

Crop Recommendation System Using Machine Learning for Enhanced Agricultural Productivity

K. Usha Rani¹, V. Ramakrishna^{*2}, S.V. Sudheer Kumar² and P. Bhargavi¹

¹Department of Computer Science, Sri Padmavati Mahila Visvavidyalayam, Tirupati, India

²Department of ECE, SOET, Sri Padmavati Mahila Visvavidyalayam, Tirupati, India

(Received 18 August, 2025; Accepted 21 October, 2025)

ABSTRACT

Agriculture which forms a major sector of the world economy is increasingly becoming problematic due to climate change, varied soils, and a lack of access to modern technologies. To assist farmers in making better decisions, this paper proposes a machine learning (ML)-based crop recommendation system. The predictive models are trained using past data on cropping, soil parameters (pH, nitrate, potassium and nitrogen), and weather parameters (temperature, precipitation and humidity). The algorithms considered in the evaluation are Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Decision Tree (DT), Random Forest (RF), Bagging (BG), Ada Boost (AB), Gradient Boosting (GB), Extra Trees (ET), and Logistic Regression (LR). Information cleaning techniques such as normalization and noise removal are applied to improve the model. To ensure that models are more robust and can be applied in a broader variety of situations, cross-validation methods are used to further refine the models. Depending on the environmental and agronomic conditions, the system suggests crop recommendations. The results obtained compare ensemble-based methods, and the random forest in particular, to be better than other models. This application of machine learning (ML) enhances crop planning, productivity, and supports the idea of environmentally friendly farming.

Key words: Crop selection, Agricultural Data, Climate factors, Soil properties, Predictive modeling

Introduction

The research (Prity *et al.*, 2024) assesses nine unique machine learning algorithms for the purpose of crop recommendation. It utilizes refined historical data on climate, soil, and yield. In delivering customized recommendations to enhance agricultural productivity and food safety, Random Forest has shown strong performance in multiple studies. To forecast how much crop would grow, several studies (Lahza *et al.*, 2023) have used records from several farms and layers. There have been extremely few errors and high accuracy in Random Forest models. These models improved during the course of the season as more data became available, which helped them

perform better. Worldwide, agriculture might be lost by 1.8% to 2.4% by 2030 due to urbanization; almost 80% of this loss will occur in Asia and Africa.

This reduction poses a significant danger to the livelihoods of rural residents and the productivity of agricultural output. Research suggests (Bren d'Amour *et al.*, 2017) that these changes might have a significant influence on food security in certain regions. Soil degradation and increased emissions of greenhouse gases are results of land use changes. Soil organic carbon is most rapidly depleted in grasslands and plantations, particularly after deforestation. Compared (Padbhushan *et al.*, 2022) to other types of modifications, they significantly reduce soil carbon storage.

Recent research (Anbananthen *et al.*, 2021) has demonstrated that hybrid machine learning methods, such as stacked generalization, are more effective at predicting crop yields than do solitary models. In particular, the accuracy of predictions has been demonstrated to be high for Random Forest and Gradient Boosted Tree regressors. Corn grain yield prediction was carried out using six machine learning models based on satellite data, where the Random Forest algorithm showed the best performance, particularly using the GREEN band and GEMI vegetation index, as demonstrated by relevant research (Pinto *et al.*, 2022). A comparative analysis (Thanh *et al.*, 2017) of Random Forest (RF), k-Nearest Neighbors (kNN), and Support Vector Machine (SVM) classifiers for land cover classification indicated that SVM attained the highest overall accuracy and exhibited the least sensitivity to variations in sample size, whereas all three classifiers performed effectively with larger datasets. When compared to more traditional classification methods, such as the Maximum Likelihood Method (MLM), Random Forests, and Support Vector Machines (SVM), machine learning techniques do far better when it comes to crop identification. The total accuracy of traditional approaches was 88.96%, while machine learning models were able to attain 98% accuracy. Machine learning models demonstrated high resilience by achieving over 95% identification accuracy across diverse crop kinds utilizing optimum feature combinations (Feng *et al.*, 2019). One way that machine learning might aid precision agriculture (Shahab *et al.*, 2025) is by providing farmers with tailored crop recommendations that take soil and weather into account. Random Forest, Gradient Boosting, XGBoost, LightGBM, Support Vector Machine and Decision Tree are among the algorithms that are examined, along with soil nutrients (potassium, nitrogen, and phosphorus), pH, temperature, humidity, and precipitation. The system's objective is to assist farmers in doing more while simultaneously addressing issues such as soil deterioration, excessive chemical use, and inefficient use of land and water. It uses a variety of data sources in conjunction with prediction algorithms to provide suggestions that are both timely and sensitive to context. When compared to the other models, XGBoost performed higher across the board, including accuracy, precision, recall, and F1-score. Based on the present condition of the fields, the algorithm also recommends and rates the best

five crops.

This aids farmers in making prudent and future-oriented choices. Previous research in crop recommendation has typically applied only two to six machine learning algorithms, limiting the depth of comparative analysis and robustness of results. Moreover, few studies evaluate a broad set of models on a single, consistent dataset, making it challenging to identify the most effective algorithms for practical use. In an effort to resolve this, the current work compares nine machine learning models of Logistic Regression, SVM, KNN, Decision Tree, Random Forest, Bagging, AdaBoost, Gradient Boosting, and Extra Trees on a dataset found on Kaggle consisting of soil information, weather-related conditions, crop produce, and a farmer preference. This paper is one of the earliest to compare all nine algorithms on the same dataset to recommend crops, and performance was measured in terms of accuracy, precision, recall, and F1-score.

Methodology

The objective of this work is to create and deploy an intelligent system of crop recommendations based on the predictive nature of different machine learning (ML) algorithms. The suggested framework is divided into various key steps to provide effective and quality crop recommendations. First, information is gathered using validated data collection sources and it is preprocessed by methods like normalization, dealing with missing values, and feature selection to improve data quality and model maturity. After that several ML algorithms are used to predict crops, and each of them goes through a systematic model-training and testing phase. The efficiency of all models will be measured with complete measures such as accuracy, precision, recall, and F1-score to choose on the most efficient procedure to use in practice. The entire process of data gathering up to the final generation of a recommendation is systematically outlined in Fig. 1, which shows the systematic approach taken to develop the system.

Data Collection

The dataset, namely Crop and Soil Data Set, which was used in this study, was acquired on the Kaggle repository (Badshah *et al.*, 2024). The 8,000 entries each have nine attributes, and they contain data about the type of crop, environmental elements, soil nutrients and fertilizers. This data has been gathered

meticulously to generate algorithms that can recommend crops and fertilizers. It allows studying soil chemistry, weather pattern, and agriculture practices simultaneously.

The dataset consists of three major parts. Environment conditions include temperature, humidity and soil moisture, which are examples of variables that can be used to determine external factors affecting development of crops. These factors are important when considering the response of plants to climate change. The number of macronutrients in the soil indicating fertility includes potassium, phosphorus and nitrogen. The key factors influencing the plant metabolism, its potential yield, and overall soil productivity are summarized below. Two types of agronomic data are soil type and crop type that demonstrate different types of crops and soil such as sandy soil, clay soil, and loamy soil. Taken together, these traits make the data effective in predicting where to expand and utilize it optimally.

Table 1. Description of dataset attributes

Attribute	Range / Values
Temperature	20.0 – 40.0 °C
Humidity	40.02 – 80.0 %
Moisture	20.0 – 70.0 %
Soil Type	e.g., Sandy, Loamy, Clay
Crop Type	e.g., Rice, Wheat, Maize
Nitrogen (N)	0 – 46 units
Phosphorus (P)	0 – 46 units
Potassium (K)	0 – 23 units
Fertilizer Name	e.g., Urea, DAP, Compost

The Crop Recommendation System then correctly forecasts on which crops will do well due to environmental and soil data. The MinMax Scaler (Eddamiri *et al.*, 2024) is used to normalise the features of the dataset such that they are comparable across the different ranges of characteristics. The model is then trained to make more precise predictions through categorization, like Random Forest and K-Nearest Neighbors. This information is very useful in helping farmers make decisions in many fields since it covers a wide variety of soil types and weather conditions. The entire dataset attributes used in this study are listed as below and the flow diagram is shown in Fig. 1.

Nitrogen (N): This property indicates the quantity of nitrogen in the soil and is measured in kilograms per hectare (kg/ha). Plants need nitrogen to grow,

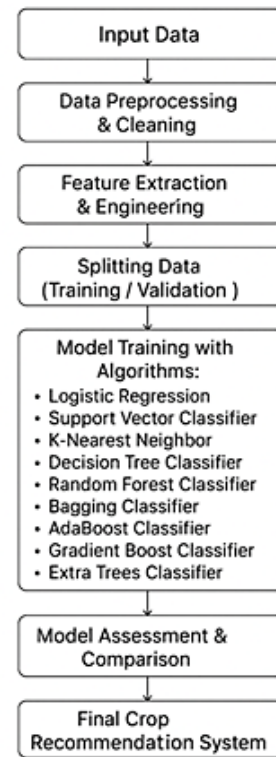


Fig. 1. Flow Diagram

and it is especially beneficial to green plants such as wheat and spinach.

P (phosphorus): The soil phosphorus level is measured in kilograms per hectare. In legumes and root crops such as peas and carrots this trait has a critical role because it determines the growth of roots.

K (potassium): The more potassium there is in plants (kg/ha) the more resistant it is to disease and the better the fruit. This especially applies to crops that bear fruits like tomatoes and bananas.

pH: The pH of the soil, i.e., its acidity or alkalinity on a 0 to 14 scale, is usually between 3.5 and 8.0. Most cereals can grow in neutral soils (6.5-7.5) although tea and barley perform well on acidic and alkaline soils, respectively.

Organic Matter: This attribute is the quantity of organic carbon within the soil. High organic matter is advantageous to fruits and vegetables as it holds more water and the nutrients become more available.

Moisture: This value represents the ability of soil to retain water. Crops such as rice thrive in high moisture, and millets and sunflowers thrive in low moisture.

Temperature: A measure of the conditions which have a direct effect on crop growth, this requires the temperature to be shown in degrees Celsius.

Whereas wheat and cabbage do well in colder climates, rice and maize do well in warmer climates.

Rainfall: This characteristic shows the amount of precipitation in millimeters (mm). Dry rainfall is advantageous to crops that can withstand drought like millets, but high rainfall favors the growth of crops like rice.

Geographical Aspects: This includes factors like climatic zones, latitude and height. Coconuts are cultivated near the shore, and coffee in the highlands.

Crop Prediction Using Machine Learning Techniques

To get precise crop recommendations based on significant agronomic factors, including soil macronutrients (potassium, phosphorus, and nitrogen), soil pH, and the present environment, this work makes use of a wide range of machine learning methods. The chosen algorithms are grouped based on their prediction accuracy, ease of usage, and ease of comprehension. Logistic Regression (LR) establishes the benchmark for linear connections, whereas Support Vector Machines (SVM) are excellent at categorizing nonlinear objects in higher dimensional scenarios. K-Nearest Neighbors (KNN) provides learning based on examples and is hence successful in local decision boundaries, even if Decision Trees (DT) preserve the separation of data into their own rules. Models can be less susceptible to overfitting by using ensemble techniques like Random Forest (RF), Bagging (BG), AdaBoost (AB), Gradient Boosting (GB), and Extra Trees (ET). This improves accuracy and captures intricate interactions, which in turn improves generalization. The system can manage the variety of soil profiles, weather patterns, and crop requirements in various agricultural regions thanks to this collection of algorithms.

Support Vector Machine

Because Support Vector Machines (SVMs) provide dependable classification models, they are useful for crop recommendations (Rajakumaran *et al.*, 2024). Choosing the optimal hyperplane to separate various crops according to soil and environmental factors is the primary objective.

For a given input x (nutrients, pH, temperature, rainfall), the SVM predicts the crop using

$$f(x)=\text{sign}(w^T x+b),$$

where w and b define the decision boundary.

SVM employs soft margins and kernels to deal with non-linear patterns when the data is complicated. After it has been taught, the model may look at soil factors and suggest the best crops to grow. This makes farming more productive, uses resources more efficiently, and is better for the environment.

Decision Tree (DT) for Crop Recommendation

A supervised learning system called a Decision Tree (Motamedi *et al.*, 2024) illustrates the decision-making process with a structure resembling a tree. While leaf nodes display the final outcome, which is often the anticipated crop type, internal nodes represent tests or conditions on input factors such as soil pH, nitrogen, phosphorus, potassium levels, or weather.

The dataset is repeatedly divided into smaller groups based on the most crucial characteristics in order to construct the tree. To find the best feature at each node, people often utilize splitting criteria like Gini Impurity, Entropy, or Information Gain. These measurements tell us how pure or informative the resulting subsets are, which makes sure that the tree has the most homogeneous branches possible.

Decision Trees are particularly suitable for agricultural applications because they handle both categorical and numerical data, require minimal preprocessing, and can naturally capture non-linear relationships between soil-climatic factors and crop suitability. However, they are sensitive to noise and small variations in data, which may result in different tree structures. Pruning techniques are often applied to mitigate overfitting and improve generalization.

Despite these limitations, Decision Trees remain a widely used and effective tool for crop recommendation, offering clear interpretability and a straightforward decision-making process that aligns well with real-world agricultural planning.

Random Forest (RF)

Random Forest is an ensemble learning method (Elbasi *et al.*, 2023) that builds many decision trees and combines their results. Each tree is trained on a random part of the dataset, and at every split, only a random set of features is checked. This randomness makes the trees diverse and improves overall

accuracy.

For classification, the final output is chosen by majority vote:

\hat{y} =most frequent class from all trees

For regression, the result is the average of predictions:

$$\hat{y} = \frac{1}{T} \sum_{t=1}^T h_t(x)$$

where T is the number of trees and $h_t(x)$ is the prediction of each tree.

In crop recommendation, Random Forest uses soil nutrients (N, P, K), pH, rainfall, temperature, and humidity to suggest the best crop. It is more stable than a single decision tree, reduces overfitting, and can also show which factors are most important in making decisions.

Logistic Regression (LR)

Logistic Regression is a statistical machine learning method (Barbedo *et al.*, 2019) commonly used for classification problems, including crop recommendation. It works well in binary cases, for example, deciding whether a specific crop is suitable or not for given soil and climate conditions.

The model learns a linear relationship between features (like soil nutrients, pH, rainfall, and temperature) and the output class. It then applies the logistic (sigmoid) function to convert this linear combination into a probability between 0 and 1:

$$P(y = 1|x) = \frac{1}{1 + e^{-(w^T x + b)}}$$

where w are the learned weights, xxx is the feature vector, and b is the bias.

In crop recommendation, LR helps evaluate whether environmental and soil factors support the growth of a given crop. It is simple, interpretable, and computationally efficient, though it may be less effective for highly non-linear relationships compared to more complex algorithms.

K-Nearest Neighbors (KNN)

A popular non-parametric method for classification and regression applications is K-Nearest Neighbors (Manju *et al.*, 2024). It works on the principle of similarity, meaning that it uses the closest instances in the feature space to generate an educated judgment about a new data point.

Based on a majority vote among the k nearest neighbors, the procedure assigns a class label. Re-

gression takes the mean of the data of those neighbors to create predictions. The parameter 2 is very important. Although the predictions with larger numbers are more consistent but less flexible, the smaller numbers are more receptive to any alterations in the area.

KNN has been called a lazy learner because it does not create an explicit training model but only does calculations when it produces predictions. This can make the process easier, although on large data sets, it can be very time-consuming. Given that KNN uses comparisons of distances, feature scaling or normalization is advisable to deliver meaningful results.

KNN is a useful approach when the variables to be considered are multiple (such as crop advice where multiple agricultural parameters need to be taken into account). This is because it is simple to use, flexible, and can support complex data structure.

Ada Boost

Adaptive Boosting (Ada Boost) (Yoon *et al.*, 2023) is a popular way for groups to learn that uses a lot of poor learners to create a better prediction model. Decision stumps, or shallow decision trees, are often used for learners who don't do well. As the algorithm goes over each learner again, they pay greater attention to the examples that the previous ones missed.

By adjusting the weights of the training samples as necessary, AdaBoost ensures that more difficult examples receive more attention in subsequent rounds. Either a weighted vote (for classification) or a weighted average (for regression) of all the students makes the final forecast. Ada Boost is popular because it is easy to use, effective, and may improve accuracy significantly over single models. However, because they could be given higher weight during training, it may be susceptible to noisy data and outliers. However, it remains a preferred option for many machine learning jobs, including crop recommendation, where improving prediction accuracy is crucial.

Gradient Boosting

In machine learning, gradient boosting (Huber *et al.*, 2022) is a powerful ensemble learning technique that is frequently applied to both regression and classification problems. By sequentially assembling several weak learners, often shallow decision trees, it creates

a potent prediction model.

In every iteration, a new tree is trained to correct the errors of the previous trees. This makes the model more accurate over time. The main goal is to make a given loss function better over time, with each new learner fixing the faults that the current ensemble committed. To avoid overfitting and make the model more generic, you need to carefully choose the learning rate, maximum tree depth, and number of estimators.

Gradient Boosting is well-known for being very accurate, being able to use a wide range of loss functions, and being able to represent complicated relationships. But if it isn't regularized enough, it could be hard to calculate and sensitive to noisy data. Even with these problems, it is still one of the best algorithms and is frequently used in fields where making accurate predictions is very important, like marketing, finance, healthcare, and farming.

Testing and Training

It is very important to care for class imbalance when building a model since datasets that are not balanced might make models inappropriately group minority classes and favor majority classes. To make the learning more fair, we used a down-sampling approach to balance out the distribution of classes.

It also used early stopping to make it more generic and get rid of overfitting. Training ceases as soon as the validation error stops becoming better. After that, the machine learning algorithms were taught to use the processed input data to figure out the best ways to grow crops.

Twenty percent of the dataset was set aside for independent testing, while the other eighty percent was utilized to teach the models. We can objectively evaluate prediction performance because of this division, and we can be sure that the findings we show show that the model can work with new data.

Performance Metrics

To thoroughly assess the efficacy of the machine learning models, many standard measures were utilized, including the Confusion Matrix (CM), Receiver Operating Characteristic (ROC) curve, Precision, Recall, F1-score, and Accuracy (Praharsha *et al.*, 2024).

Confusion Matrix (CM)

A CM provides a tabular visualization of the classifier's outcomes by comparing predicted labels

against actual labels. It consists of:

- True Positives (TP): Correctly predicted positive cases
- False Positives (FP): Incorrectly predicted as positive
- True Negatives (TN): Correctly predicted negative cases
- False Negatives (FN): Incorrectly predicted as negative

$$CM = \begin{bmatrix} TP & FP \\ FN & TN \end{bmatrix}$$

Receiver Operating Characteristic (ROC) Curve

The ROC curve plots the True Positive Rate (TPR) against the False Positive Rate (FPR) at varying thresholds:

$$TPR = \frac{TP}{TP + FN}, FPR = \frac{FP}{FP + TN}$$

The Area Under the Curve (AUC) provides a scalar value to summarize ROC performance, where higher AUC indicates better discriminative ability.

Precision (P)

Precision quantifies the correctness of positive predictions, i.e., the proportion of correctly predicted positives among all predicted positives:

$$Precision = \frac{TP}{TP + FP}$$

Recall (R) / Sensitivity / True Positive Rate

Recall measures the ability of the model to correctly identify actual positives:

$$Recall = \frac{TP}{TP + FN}$$

F1-Score

The F1-score represents the harmonic mean of Precision and Recall, balancing the trade-off between them:

$$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}$$

Accuracy (Acc)

Accuracy gives the overall proportion of correctly classified instances:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Results and Discussion

The experimental results of applying machine learning algorithms to the suggested crop recommendation system are examined in detail in this section. Numerous performance metrics were used in the analysis, including the Receiver Operating Characteristic (ROC) curve, F1-score, accuracy, precision, recall, and confusion matrix (CM). To evaluate how the true positive rate (TPR) and false positive rate (FPR) varied with various thresholds, we examined the ROC curves. Because they were better at differentiating between crop classes, models with larger Area Under the Curve (AUC) values were considered more reliable.

Recall and precision demonstrated how effectively the system struck a compromise between preventing too many false alarms and obtaining all the significant positive cases. The effectiveness of the entire classification procedure was evaluated using accuracy scores. When there were too many or too few in each class, the F1-score—the harmonic mean of accuracy and recall—was a fair method to assess how well the models performed. By identifying the types of misclassifications that existed, the confusion matrix also enabled a more thorough investigation of errors. Suppose, for instance, that a crop was mistakenly identified as being in a different crop class. The system's high sensitivity to samples from the minority class demonstrated its shortcomings.

Visualization of Model Responses for Crop Recommendations

Used several visualization techniques to see how beneficial the proposed crop recommendations models were and to check the correctness of the classifications and the links between the crops.

Figure 4's confusion matrix shows how accurate the classification of the different sorts of crops is overall. The number of true positives, false positives, true negatives, and false negatives for each class are displayed to indicate how accurate the model's predictions are. This graphic also helps to figure out which classes are most commonly misclassified and make the model even better.

Additionally, the trade-off between the True Positive Rate (TPR) and the False Positive Rate (FPR) at various categorization levels is depicted by the Receiver Operating Characteristic (ROC) curves in Figure 5. The models' ability to distinguish between different crops is measured by the AUC values. The

models are better able to predict which crops belong to which category when the AUC values are higher.

The correlation matrix for every crop is displayed in Figure 6. In order to determine the relationships between crop pairings based on their feature sets, Pearson correlation coefficients were applied. Crop relationships are depicted in the diagonal entries (e.g., "Rice to rice, Maize to maize, Jute to jute"). The off-diagonal entries illustrate the relationships between crops in various agronomic and environmental contexts. This matrix is a helpful tool for understanding crop interdependencies and potential interactions, particularly with regard to soil composition, climate change, and farming practices.

Quick Insights from Table

- **Random Forest (RF):** Best performer (Accuracy ~99.3%, CV Accuracy ~99.4%).
- **Extra Trees (ET):** Almost as good as RF (~99.0%).
- **Decision Tree (DT):** Strong (98.8%), but slightly behind ensemble methods.
- **Gradient Boosting (GB):** Very reliable (~98.1%).
- **Logistic Regression (LR) & SVM:** Solid baseline models (~96–97%).
- **KNN:** Acceptable but not as strong (~95.6%).
- **Bagging (BG):** High performance (~99%).
- **AdaBoost (AB):** Clearly under performed (~14%). Likely due to parameter issues or dataset imbalance

The comparative evaluation of nine machine learning algorithms for crop recommendation revealed significant variations in predictive performance. The baseline classifiers—Logistic Regression, Support Vector Machine (SVM), Decision Tree, and K-Nearest Neighbor (KNN)—showed strong and consistent results, though with some differences in robustness. Logistic Regression achieved an accuracy of 98.4%, with precision and recall values of 98.5% and 98.4%, respectively. SVM slightly outperformed Logistic Regression in terms of precision (98.7%) but achieved the same accuracy (98.4%). The Decision Tree matched Logistic Regression almost exactly, reaching an accuracy of 98.4%, with a nearly identical balance of precision and recall. KNN, in contrast, lagged slightly behind, with an accuracy of 97.7%, a precision of 97.8%, and recall of 97.7%, highlighting its sensitivity to feature overlap and noise.

Ensemble methods clearly dominated the experi-

ment. Both Random Forest and Extra Trees Classifier delivered the highest overall performance, each achieving an accuracy of 99.3%, with equally impressive precision (99.4%) and recall (99.3%). These findings show that tree-based ensembles had outstanding generalization capabilities in addition to capturing intricate non-linear feature correlations.. The Bagging Classifier also performed strongly, achieving an accuracy of 98.6%, precision of 98.7%, and recall of 98.6%, outperforming all baseline models. Boosting algorithms produced mixed outcomes. Gradient Boosting achieved near-optimal performance with an accuracy of 98.9%, precision of 98.9%, and recall of 98.9%, placing it among the best models in this study. However, Ada Boost performed exceptionally poorly, with an accuracy of only 13.6%, precision of 7.1%, and recall of 13.6%. This stark contrast highlights the limitations of AdaBoost in multi-class agricultural datasets, where noise, class imbalance, or weak base learners can drastically reduce performance.

In order to continue exploring the performance of the baseline models, confusion matrices of Logistic Regression, SVM, and KNN were evaluated, and they are displayed in Figures X–Y. These also give a breakdown of the misclassifications at the level of classes and outline particular crops that were problematic across models. Below Figure 2 shows the confusion matrix of nine models. The comparison of the three classifiers is given and in similar way remaining can easily analysed. Logistic Regression, Support Vector Machine (SVM) and K-Nearest Neighbors (KNN) in terms of their confusion matrices show significant differences in performance. Logistic Regression is shown to be reasonable in general and with misclassification in some category of crops, specifically black gram, moth beans, pigeon peas, maize and rice with three error in rice. The SVM model has the best classification accuracy

and almost perfect results on most classes with only few mistakes, mostly on rice (five cases), and in few cases, on maize and pigeon peas may result in error. The KNN model is also very robust, yet it has slightly more errors than SVM, in the form of jute (five misclassifications), grapes (two), rice (five) and minor confusion in pigeon peas and moth beans. One unifying theme of all the three models is that it is difficult to classify rice correctly, suggesting that its feature space significantly overlaps other crops.

In contrast, a number of crops, including apple, banana, chickpea, coconut, coffee, and mango are always perfectly and strictly determined by all models. On the whole, the comparative analysis indicates that SVM has better performance, and KNN and Logistic Regression come next, and the classification error of rice, black gram, moth beans, pigeon peas, and jute is the area where the further optimization of the feature selection or the model is needed. The Receiver Operating Characteristic (ROC) curve is a useful tool to evaluate classification models, as it shows the trade-off between the true positive rate and false positive rate. In this study, ROC curves were generated for all algorithms but compared Logistic Regression, Support Vector Machine (SVM), and K-Nearest Neighbors (KNN) classifiers across different crop categories. With the Area Under the Curve (AUC) values for each crop equal to 1.00, the findings demonstrate that all three of the models displayed in Figure 3 performed exceptionally well. This indicates flawless categorization skills with no incorrect categories. While KNN also performs flawlessly, albeit with a little more curve fluctuation, SVM and logistic regression have nearly identical ROC curves, attaining maximum accuracy at extremely low false positive rates. All things considered, the investigation demonstrates that every model produces extremely accurate forecasts for every kind of crop.

Table 2. Comparative Performance of Algorithms

Algorithm	Precision	Recall	F1-Score	Accuracy
Logistic Regression	0.984969	0.984091	0.983993	0.984091
SVM	0.986580	0.984091	0.983908	0.984091
KNN	0.978474	0.977273	0.977231	0.977273
Decision Tree	0.984699	0.984091	0.984128	0.984091
Random Forest	0.993506	0.993182	0.993178	0.993182
Bagging	0.987361	0.986364	0.986324	0.986364
AdaBoost	0.070574	0.136364	0.080303	0.136364
Gradient Boosting	0.989742	0.988636	0.988723	0.988636
Extra Trees	0.993506	0.993182	0.993178	0.993182

An agricultural feature association map of temperature, humidity, pH, rainfall, nitrogen (N), phosphorus (P), and potassium (K) may be seen below figure 4. These values range from -1 to +1, where high values indicate a relationship between two traits that is low or non-existent, negative values indicate the opposite way, and high values are near zero. Phosphorus (P) and potassium (K) have the strongest correlation in the matrix (0.74), suggesting that they frequently change together. Since they have very little correlation with the majority of the factors, other variables like temperature, humidity, pH, and rainfall behave more like independent variables. This aids in understanding the characteristics that are connected to one another. The findings show that certain models perform better than others. Random Forest and Extra Trees produced nearly perfect results, demonstrating their ability to integrate a number of subpar learners into a far more powerful and accurate classifier. Their success can be attributed to their resistance to

overfitting, non-linear connection factor, and variance reduction. Because of these qualities, they are particularly well-suited for agricultural prediction problems, where soil and environmental factors frequently interact in intricate, non-linear ways. Gradient Boosting did remarkably well as well, but little worse than Random Forest and Extra Trees. Sequential error correction is one of its strong points, but the findings imply that adjusting hyperparameters could be necessary to attain the best accuracy. When computational simplicity is chosen over hyperparameter complexity, bagging is a good choice since it produced competitive results while balancing robustness and generalization. Since these very simple models already obtained accuracies above 98%, the baseline models—Logistic Regression, SVM, and Decision Tree—showed that the dataset itself had a high degree of class separability. Such performance suggests that these models might still provide farmers with useful advice even in resource-constrained settings where

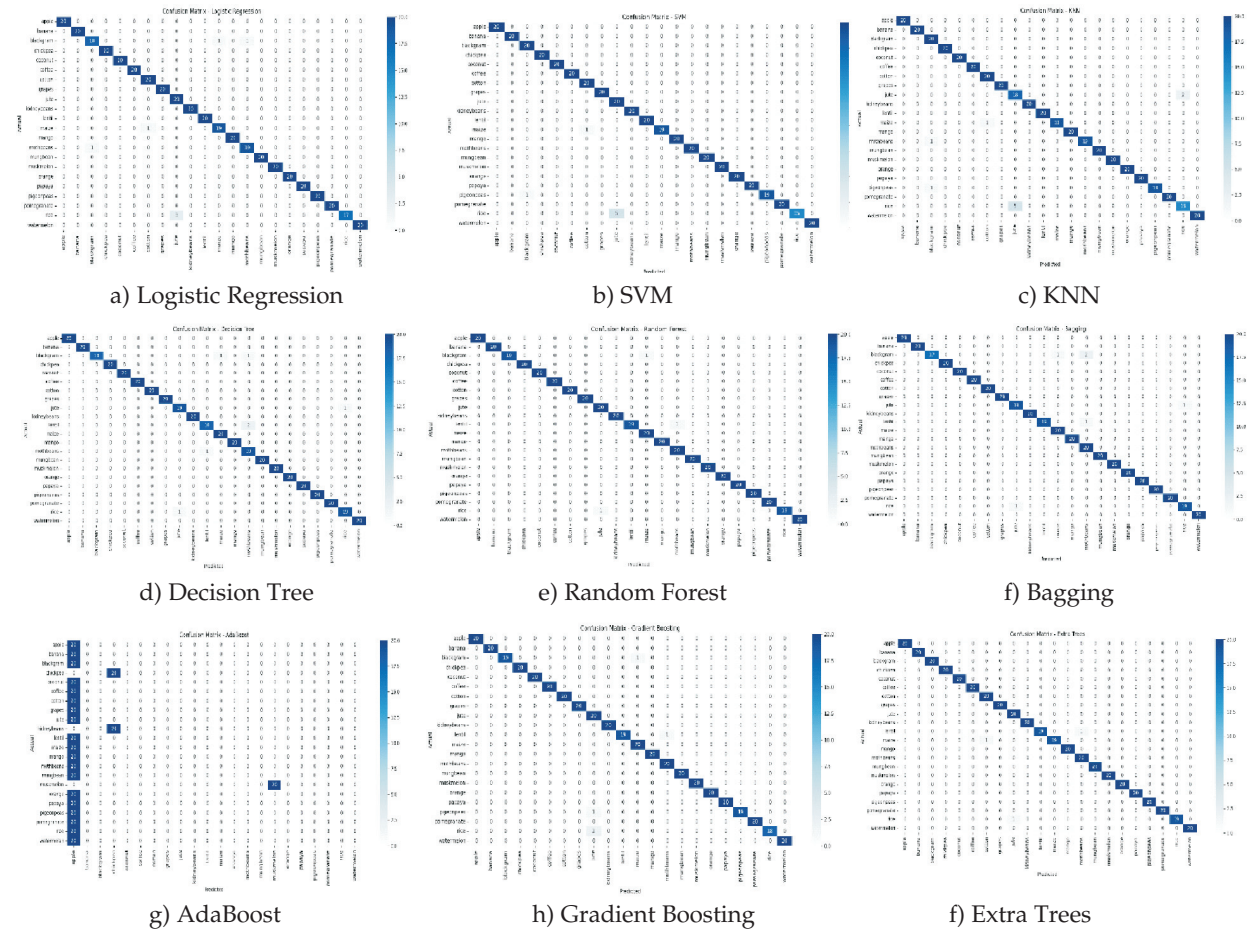


Fig. 2. Confusion Matrix

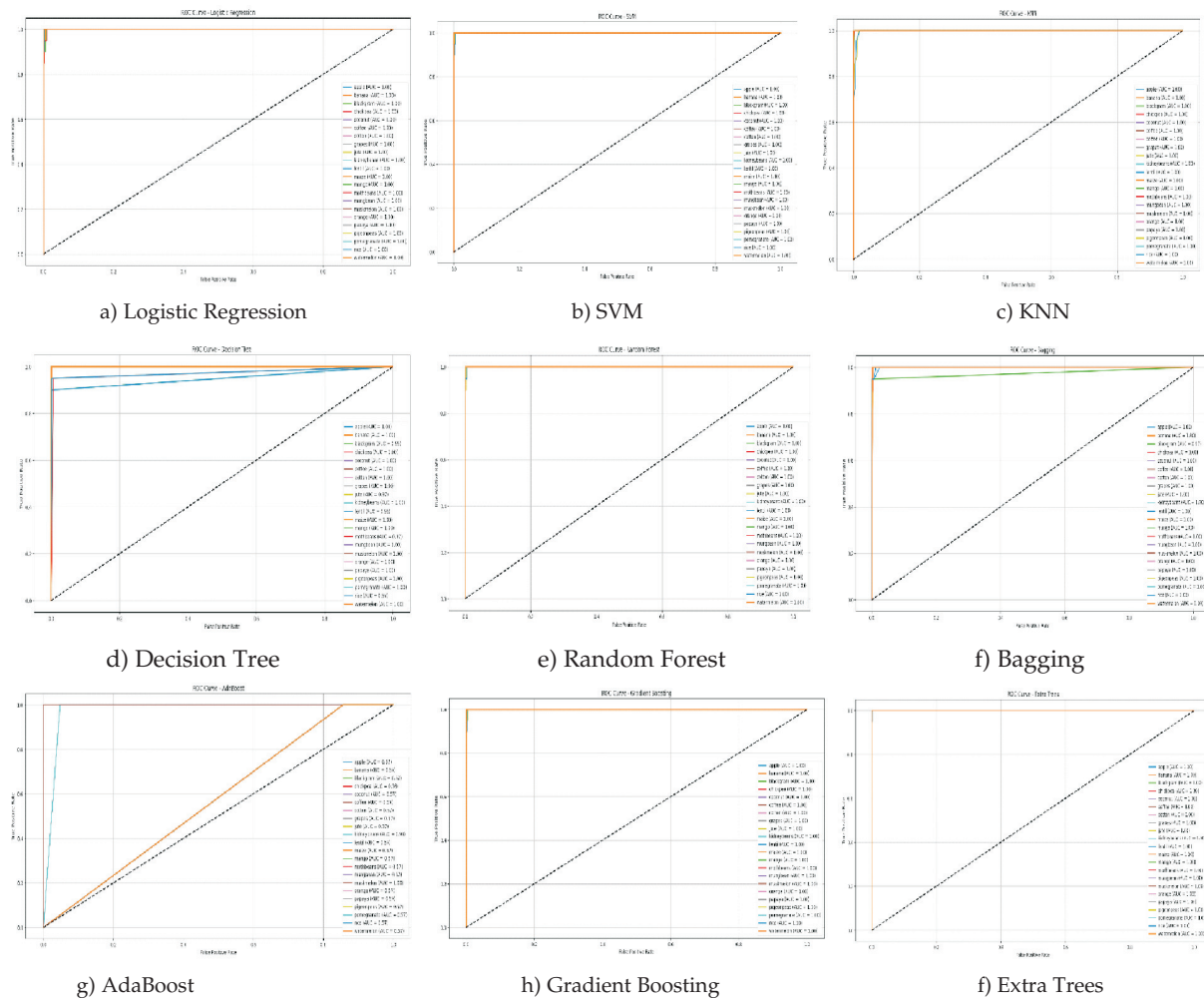


Fig. 3. ROC Curves

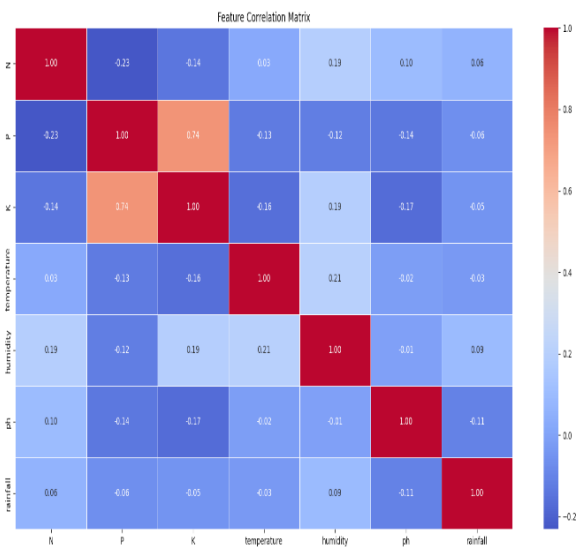


Fig. 4. Feature Correlation Matrix

sophisticated ensembles cannot be used. KNN's somewhat worse performance, however, highlights its drawback in noisy and high-dimensional agricultural datasets, where distance measurements lose their ability to discriminate. The most startling discovery is AdaBoost's failure. Its 13.6% accuracy, which is significantly lower than that of any other model, indicates that either the algorithm had trouble handling multi-class distribution and imbalance, or the weak learners it utilized were not appropriate for the data. This finding emphasizes that not all ensemble approaches are equally relevant, necessitating careful algorithm selection and tweaking in agricultural prediction systems.

Conclusion

With accuracies above 99% and balanced precision

and recall, this study shows that ensemble tree-based classifiers—Random Forest and Extra Trees in particular—are the most dependable and accurate algorithms for crop selection. Gradient Boosting and Bagging also emerged as strong alternatives, while baseline models such as Logistic Regression and SVM remain viable for contexts where computational resources are limited. AdaBoost proved inappropriate in its default configuration, whereas KNN provided competitive but marginally worse performance. Looking forward, the integration of these high-performing models into real-time decision support systems for farmers represents a practical and impactful direction. Such systems could incorporate localized soil, weather, and crop history data to deliver precise recommendations, reducing trial-and-error cultivation and optimizing resource use. Moreover, future research should explore hybrid models, explainable AI approaches, and transfer learning techniques to enhance transparency and adaptability across diverse agricultural regions. By bridging advanced machine learning with farmer-centric applications, this line of research can contribute directly to sustainable agriculture, food security, and farmer empowerment.

Acknowledgements

The authors gratefully acknowledge that the funding for this research was provided by the Pradhan Mantri Uchchar Shiksha Abhiyan (PM- USHA), under the Multi-Disciplinary Education and Research Universities (MERU) grant, sanctioned to Sri Padmavati Mahila Visvavidyalayam, Tirupati.

Conflict of Interest- None

References

- Anbananthen, K.S.M., Subbiah, S., Chelliah, D., Sivakumar, P., Somasundaram, V., Velshankar, K.H. and Khan, M. 2021. An intelligent decision support system for crop yield prediction using hybrid machine learning algorithms. *F1000Research*, 10, p.1143.
- Badshah, A., Alkazemi, B.Y., Din, F., Zamli, K.Z. and Haris, M. 2024. Crop classification and yield prediction using robust machine learning models for agricultural sustainability. *IEEE Access*.
- Barbedo, J.G.A. 2019. Detection of nutrition deficiencies in plants using proximal images and machine learning: A review. *Computers and Electronics in Agriculture*. 162: 482-492.
- Bren d'Amour, C., Reitsma, F., Baiocchi, G., Barthel, S., Güneralp, B., Erb, K.H., Haberl, H., Creutzig, F. and Seto, K.C. 2017. Future urban land expansion and implications for global croplands. *Proceedings of the National Academy of Sciences*. 114(34): 8939-8944.
- Eddamiri, S., Bassine, F.Z., Ongoma, V., Epule Epule, T. and Chehbouni, A. 2024. An automatic ensemble machine learning for wheat yield prediction in Africa. *Multimedia Tools and Applications*. 83(25): pp.66433-66459.
- Elbasi, E., Zaki, C., Topcu, A.E., Abdelbaki, W., Zreikat, A.I., Cina, E., Shdefat, A. and Saker, L. 2023. Crop prediction model using machine learning algorithms. *Applied Sciences*. 13(16): 9288.
- Feng, S., Zhao, J., Liu, T., Zhang, H., Zhang, Z. and Guo, X. 2019. Crop type identification and mapping using machine learning algorithms and sentinel-2 time series data. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*. 12(9): 3295-3306.
- Huber, F., Yushchenko, A., Stratmann, B. and Steinhage, V. 2022. Extreme Gradient Boosting for yield estimation compared with Deep Learning approaches. *Computers and Electronics in Agriculture*. 202: p.107346.
- Lahza, H., Kumar, K.R.N., Sreenivasa, B.R., Shawly, T., Alsheikhy, A.A., Hiremath, A.K. and Lahza, H.F.M. 2023. Optimization of crop recommendations using novel machine learning techniques. *Sustainability*. 15(11): 8836. <https://doi.org/10.3390/su15118836>
- Manju, G., Syam Kishor, K.S. and Binson, V.A. 2024. An IoT-Enabled Real-Time Crop Prediction System Using Soil Fertility Analysis. *Eng 5*, no. 4: 2496-2510.
- Motamedi, B. and Villányi, B. 2024. A predictive analytics model with Bayesian-Optimized Ensemble Decision Trees for enhanced crop recommendation. *Decision Analytics Journal*. 12: 100516.
- Padbhushan, R., Kumar, U., Sharma, S., Rana, D.S., Kumar, R., Kohli, A., Kumari, P., Parmar, B., Kaviraj, M., Sinha, A.K. and Annapurna, K. 2022. Impact of land-use changes on soil properties and carbon pools in India: A meta-analysis. *Frontiers in Environmental Science*. 9: 794866.
- Pinto, A.A., Zerbato, C., Rolim, G.D.S., Barbosa Júnior, M.R., Silva, L.F.V.D. and Oliveira, R.P.D. 2022. Corn grain yield forecasting by satellite remote sensing and machine-learning models. *Agronomy Journal*. 114(5): 2956-2968..
- Praharsha, C.H., Poulouse, A. and Badgujar, C. 2024. Comprehensive investigation of machine learning and deep learning networks for identifying multispecies tomato insect images. *Sensors*. 24(23): 7858.
- Prity, F.S., Hasan, M.M., Saif, S.H. et al. Enhancing Agricultural Productivity: A Machine Learning Approach to Crop Recommendations. *Hum-Cent Intell*

- Syst.* 4: 497–510 (2024). <https://doi.org/10.1007/s44230-024-00081-3>.
- Rajakumaran, M., Arulselvan, G., Subashree, S. and Sindhuja, R. 2024. Crop yield prediction using multi-attribute weighted tree-based support vector machine. *Measurement: Sensors*. 31: 101002.
- Shahab, H., Naeem, M., Iqbal, M., Aqeel, M. and Ullah, S.S. 2025. IoT-driven smart agricultural technology for real-time soil and crop optimization. *Smart Agricultural Technology*. 10: 100847..
- Thanh Noi, P. and Kappas, M. 2017. Comparison of random forest, k-nearest neighbor, and support vector machine classifiers for land cover classification using Sentinel-2 imagery. *Sensors*. 18(1): 18.
- Yoon, H.I., Lee, H., Yang, J.S., Choi, J.H., Jung, D.H., Park, Y.J., Park, J.E., Kim, S.M. and Park, S.H., 2023. Predicting models for plant metabolites based on PLSR, AdaBoost, XGBoost, and LightGBM algorithms using hyperspectral imaging of *Brassica juncea*. *Agriculture*. 13(8): 1477.