithms, whereas if we talk about the ensemble methods like Random Forest (RF), Gradient Boosting Machines (GBM), and various others, they are used to improve prediction accuracy and robustness.

This case study aims to reach a site of exploration where the effects of SMOTE and ensemble methods are playing a role in enhancing the accuracy of ML algorithms, particularly in phases where we could not obtain a balanced dataset. We will check how the machine learning algorithms help us face a common problem, which is dealing with a dataset that is not balanced. In this introduction, we offer an overview of SMOTE and ensemble methods, in which we discuss each algorithm's contribution to machine learning performance.

Soil fertility refers to the capacity of soil fertility to provide essential plant nutrients, and it also includes some additional nutrients like nitrogen, phosphorous, and potassium. It also ensures that crops receive the essential elements for growth. On the other hand, it also plays a major role in producing crops. It promotes higher agricultural productivity, contributing to food availability and accessibility. Since the increasing population there limits the area available to cultivate, it is very important to increase soil fertility to ensure food security. Fertile soil requires less use of pesticides and fertilizers, and methods like cover crops and crop rotation enhance soil fertility naturally, which is equally beneficial for both the environment and human health. Good fertility is more stubborn to the impacts of climate change, such as droughts and extreme weather conditions, because of the better water-holding capacity of fertile soil. It is very necessary to keep the soil fertile, adding an amount of good compost or fertilizers to refresh the nutrients and ensuring the plants flourish with healthy components, minerals, and nutrients. That ensures healthy plants and a chance to flourish.

Nowadays, agriculture plays a vital role in India, where the majority of people depend on it. Soil fertility is one of the important factors for enhancing productivity, as well as improving sustainability. In earlier days, farmers used to produce the crops using the traditional approaches. However, in the 21st century, farmers are well equipped to handle the agricultural farm with the latest technologies. The role of AI (especially machine and deep learning) dominates in agribusiness for enhancing the accuracy and efficiency of soil fertility evaluations. In this paper, our main objective is to enhance the model's accuracy for soil fertility assessment. Our paper also discusses the limitations of traditional soil fertility evaluation techniques and highlights the advanced machine learning algorithms to handle these issues. We demonstrate how machine learning classifiers such as K-Nearest Neighbors (KNN), Decision Trees (DT), naive Bayes (NB), and logistic regression (LR) are used to improve the precision and reliability of soil fertility predictions. This approach not only handles precision agriculture but also provides insights for better soil management. Our research work is novel and significant, as it finds the gap between theoretical research and practical applications in soil management. By utilizing various ML techniques, we strive to provide comprehensive and innovative solutions of soil fertility management that can be applied in the real-time agricultural context. By integrating various machine

learning techniques, we strive to offer a comprehensive and innovative solution for soil fertility assessment that can be applied in real-world agricultural contexts.

2. Literature Review

Many papers have undergone the processes of soil fertility identification by identifying constituents of soil. The study in ref. [1] presented a Random Forest classifier, as it offered the best result or greater accuracy, as compared to other classifiers considering the dataset. The main objective of this was to build a robust model that offers a high accuracy score to obtain the required type of soil when different types of soil are considered for growing a particular crop and to help farmers maximize their yield using that soil. In the study from [2], Random Forest offered the best accuracy, which was almost 100% accuracy, compared to other models, such as neural networks, SVM, and naïve Bayes, considering the dataset, which was divided into three parts, i.e., crop dataset, soil dataset, and yield dataset, in which mainly N, P, and K values are taken for calculating the accuracy. The Random Forest classifier used here was an efficient model for the particular dataset, but it may not be the same for other datasets that have larger sample sizes; in that case, neural networks can be used, along with some hyperparameter tuning to obtain a more precise accuracy score.

The algorithms used were Random Forest (RF), support vector machine (SVM), ANN, K-NN, and some regression-based models [3]. The objective of the above machine learning model was to predict soil properties, such as chemical properties, and nutrients and then recommend fertilizers using the ML algorithm. The accuracy was between 85% and 91%, as the SVM, RF, and ANN contributed high accuracies, and some improvements can be made using the algorithms specifically, as there were many other algorithms used too. Also, performance evaluation should be worked on. To address the low accuracy of standard models, this work developed a novel prediction model based on big data statistics and near-infrared (NIR) analysis, resulting in a threefold improvement in accuracy [4]. However, there was room for development by including external elements in data gathering and improving geographic applicability. The study intended to create an accurate forecast model for soil nutrient concentration by combining the improved genetic algorithm (IGA) with back-propagation (BP) neural network models [5]. This refined approach increased the accuracy of soil nutrient prediction, with coefficients of determination (R2) better than 0.88 for a variety of nutrients. However, there is room for improvement by using advanced techniques, such as convolutional neural networks (CNNs), to increase the model's predictive capabilities. The study examined soil spectral reflectance at various moisture levels to better understand moisture's impact on soil spectral data [6]. It created prediction models for soil nitrogen, phosphorus, and potassium utilizing hyperspectral data and algorithms, such as PLS, SVR, Si-PLS, and Si-SVR. The Si-SVR model had the highest accuracy, with RPD values greater than 2.0 for all nutrients, including 2.86 for nitrogen prediction, but it might be improved by integrating soil pH, texture, and organic matter content. The study in ref. [7] discussed how the crop can be recommended using machine learning. The authors also developed the IoT-enabled soil nutrient analysis. They collected the data from different sources.

The study in ref. [8] discussed how machine learning and deep learning are useful for handling soil fertility management. They used both ML and Dl classifiers and found that NB and CNN outperformed, i.e., 99% and 87%. The study in ref. [9] used both data mining and machine learning classifiers to exhibit soil fertility management. They found an accuracy of 0.87 in Bayesian classification. The study in ref. [10] discussed how ML classifiers are used for soil fertility 0 and recommended the crop. They used several classifiers, including KNN, RF, NB, Lasso, Ridge, etc. Out of them, KNN and RF obtained 83% and 82% accuracies. The study in ref. [11] used ANN techniques for soil fertility management. They estimated the performance measures like MSE and R2. Both the training and testing accuracies they obtained were 0.96. The study in ref. [12] discussed the SMOTE technique for crop recommendation. They used before SMOTE and after SMOTE to check whether or not

there was a possibility to enhance the accuracy. They also discussed in the paper how to handle soil fertility management, but they primarily focused on the recommendation system. The study in ref. [13] discussed the regression approach, as well as conducted the hypothetical test for soil organic prediction. They used machine learning regression classifiers like LR, ENT, MLR, Ridge, Lasso, etc., and found that RF performed well and was about 0.74 R-square.

The below mentioned Table 1 discusses how the proposed model is better than the existing one for soil fertile management.

Table 1. Comparison with other state-of-the-art approaches.

Ref#	Existing Approaches	Our Approach
[8]	Soil fertility assessment and crop recommendation. They used ML and DL techniques and obtained an accuracy of 87%.	We used only machine learning classifiers like KNN, DT, and NB. Our proposed model exhibited an 89% accuracy for DT.
[9]	They used the following techniques for soil fertility management: 1. Data mining; 2. The machine learning classifier. They found an accuracy of 0.87 in Bayesian classification.	Our model discussed only the machine learning classifiers, and DT was found to be the most prominent classifier for soil fertility management.
[10]	They used machine classifiers, including KNN, RF, NB, Lasso, Ridge, etc. Out of them, KNN and RF obtained 83% and 82% accuracies.	Our paper discussed only the machine learning classifiers and found that the accuracy of DT was 89%. We also found that the precision and recall of logistic regression and KNN was 90%.
[11]	They used several regression approaches to predict the soil's organic quality. The RF performed well, and it was about 0.74 R-square.	Our model obtained 89% accuracy for DT and 90% precision and recall for KNN and NB classifiers. Our approach was unique in comparison to the mentioned state-of-the-art method.

3. Proposed Model

In Section 3, we proposed one model that assessed whether the soil was fertile or not. The entire research was conducted in the Gunupur Agriculture Department, Odisha, during the period from January 2023 to December 2023. The study was conducted under the supervision of the agriculture department, and they provided the necessary facilities and resources for data collection.

3.1. Model Overview Our Proposed Model Demonstrated the Machine Learning Classifiers to Enhance the Accuracy and Reliability of Soil Fertility Prediction

1. Logistic regression (LR): This classifier is a supervisor machine learning classifier, and it is used to analyze the relationship between soil properties and fertility metrics.
2. K-Nearest Neighbors (KNN): This classifier is suitable for classification. In our dataset, the class label was a binary class. Hence, we used it for the binary class classification solution.
3. Decision Trees (DT): This classifier is suited to the model for the decision-making process for soil fertility classification.
4. Naive Bayes (NB): Similarly, this classifier is suitable for probabilistic classification based on feature independence assumptions.

Figure 1 demonstrates the complete procedure for identifying whether or not the soil is fertile.
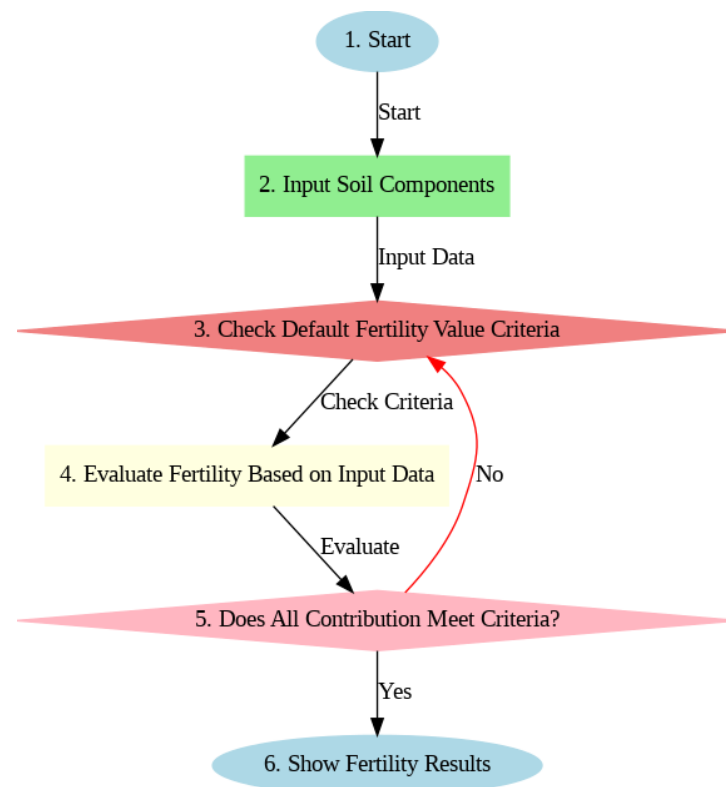
Figure 1. Rough sketch of soil fertility management.

The above flowchart works as follows:

(i) **Start**: The process begins here.
(ii) **Input Soil Components**: At this stage, various soil components are inputted for analysis.
(iii) **Check Fertility Value Criteria**: The inputted soil components are checked against predefined fertility value criteria.
(iv) **Evaluate Fertility based on Input Data**: Based on the criteria check, the fertility of the soil is evaluated.
(v) **Do All Inputs Contribute**: This decision point determines if all the inputted components contribute to fertility.

If **Yes**: The process moves forward to the final step.
If **No**: It implies that the evaluation needs to be revisited with the input data.

(vi) **Show the Fertility**: This is the final step where the fertility of the soil is displayed or reported.

Figure 2 demonstrates how the machine learning classifiers work in stage 4.

3.2. Research Context and Setup

We conducted the study in the year 2023 (January to December), and the location was Gunupur, India. The entire research was carried out at the research unit of GIET University, Gunupur, Odisha, India, where the study was performed. We collected the data and tested our model.

Our proposed model's objective was to check whether the soil was fertile or not, cohesively, using the machine learning classifiers. The data used in this paper were collected from January 2023 to December 2023. We estimated the performance of the model using the evaluation of metrics.
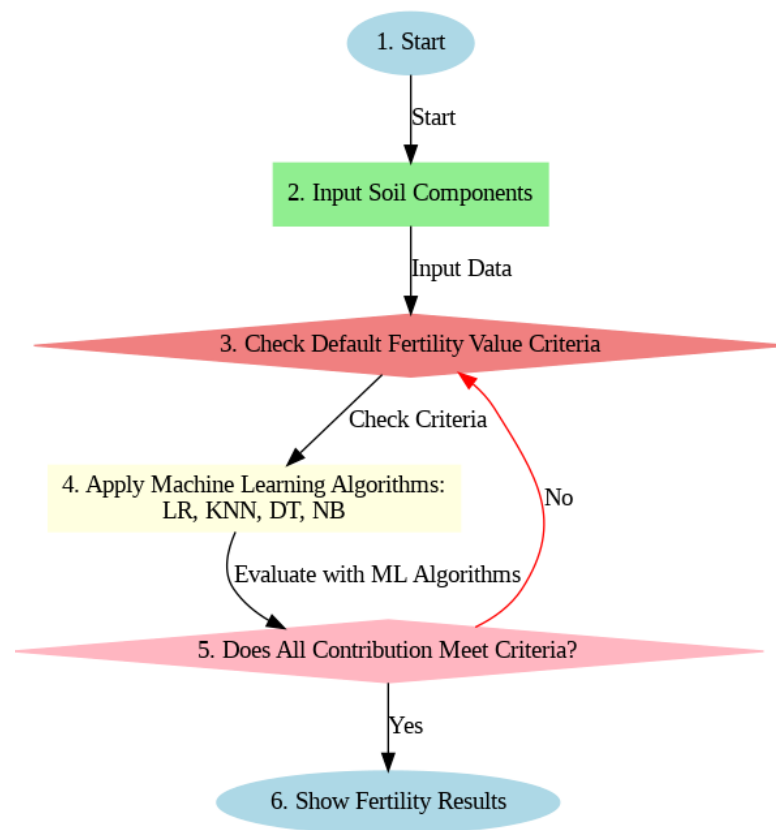
Figure 2. Proposed model for soil fertility management.

4. Results and Discussion

Table 2 presents the analysis of the classification models. Here, we portrayed the performance matrix of accuracy precision and recall values of each algorithm used in the model before implementing the SMOTE technique. The table below demonstrates the tabular results of the classifier.

Table 2. Performance metrics of different classifiers.

Algorithms	Accuracy	Precision	Recall
Logistic Regression	0.83	0.90	0.83
KNN	0.88	0.90	0.88
NB	0.77	0.89	0.77
Decision Tree	0.89	0.89	0.88

RQ: How do the different machine learning classifiers work for soil fertility based on the different performance metrics (especially accuracy, precision, and recall)?

RQ: How do you measure and impact the performances of all the machine learning classifier training data for soil fertility observed in the learning curve?

The solution to the research question: The above research question can be addressed using the learning curve. It discusses the performance of each machine learning classifier training data. It obtains the training and cross-validation scores against each training sample.

Figures 3–5 depict the learning curves that indicate the learning during the training of a machine learning model. We plotted the graph and demonstrated the performances of the classifiers. We measured the training score, as well as the cross-validation scores, to obtain the progress of the classifiers.
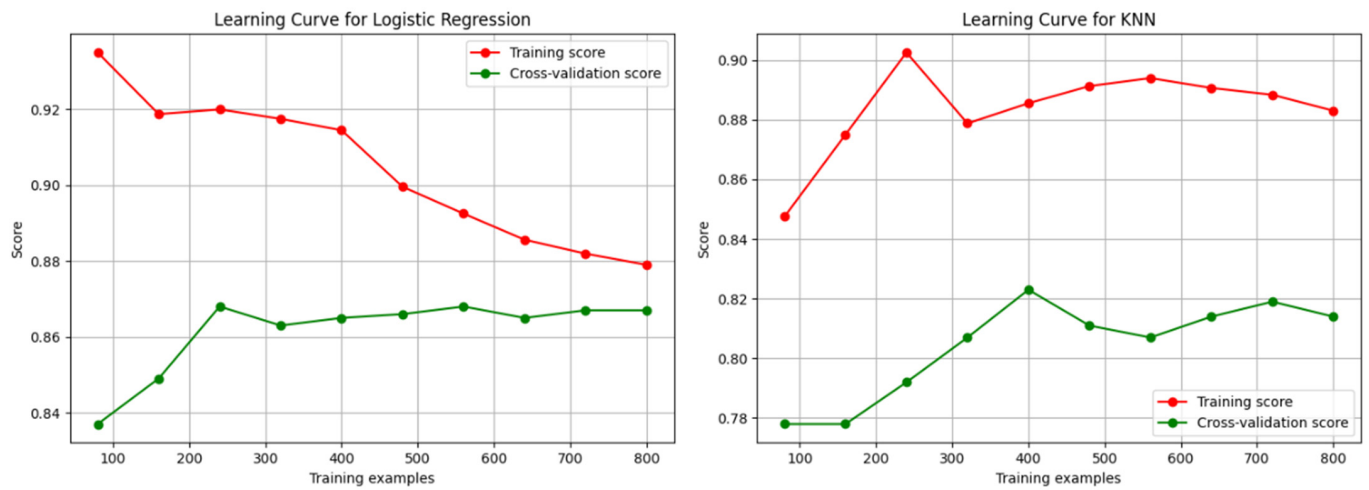
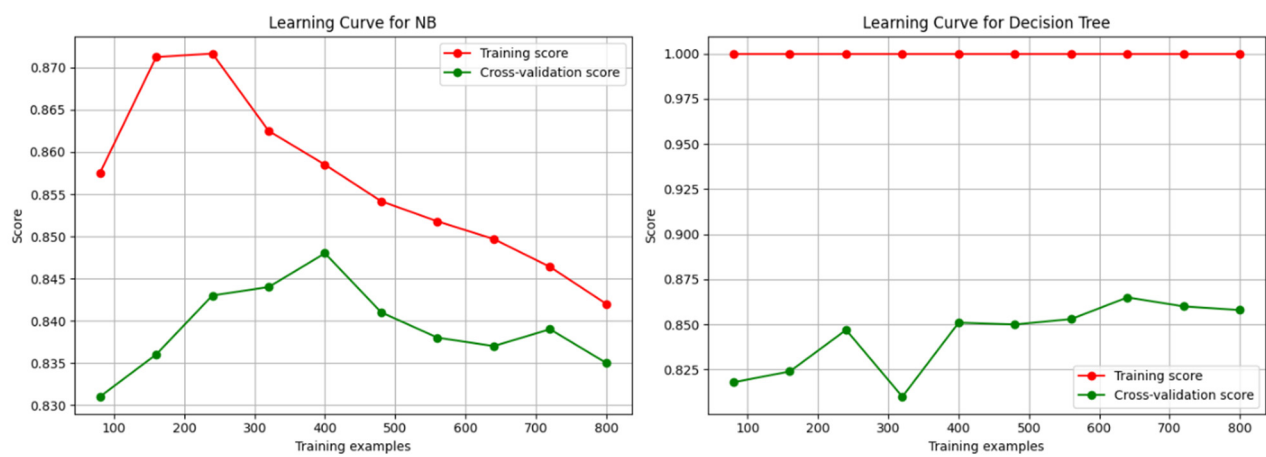**Figure 3.** The learning curves of logistic regression and KNN.

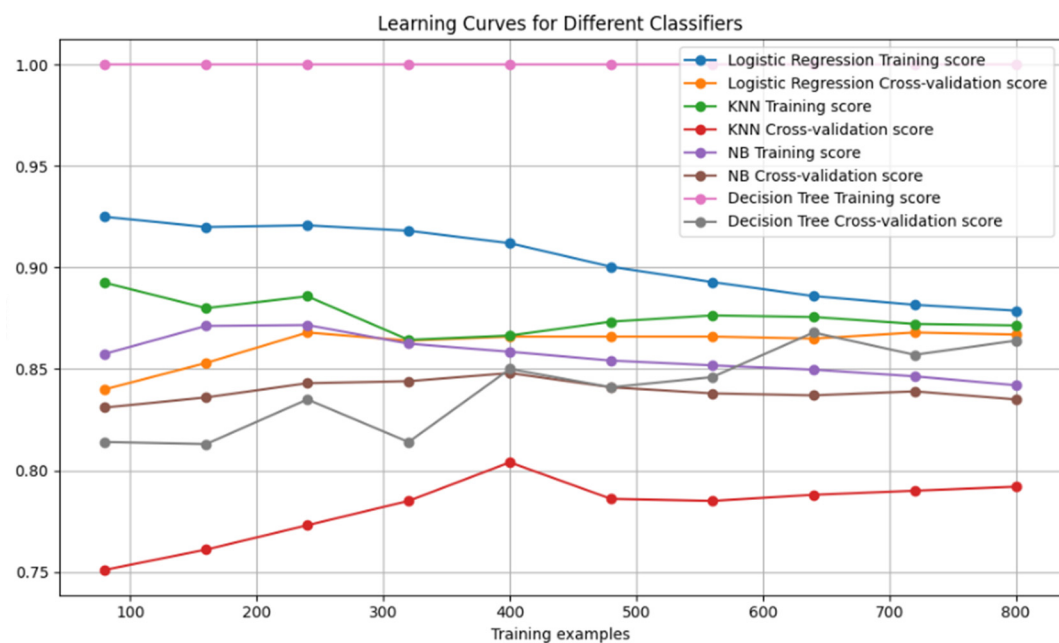**Figure 4.** The learning curves of NB and KNN classifiers.

Figure 5. The learning curves for all the classifiers.

RQ: Which algorithm performs well for soil fertility?

Solution to RQ: The above RQ discussed the performance of different classifiers. We have done an extensive comparison of the performance metrics and demonstrated which one performs well. The below-mentioned Figure 6 depicts the performances of all the algorithms used in our results. The results of LR and DT provide the prominent results compared to others.
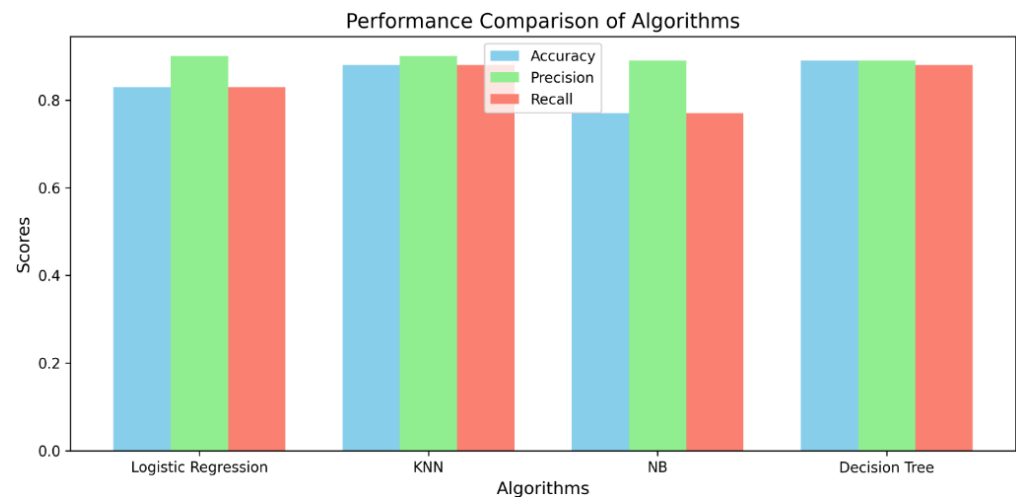
Figure 6. Algorithm performances.

The highest accuracy and precision of the DT was 0.89, and the recall was 0.88. Similarly, The highest accuracy and precision of the Decision Tree (DT) were 0.89, with a recall of 0.88. Similarly, the highest precision was achieved by both the Logistic Regression (LR) and K-Nearest Neighbors (KNN) classifiers.

In Figure 7, we created the boxplot visualization for all the algorithms that we used in our model. We identified some of the algorithms that performed well and some that did not.
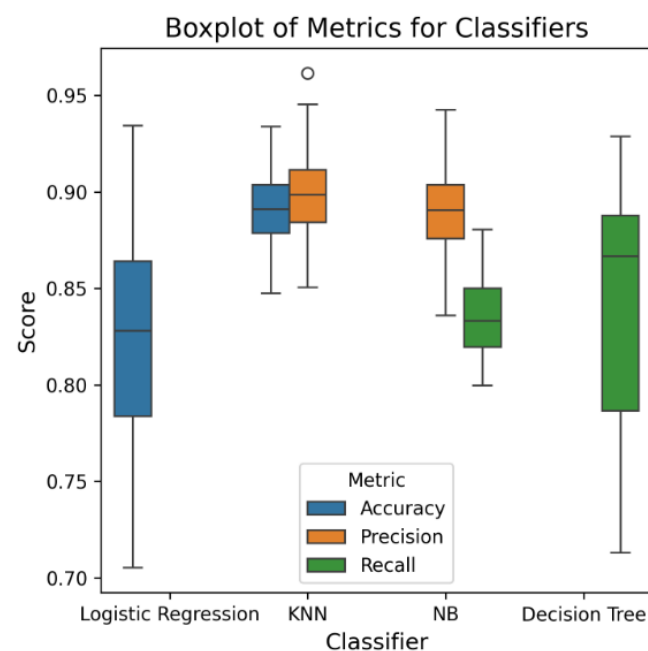
Figure 7. Boxplot representation of different metrics.

Figure 8 shows the outlier detection from the dataset and indicates a box plot for outlier detection. The box plot is another graph that was used for outlier detection, which offered a pretty clear understanding of the outliers present in the dataset, as one can see the outliers clearly. (The outliers were present outside the lower and upper whiskers.)
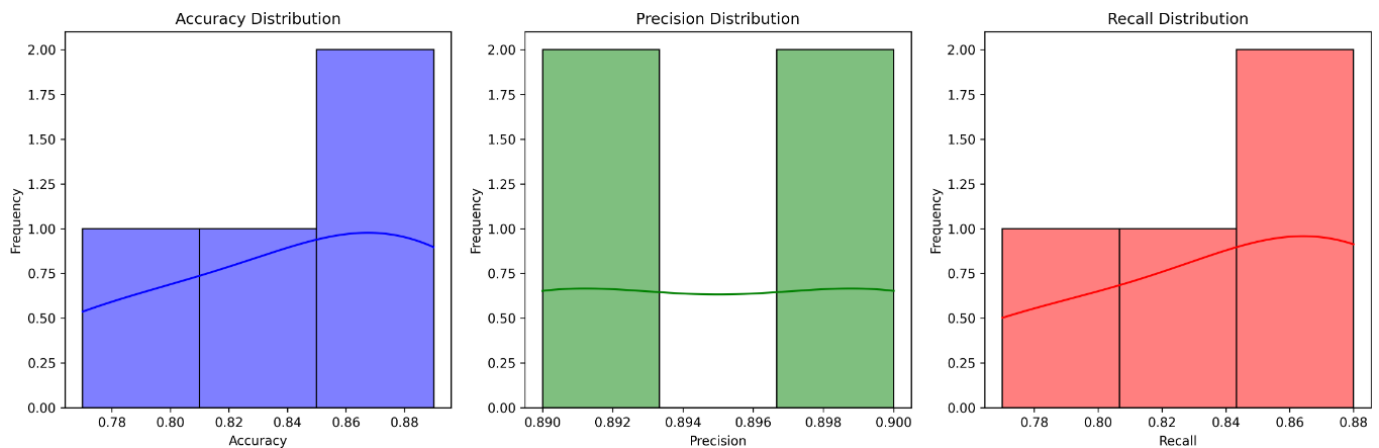

Figure 8. Outlier detection.

Table 3 and Figure 9 show a counter plot graph. This is a counter plot representing the count of each output, whether the soil was "Not Fertile, Medium Fertile, or Highly Fertile", represented as 0, 1, or 2 on the X-axis. Figure 10 shows a correlation matrix that presented the relation between each feature with others, signifying how strong and in what direction two or more variables were related. (Here, in the matrix above, we can see that nitrogen (N) had a relation of 0.12 with boron (B), which is a strong relationship.) It was similar to all other features.

Figure 11 shows a feature importance plot of the data. The graph above is a feature importance plot that shows the importance of a particular feature in the dataset analysis. In the graph above, nitrogen (N) was the most important feature for predicting soil fertility, followed by phosphorus and so on.
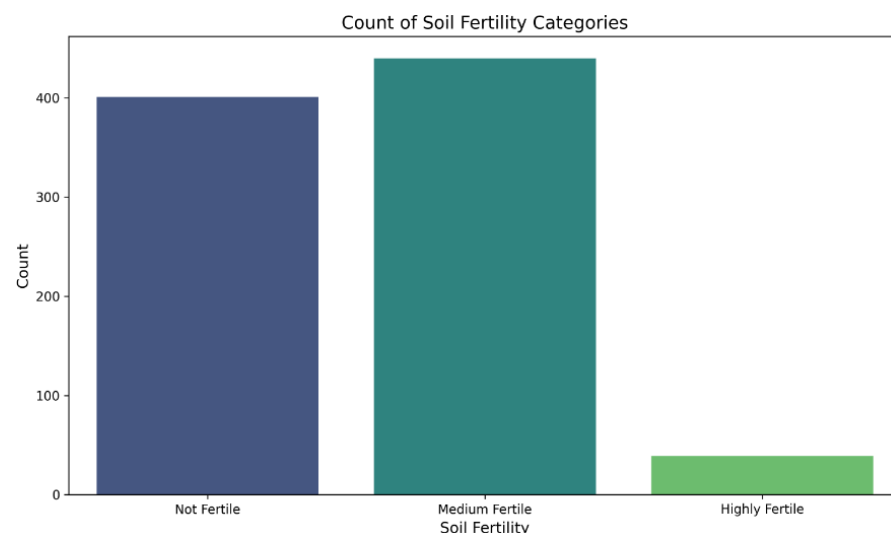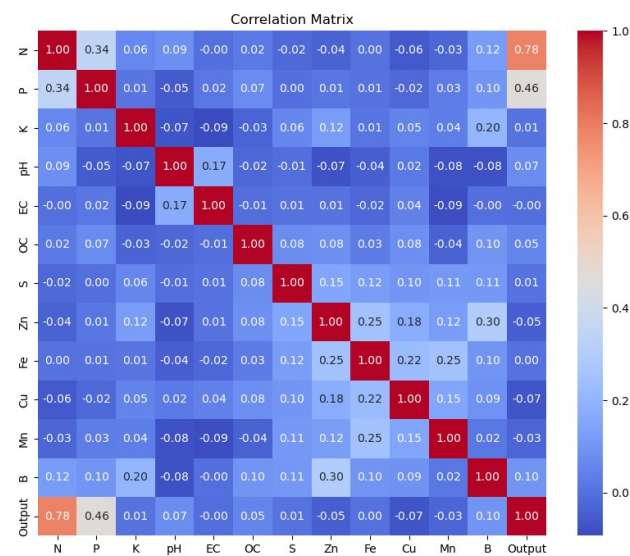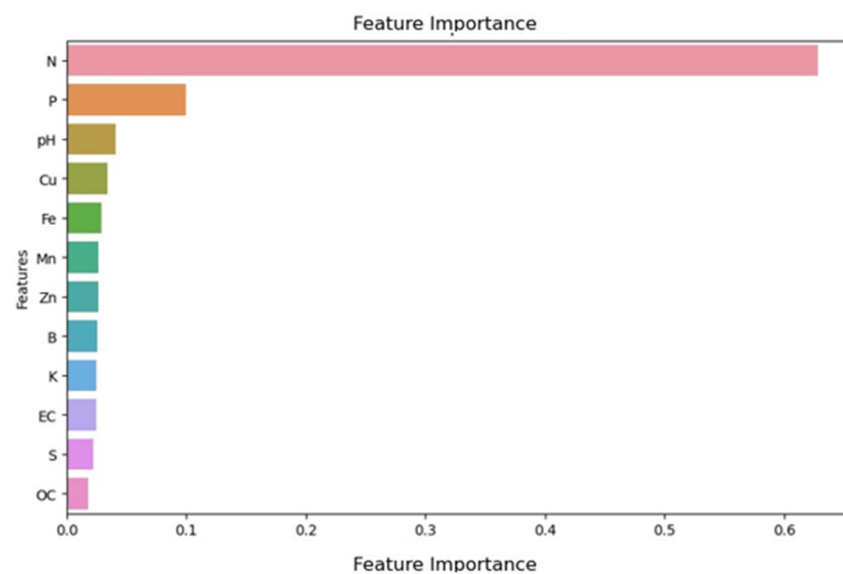

Figure 9. Fertility check.

Table 3. Verifying the data.

	Accuracy	Precision	Recall
Count	4.0000	4.0000	4.0000
Mean	0.84	0.89	0.84
Std	0.05	0.005	0.05
Min	0.77	0.89	0.77
25%	0.81	0.89	0.81
50%	0.85	0.89	0.85
75%	0.88	0.9	0.88
Max	0.89	0.9	0.88

Figure 10. Correlation plot of the data.

Figure 11. Feature importance plot.

5. Conclusions

In the study above, research on identifying soil fertility through machine learning algorithms presented several avenues for enhancement and expansion in agricultural

practices. We did not perform the machine learning classifier to detect whether the soil can be fertile or not. Here, we considered accuracy as the major criterion. One proposed model was developed to identify the soil type, as well as to predict it. Fine-tuning model architectures, including deep learning and ensemble methods, tailored to specific soil types and environmental conditions, could further enhance performance. Additionally, the development of real-time monitoring systems using IoT devices and sensor networks can enable continuous soil fertility assessment and provide timely recommendations to farmers.

In conclusion, by addressing these challenges and pursuing these avenues, machine learning-based soil fertility prediction models can contribute to optimizing fertilization strategies, promoting sustainable agriculture, and ensuring food security in the future.

References

1. Keerthan Kumar, T.G.; Shubha, C.A.; Sushma, S.A. Random forest algorithm for soil fertility prediction and grading using machine learning. *Int. J. Innov. Technol. Explor. Eng.* **2019**, *9*, 1301–1304.
2. Yadav, J.; Chopra, S.; Vijayalakshmi, M. Soil analysis and crop fertility prediction using machine learning. *Mach. Learn.* **2021**, *8*, 41–49.
3. Motia, S.; Reddy, S.R.N. Exploration of machine learning methods for prediction and assessment of soil properties for agricultural soil management: A quantitative evaluation. *J. Phys. Conf. Ser.* **2021**, *1950*, 012037. [CrossRef]
4. Cui, Y.; Yue, Q. Regional soil nutrient content prediction model based on big data. *J. Phys. Conf. Ser.* **2023**, *2555*, 012005. [CrossRef]
5. Liu, Y.; Jiang, C.; Lu, C.; Wang, Z.; Che, W. Increasing the accuracy of soil nutrient prediction by improving genetic algorithm backpropagation neural networks. *Symmetry* **2023**, *15*, 151. [CrossRef]
6. Yu, S.; Bu, H.; Dong, W.; Jiang, Z.; Zhang, L.; Xia, Y. Construction and evaluation of prediction model of main soil nutrients based on spectral information. *Appl. Sci.* **2022**, *12*, 6298. [CrossRef]
7. Senapaty, M.K.; Ray, A.; Padhy, N. IoT-enabled soil nutrient analysis and crop recommendation model for precision agriculture. *Computers* **2023**, *12*, 61. [CrossRef]
8. Vandana, W.M.; Kavya, B. Soil fertility assessment and crop recommendation for sustainable farming using machine learning and deep learning. In Proceedings of the 4th International Conference on Data Engineering and Communication Systems (ICDECS), Bangalore, India, 22–23 March 2024; Volume 2024, pp. 1–3. [CrossRef]
9. Inoyatova, M.; Ziyadullaev, D.; Muhamediyeva, D.; Aynaqulov, S.; Ziyaeva, S. Data mining for assessing soil fertility. *E3S Web Conf.* **2024**, *494*, 02012. [CrossRef]
10. Asif, M.; Wahid, A. Leveraging Machine Learning for Soil Fertility Prediction and Crop Management in Agriculture. Available online: https://www.researchsquare.com/article/rs-4310747/v1 (accessed on 2 June 2024).
11. El Behairy, R.A.; El Arwash, H.M.; El Baroudy, A.A.; Ibrahim, M.M.; Mohamed, E.S.; Rebouh, N.Y.; Shokr, M.S. An accurate approach for predicting soil quality based on machine learning in drylands. *Agriculture* **2024**, *14*, 627. [CrossRef]
12. Senapaty, M.K.; Ray, A.; Padhy, N. A decision support system for crop recommendation using machine learning classification algorithms. *Agriculture* **2024**, *14*, 1256. [CrossRef]
13. Swain, V.K.; Padhy, N.; Ray, T.; Biswal, S.; Patra, A.; Viswaroopanand, B.S.; Sahu, K.K.; Baral, A. Agricultural Pest Classification Using Transfer Learning: A Process Control and Monitoring Perspective. *Proceedings* **2024**, *105*, 57. [CrossRef]