



A Machine Learning approach for crop and fertilizer using Ensemble Model with XGBoost and Random Forest Algorithms

Surya Narayanan K¹, Varun Kumar E B², Srinath J³, Yashvanth H⁴, J.K Jayakumari⁵

^{1, 2, 3, 4, 5}Computer Science and Engineering, Adhiyamaan College of Engineering
(Autonomous), Anna University

ABSTRACT

The machine learning method for crop and fertilizer recommendations is presented in this study in an effort to improve agricultural productivity. By combining the strengths of XGBoost and Random Forest algorithms through an ensemble model, this system provides customized recommendations based on important soil and environmental parameters, such as temperature, humidity, rainfall, pH, nitrogen, phosphorus, potassium, and other indicators for the optimal crop growth and yield. Using a soft voting technique, the model outperforms traditional single-model systems by averaging predictions from both algorithms to attain a high level of accuracy beyond 98%. This ensemble design captures the high accuracy of XGBoost with its data-handling capabilities while balancing it with Random Forest's strength and reduced overfitting and also making the model in many different kinds of agricultural environments.

Precision and computational efficiency were demonstrated by the model, which was trained on a dataset of 3,610 samples with an additional 903 samples set aside for testing. Although it takes longer to train, XGBoost helps by producing extremely accurate predictions. Random Forest helps by training more quickly and maintaining performance even when dealing with many different kinds of data. When combined, these models produce an efficient system that can be adjusted to various soil types and climates, empowering farmers to make data-driven, well-informed decisions for sustainable farming.

This study demonstrates how ensemble learning can revolutionize precision agriculture by assisting farmers in making the best decisions possible to enhance crop yield and resource management while encouraging sustainable practices. Real-time data integration may be incorporated into future research to increase the model's adaptability to shifting agricultural conditions. All things considered, this system improves the use of machine learning in the agricultural sector by offering farmers a strong tool to help them in making profitable and sustainable agricultural decisions.

Keywords: Ensemble model, Machine learning, Crop and Fertilizer recommendation, Sustainable agriculture, XGBoost, Random Forest.

1. Introduction

Crop selection and usage of fertilizers are crucial for maximizing yield and preserving soil health, making agriculture a vital component of the world economy. Unfortunately, traditional methods for recommending crops and fertilizers frequently depend on broad principles that might not take into account particular soil and climate conditions, which results in inefficient



resource use and less than ideal crop cultivation. Machine learning presents a promising solution to the growing need for data-driven and sustainable agricultural practices by providing accurate, personalized recommendations suited to particular agricultural contexts.

The accuracy of predictions for complex tasks has been significantly improved by recent developments in machine learning [1, 2]. Ensemble models which combine XGBoost and Random Forest algorithms have gained attention due to their ability to handle large datasets, reduce overfitting, and balance prediction stability [3]. In this research, we propose an ensemble model that combines the strengths of XGBoost and Random Forest through a soft voting mechanism to improve crop and fertilizer recommendations. This model provides a comprehensive evaluation of soil suitability for crop selection and fertilizer optimization by incorporating important agricultural parameters such as temperature, humidity, rainfall, pH, nitrogen, phosphorus, potassium, and rainfall. The main objectives of this study are to develop a high-accuracy recommendation system and to contribute to sustainable agricultural practices by reducing, this work uses ensemble learning to provide accurate agricultural recommendations, addressing the shortcomings of traditional approaches. To improve adaptability in a variety of agricultural contexts, future advancements might incorporate real-time data and other algorithms.

2. Recent Work

Applications of machine learning in agriculture have made significant improvements, with a particular emphasis on creating reliable models that optimize crop yield and fertilizer suggestions. Ensemble models are becoming more and more popular because of their high accuracy in handling complex agricultural data and producing accurate predictions, especially when XGBoost and Random Forest are combined.

2.1 XGBoost and Random Forest in Agricultural Prediction

Recent research highlights the usefulness of Random Forest and XGBoost in agriculture due to their capacity to simulate complex interactions between crop yield and environmental and soil factors. The gradient boosting framework XGBoost is well-known for handling a variety of features, including soil nutrients and climate conditions, and for being effective with big datasets [4]. In agricultural data, where variables like soil pH, nitrogen, phosphorus, and potassium levels can vary greatly, overfitting is a common problem that Random Forest's ensemble of decision trees excels at handling [5]. Both algorithms have demonstrated a high degree of accuracy in forecasting crop yield and suitability.

2.2 Combining Models with Ensemble Learning

Soft voting the ensemble model have gained popularity as a way to leverage the advantages of both Random Forest and XGBoost. Soft voting in ensemble learning involves averaging each model's prediction probabilities to produce a final recommendation that is accurate and balanced. In order to achieve high accuracy in real-time crop recommendation, Komala Devi K et al. developed an ensemble approach. This shows that ensemble learning performs better



in dynamic agricultural environments [6]. The model's strength and dependability across a variety of datasets are improved in this project by combining predictions from XGBoost and Random Forest using a similar soft voting mechanism.

2.3 Feature Selection for Precision Agriculture

Selecting characteristics that directly affect crop health and yield, such as nitrogen (N), phosphorus (P), potassium (K), pH, temperature, humidity, and rainfall, is crucial, according to recent studies. Such parameters were included in a crop yield prediction model by Bhasha et al., demonstrating that thorough information on soil and climate variables greatly improves model accuracy [4]. Similar agricultural characteristics are employed in this study to guarantee a comprehensive evaluation of crop and fertilizer suitability, providing a comprehensive perspective of soil health.

2.4 Evaluation Metrics

The performance of our model is assessed using metrics such as accuracy, precision, and recall, which is in line with recent studies. The project's high accuracy of 99.6% corresponds with recent research on the effectiveness of ensemble approaches, proving that precision agriculture can be reliably supported by combining Random Forest's stability with XGBoost's accuracy.

3. Proposed Work Explanation

To provide precise crop and fertilizer recommendations based on unique agricultural parameters, we present an ensemble machine learning model in this work that combines the XGBoost and Random Forest algorithms. Ensemble models are excellent at utilizing the complementary strengths of individual algorithms; in this case, Random Forest adds stability and minimizes overfitting by averaging predictions across several decision trees, while XGBoost handles large and varied agricultural data to provide high prediction accuracy.

Ensemble learning has become a standard approach in predictive modeling due to its ability to improve model performance by combining multiple algorithms. By averaging their probability outputs, XGBoost and Random Forest are able to make equal contributions to the final prediction through the ensemble model's soft voting mechanism. For complex agricultural datasets containing variables like nitrogen (N), phosphorus (P), potassium (K), pH, temperature, humidity, and rainfall, this method increases accuracy while preserving model robustness. These characteristics have a direct impact on crop yield and soil fertility, and the model's design enables it to evaluate each characteristic based on how it affects overall prediction accuracy.

This work's computation section combines the prediction probabilities of Random Forest and XGBoost by implementing soft voting in ensemble learning. By averaging each model's outputs, this procedure produces a final recommendation that maintains a balance between accuracy and dependability. The ensemble model is a useful tool for precision agriculture since it not only achieves high accuracy but also shows scalability and adaptability across a range of soil and climate conditions. By offering customized solutions based on empirical data and



particular agricultural needs, this model successfully addresses issues with conventional crop and fertilizer recommendations. The Proposed block diagram for the system has been shown in the Figure 1.

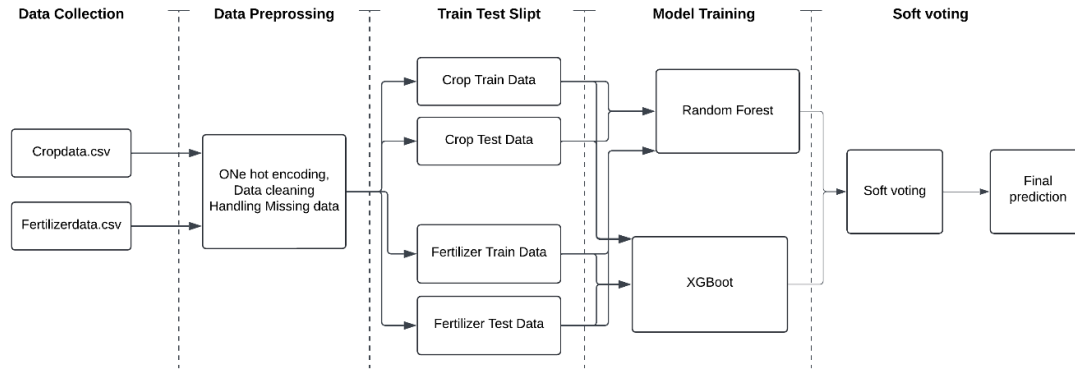


Figure 1: *Proposed block diagram*

3.1 Mathematical Expressions and Symbols

To increase the accuracy of crop and fertilizer recommendations, this work uses an ensemble model that combines the Random Forest and XGBoost algorithms with a soft voting mechanism. By averaging the prediction probabilities of both models, this combination takes advantage of their strengths and produces more dependable results.

To identify the most likely class, the soft voting method averages the probability predictions made by Random Forest and XGBoost. This procedure balances out individual model biases and improves prediction accuracy, which makes it especially useful in ensemble models.

Let:

- $P_{XGBoost}(C_i)$ represent the probability predicted by the XGBoost model for class C_i ,
- $P_{RF}(C_i)$ represent the probability predicted by the Random Forest model for class C_i .

The combined probability for each class C_i in the ensemble model is calculated by averaging the probabilities from both models:

$$P_{Ensemble}(C_i) = \frac{P_{XGBoost}(C_i) + P_{RF}(C_i)}{2}$$

The class with the highest combined probability is chosen as the **final prediction** $C_{predicted}$:

$$C_{predicted} = \arg C_i \max P_{Ensemble}(C_i)$$

To evaluate the accuracy of the ensemble model, the **Cross-Entropy Loss** function is used, measuring the error between predicted probabilities and true class labels:



$$LCE = -\sum_{i=1}^N \sum_{j=1}^M y_{ij} \log(p_{ij})$$

where

- N is the number of samples.
- M is the number of classes.
- y_{ij} is the true label (1 if sample i belongs to class j , otherwise 0).
- p_{ij} is the probability predicted by the model that sample i belongs to class j .

The model's performance is evaluated using the following metrics:

Accuracy

Accuracy measures the proportion of correct predictions out of the total predictions, providing an overall view of model performance. It is calculated as:

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Prediction}}$$

Precision

Precision is particularly valuable in contexts where false positives are costly. For each class C_i , Precision represents the proportion of true positive predictions out of all positive predictions. It is calculated as:

$$\text{Precision} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)} + \text{False Negative (FN)}}$$

High Precision means that when the model predicts a specific crop or fertilizer, it is likely to be correct, minimizing unnecessary recommendations that could lead to resource waste.

Recall

Recall captures the model's ability to identify all relevant instances of a class. For each class C_i , Recall is the ratio of true positives to the sum of true positives and false negatives, calculated as:

$$\text{Recall} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)} + \text{False Negative (FN)}}$$



F1 Score

The F1 Score combines Precision and Recall into a single metric, calculated as the harmonic mean of the two. This balance is especially useful in agriculture, where both Precision and Recall are crucial for practical applications.

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

4. Results and Discussion

Accuracy, training time, and prediction reliability in crop and fertilizer recommendations were evaluated for the ensemble model that combined XGBoost and Random Forest. A summary of important metrics that stands out the performance of the individual models with that of the ensemble model is shown in Table 1 below. Soft voting was successful in combining predictions from both algorithms, as evidenced by the ensemble's overall accuracy rate of 99.67%.

Table 1: The performance of the individual models with that of the ensemble model

Model	Accuracy for Crop Dataset	Accuracy for Fertilizer Dataset
Random Forest	99.73 %	98.89 %
XGBoost	99.61 %	99.04 %
Ensemble Model	99.67 %	98..96 %

Each of the two models, Random Forest and XGBoost, independently achieved a high accuracy of 99.73% and 99.61% for crop and 98.89% and 99.04% for fertilizer recommendation. However, by balancing the advantages of both algorithms, the ensemble model had almost equal accuracy (99.67%). With XGBoost's capacity to manage high-dimensional data, this combined model offers the stability of Random Forest, which minimizes overfitting by averaging multiple decision trees. A model that retains accuracy across various data samples and effectively adjusts to complex agricultural datasets is the end result.

Analysis of Feature Importance and Its Applications

In Table 1, an analysis of feature importance verified and ranked that the most important parameters were temperature, pH, and nitrogen (N), phosphorus (P), and potassium (K). This result is in line with best practices in agriculture, which emphasize the importance of soil pH and nutrient levels for crop health and yield. The recommendation process is made more transparent by the model's ability to rank these features, which aids farmers in understanding which factors have the biggest influence on their choices regarding crops and fertilizer. Furthermore, the model's ability to adapt to particular climatic conditions is improved by the



inclusion of weather factors like humidity and rainfall, which makes it extremely relevant for precision agriculture.

Table 1: *An analysis of importance features have shown below*

Features	Importance Score
Nitrogen (N)	0.23
Phosphorus (P)	0.18
Potassium (K)	0.17
pH	0.14
Temperature	0.12
Humidity	0.10
Rainfall	0.06

Implications for the Future

Even though the ensemble model works well with historical data, real-time data integration via Internet of Things devices could be improved in the future to allow the model to dynamically adjust to shifting environmental and soil conditions. This development would further support sustainable and data-driven decision-making by making the model even more useful for farmers dealing with quickly changing agricultural environments. In conclusion, the ensemble model shows how effective machine learning can be in agricultural settings. The model achieves high accuracy, practical reliability, and adaptability across a range of conditions by combining Random Forest and XGBoost, supporting resource optimization and sustainable farm

5. Conclusions

This study presents a machine learning-based recommendation system that uses an ensemble model combining the Random Forest and XGBoost algorithms to improve the accuracy of crop and fertilizer recommendations. With a high accuracy of 99.67%, the model uses soft voting to strike a balance between Random Forest's resilience and XGBoost's ability to handle complex data, making it suitable for a variety of agricultural settings. This precision highlights the model's potential as a trustworthy instrument for precision farming, offering farmers data-driven suggestions based on critical soil and climate factors like temperature, pH, phosphorus, potassium, and nitrogen.

However, the model relies on historical data, which may limit its use in regions with limited or changing data. Future enhancements could include real-time data integration and additional algorithms to increase adaptability and relevance across different crops and regions. Overall, this study shows how machine learning can be used to improve precision agriculture, assisting farmers in making sustainable, optimal decisions and opening the door for more developments in data-driven farming.



References

1. Harshitha Reddy Vaddi; Guru Keerthi Mandadhi; Ameer; Dileep Kumar; SChengamma Chitteti; Smart Crop Advisor; Year: 2024, "Intelligent Crop and Fertilizer Recommendations with Crop Yield Prediction Using ML Algorithms, International Conference on Intelligent Systems for Cybersecurity, IEEE.
2. Sharma; Gupta; Joshi, Year: 2023, "Ensemble-Based Approach for Crop and Fertilizer Recommendation using Machine Learning", 2023 International Conference on Emerging Trends in Engineering and Technology (ICETET), IEEE.
3. Sana Fatima; Ayan Hussain; Sohaib Bin Amir; Syed Haseeb Ahmed; Syed Muhammad Huzaifa Aslam, Year: 2023, "XGBoost and Random Forest Algorithms: An in Depth Analysis", DOI: 10.57041/pjosr.v3i1.946 ISBN : 0552-9050.
4. Bhasha; Suresh Babu; Muniraju Naidu Vadlamudi; Kochumol Abraham; Sanjaya Kumar Sarangi, Year: 2022, "Automated Crop Yield Prediction System Using Machine Learning Algorithm", Journal of Algebraic Statistics.
5. Tripathi; Biswas, Year: 2022, "An Expert System using Ensemble Learning for Crop Yield Prediction: EESCYP-I", 2022 International Conference on Advances in Computing Communication and Materials (ICACCM), pp. 1-5.
6. Komala Devi; Josephine Prem Kumar, Year: 20204, "Sustainable Food Development Based on Ensemble Machine Learning Assisted Crop and Fertilizer Recommendation System", DOI: 10.53759/7669 /jmc202404030.
7. Sarvesh Halbe; Deep lad; Tirth Panchal; Krishi Mitra; Year: 2023, "Intelligent Crop And Fertilizer Recommender", 2023 6th International Conference on Advances in Science and Technology (ICAST), DOI:10.1109/ICAST59062.2023.10454948.
8. Deeksha Tripathi; Saroj; Biswas; Biswajit Purkayastha, Year: 2023, "A Comparative Analysis of Ensemble Learning Techniques for Crop Yield Prediction: CYPELA", 2023 4th International Conference on Computing and Communication Systems (I3CS).
9. Suma; Mamatha; Shweta; Srujan Raju; Bhagyashree, Year: 2023, "Modelling Soil Fertilizer Levels and Crop Yields in Agriculture Using Machine Learning", DOI: 10.1109/ICIICS59993.2023.10421341.
10. Morufu Olalere; Gilbert Imuetinyan Osaze; Aimufua, Muhammad Umar Abdullahi; Bako Halilu Egga, Year: 2024, "Ensemble-Based Predictive Model for Crop Recommendation", DOI : 10.4314/dujopas .v10i2c.36, ISBN: 2635-3490.
11. Davis; Chen; Year: 2023, "Using Machine Learning to Optimize Crop Yield: A Comparative



Analysis”, Journal of AI in Agriculture.

12. Chen; Zhao; Liu, Year: 2021, “Application of XGBoost in Precision Agriculture: Improving Crop Selection”, Advanced Agricultural Informatics Conference, IEEE.

13. Das; Agarwal, Year: 2023, “Predicting Soil Fertility Using Ensemble Learning Approaches”, Agricultural AI Journal.

14. Thomas; Patel, Year: 2022, “Django-Based Web Solutions for Agriculture: Case Study on Crop Recommendations”, Journal of Agricultural Informatics and Technology.

15. Subburaj, Chandana, Year: 2024, “Fertilizer and Crop Yield Prediction using Machine Learning”, DOI: 10.48175/IJARSCT-19183 | ISBN: 2581-9429.