A Comparative Study of Supervised Machine Learning Classifiers for Crop Prediction in Smart Agriculture

Boddapati Soujanya [1*], Jonnakuti Punnami Devi [2]
[1*] Assistant Professor, [2] Assistant Professor
[1*,2] Dept. of CSE, RISE Krishna Sai Prakasam Group of Institutions, Ongole, Andhra Pradesh, India.

ABSTRACT:

The rapid advancement of smart farming technologies has highlighted the importance of data-driven approaches for improving agricultural productivity and sustainability. Crop prediction is a fundamental task in precision agriculture, where accurate identification of suitable crops based on soil and environmental conditions can significantly enhance yield and resource utilization. This study presents a comparative analysis of supervised machine learning classification algorithms for crop prediction using agricultural datasets integrated with IoT-based sensing information. The evaluated classifiers include Logistic Regression, K-Nearest Neighbors (KNN), Decision Tree, Naïve Bayes, Random Forest, Support Vector Machine (SVM), and XGBoost Classifier. The models were trained and tested on a benchmark dataset containing soil nutrient parameters and climatic features such as temperature, humidity, rainfall, and pH value. Performance was assessed using accuracy, precision, recall, and F1-score. Experimental results demonstrate that ensemble-based methods outperform traditional classifiers, with the XGBoost Classifier achieving the highest prediction accuracy and robustness, followed closely by Random Forest. The findings confirm that advanced ensemble classifiers can effectively support intelligent crop prediction and decision-making in IoT-enabled smart farming systems, contributing to increased agricultural efficiency and sustainable food production.

Keywords: Crop Prediction; Machine Learning; Smart Farming; Classification Algorithms; IoT; Precision Agriculture

1. INTRODUCTION

Agriculture plays a vital role in global food security and economic development; however, it faces increasing challenges due to climate variability, soil degradation, population growth, and inefficient resource utilization. Traditional farming practices largely depend on farmers' experience and historical knowledge, which are often insufficient to manage the complexity of modern agricultural systems. Consequently, there is a growing demand for intelligent, data-driven solutions that can support accurate and timely decision-making in agriculture [11], [12].

Recent advancements in smart farming and precision agriculture have enabled the large-scale collection of agricultural data through Internet of Things (IoT) sensors, satellite imagery, and automated monitoring systems. These technologies generate valuable information related to soil nutrients, temperature, humidity, rainfall, and soil pH. When properly analyzed, such data can significantly enhance crop planning and improve agricultural productivity [1], [2], [11], [13]. However, extracting meaningful insights from large, heterogeneous agricultural datasets remains a challenging task.

Machine learning (ML) has emerged as a powerful approach for analyzing agricultural data due to its ability to model complex and nonlinear relationships. In crop prediction, ML techniques are commonly formulated as classification problems, where the objective is to identify the most suitable crop type based on soil and environmental conditions. Accurate crop prediction enables farmers to optimize land usage, reduce production risks, minimize resource wastage, and improve yield quality [3], [14], [15].

Several supervised machine learning algorithms have been applied to crop prediction, including Logistic Regression, K-Nearest Neighbors (KNN), Decision Tree (DT), Naïve Bayes, Support Vector Machine (SVM), and ensemble-based approaches such as Random Forest and boosting algorithms [4], [5]. While traditional classifiers are computationally efficient and interpretable, ensemble methods often achieve higher accuracy by reducing bias and variance. In particular, boosting-based models have gained attention due to their robustness and superior performance on structured agricultural datasets [6], [8], [15].

Despite the extensive use of machine learning in agriculture, many existing studies focus on a limited set of algorithms or lack a comprehensive comparative analysis under a unified framework. Moreover, the relative performance of advanced ensemble classifiers compared to traditional methods remains insufficiently explored. To address these limitations, this study presents a systematic comparison of widely used supervised classification algorithms for crop prediction, aiming to identify the most effective approach for IoT-enabled smart farming systems.

The remainder of this paper is organized as follows: Section 2 presents a review of existing literature on crop prediction in smart agriculture, with emphasis on supervised classification and ensemble learning techniques. Section 3 describes the dataset, preprocessing steps, and the proposed crop prediction methodology. Section 4 reports the experimental results and provides a detailed discussion of model performance. Finally, Section 5 concludes the paper and outlines potential directions for future research.

2. LITERATURE REVIEW

The use of machine learning techniques for crop prediction and yield estimation has been widely explored in recent years. A systematic review by van Klompenburg et al. [1], [14] analyzed various machine learning models applied to crop yield prediction and emphasized the importance of integrating multi-source environmental data to achieve high prediction accuracy. Similarly, Li et al. [2] demonstrated that machine learning models trained on soil and weather data significantly improve wheat yield prediction by capturing nonlinear feature interactions.

Early studies in crop prediction often employed traditional classifiers, such as Logistic Regression and Naïve Bayes, due to their simplicity and computational efficiency. Kuradusenge et al. [3] compared Logistic Regression, Naïve Bayes, and Random Forest for crop yield prediction and reported that probabilistic classifiers perform well when the dataset is structured and relatively noise-free. However, these models struggle to represent complex dependencies among features.

Distance-based and rule-based classifiers have also been investigated. K-Nearest Neighbors (KNN) has been applied to crop classification problems where similarity among soil and climate conditions is crucial [4]. Although KNN can achieve competitive accuracy, it is computationally expensive for large datasets. Decision Tree–based methods are popular due to their interpretability and ability to model decision rules; however, standalone decision trees are prone to overfitting [5].

To overcome the limitations of single classifiers, ensemble learning techniques have been increasingly adopted. Random Forest, which aggregates multiple decision trees using bagging, has shown improved robustness and accuracy in crop prediction tasks [6]. Support Vector Machines (SVMs) have also demonstrated strong performance in agricultural classification problems, particularly when kernel functions are used to model nonlinear decision boundaries [7], [14].

More recently, boosting-based algorithms have gained prominence in agricultural machine learning applications. Chen and Guestrin [8] introduced the XGBoost algorithm, which enhances gradient boosting through regularization and efficient computation. Several studies have reported that XGBoost outperforms traditional classifiers and bagging-based ensembles in crop prediction and yield estimation tasks [9], [10], [15].

Although existing research confirms the effectiveness of machine learning in crop prediction, many studies evaluate algorithms in isolation or under inconsistent experimental settings. This highlights the need for a comprehensive comparative analysis using a unified dataset, preprocessing pipeline, and evaluation metrics. The present study addresses this gap by systematically comparing Logistic Regression, KNN, Decision Tree, Naïve Bayes, Random Forest, SVM, and XGBoost Classifier for crop prediction based on soil and environmental features.

3. MATERIAL AND METHODS

This section describes the dataset used, data preprocessing steps, machine learning models, experimental setup, and evaluation metrics employed for crop prediction.

3.1 Dataset Description

The study utilizes an agricultural crop prediction dataset containing 2200 samples representing different crop types. Each sample corresponds to a specific combination of soil and climatic conditions, with the objective of predicting the most suitable crop class [6], [14].

The dataset used in this study consists of seven numerical input features that represent essential soil and environmental characteristics influencing crop growth. These features include the nitrogen (N), phosphorus (P), and potassium (K) content of the soil, which are critical macronutrients required for healthy plant development. In addition to soil nutrients, climatic parameters such as temperature (°C), humidity (%), and rainfall (mm) are incorporated to capture the environmental conditions affecting crop suitability. The soil pH value is also included, as it plays a significant role in nutrient availability and crop adaptability. The output variable of the dataset is the crop type, formulated as a multi-class classification problem with 22 distinct crop categories,

where each instance is labelled with the most suitable crop corresponding to the given soil and environmental conditions.

3.2 Data Preprocessing

The collected agricultural data were preprocessed to improve model reliability and prediction accuracy. Data cleaning was first performed to handle missing values and remove duplicate or inconsistent records. To address differences in feature scales, standardization was applied, which is particularly important for algorithms such as K-Nearest Neighbors and Support Vector Machine that are sensitive to feature magnitudes. Finally, the dataset was split into 70% training data and 30% testing data, ensuring effective model learning and unbiased performance evaluation.

3.3 Proposed Methodology

The methodology can be visualized as a linear pipeline that transforms environmental "noise" into precise agricultural decisions showed in Table 1.

Table 1. Workflow Stages of the Proposed Crop Prediction System

Stage	Action	Description
Data Acquisition	Sensing	Gathering real-time data from IoT sensors (Soil moisture, pH, N−P−K levels, temperature) and historical climate datasets.
Preprocessing	Refining	Cleaning "noisy" sensor data, handling missing values, and using Normalization to ensure all features (like humidity % vs. temperature) are on a comparable scale.
Model Training	Learning	Feeding the processed data into supervised algorithms (e.g., Random Forest, SVM, or Neural Networks) where the model learns the relationship between soil/weather and crop yield.
Evaluation	Validation	Using metrics like Accuracy, Precision, Recall, and F1-Score to ensure the model isn't just guessing but actually understands the patterns.
Decision Support	Application	The final output: providing the farmer with specific crop recommendations or automated irrigation triggers via a mobile or web dashboard.

3.4 Classification Algorithms

Seven supervised machine learning algorithms were implemented and compared to evaluate their effectiveness in crop classification. Logistic Regression was used as a baseline linear probabilistic model due to its simplicity and interpretability. K-Nearest Neighbors was applied as a distance-based approach that classifies crops based on similarity to nearby samples in the feature space. Decision Tree was employed as a rule-based method that recursively partitions the data using feature thresholds, offering intuitive decision-making logic. Naïve Bayes, a probabilistic classifier grounded in Bayes' theorem, was included for its efficiency and assumption of conditional independence among features. Random Forest was utilized as an ensemble technique that combines multiple decision trees through bagging to improve robustness and reduce overfitting. Support Vector Machine was adopted as a margin-based classifier that constructs an optimal separating hyperplane using kernel functions to handle nonlinear relationships. Finally, the XGBoost Classifier was implemented as an advanced gradient boosting method that sequentially builds trees to minimize classification errors while incorporating regularization to enhance generalization performance. The working or architectures of all the classification algorithms are resented in Figure 1.

3.5 Experimental Setup

All experiments were conducted using the Python programming language and standard machine learning libraries. Each classifier was trained using identical training and testing splits to ensure a fair comparison.

3.5.1 Training Strategy

- Default hyperparameters were used initially
- Model performance was evaluated on unseen test data
- No data leakage between training and testing phases

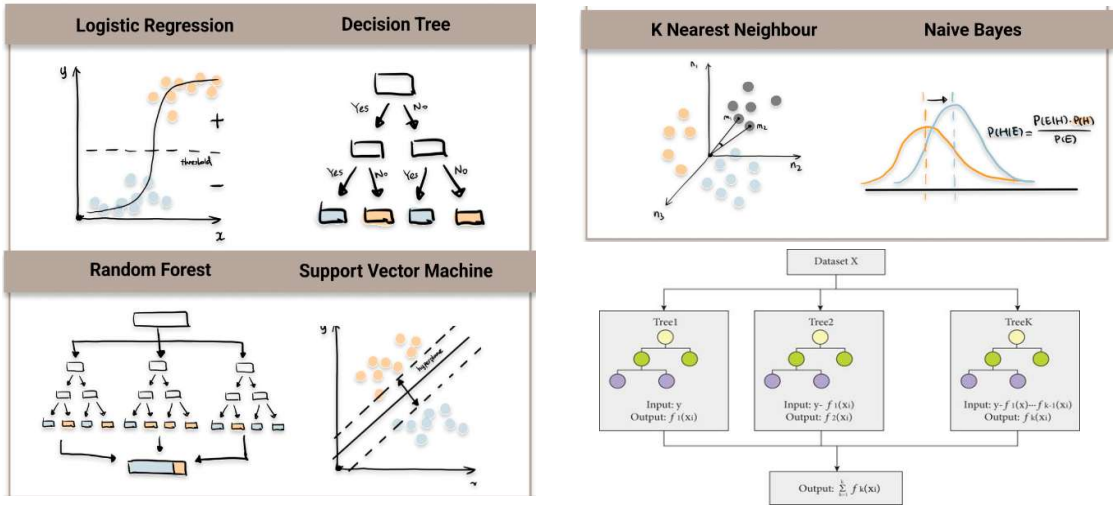

Figure 1. Working of classification algorithms

3.6 Performance Evaluation Metrics

To assess model effectiveness, multiple evaluation metrics were employed:

- Accuracy – Overall correctness of predictions
- Precision – Correctly predicted positive instances
- Recall – Ability to identify all relevant instances
- F1-Score – Harmonic mean of precision and recall
- Confusion Matrix – Detailed class-wise performance analysis

Using multiple metrics ensures robust evaluation, particularly for multi-class crop prediction problems and the performance metrics formulas are shown in Figure 2 [15], [16].

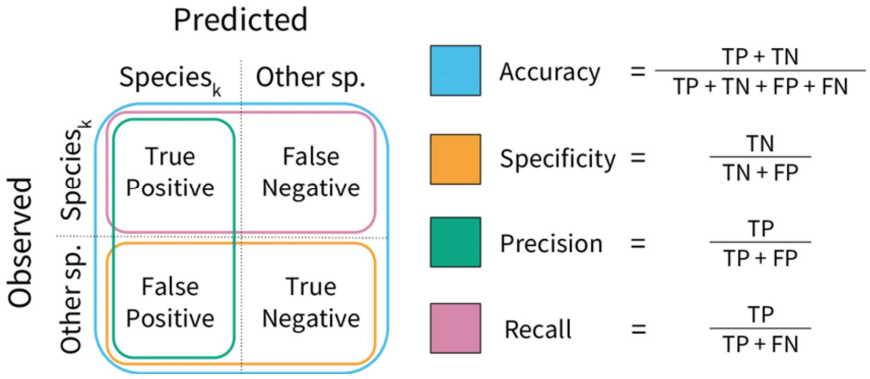

Figure 2. Performance metrics of classification algorithms

The proposed materials and methods establish a structured framework for crop prediction using supervised classification algorithms. By combining agricultural datasets, proper preprocessing, and advanced ensemble classifiers, the methodology supports accurate and scalable crop prediction for smart farming environments.

4. EMPIRICAL RESULTS AND DISCUSSION

4.1 Prediction Architecture

Figure 3 illustrates the overall architecture of the proposed crop prediction system. Agricultural and environmental data are collected from soil sensors and meteorological sources. The raw data undergo preprocessing, including cleaning, normalization, and feature scaling. The processed data are then fed into multiple supervised classification models. Each classifier predicts the most suitable crop type, and the predictions are evaluated using standard classification metrics. The best-performing model is selected to support intelligent decision-making in smart farming environments.

Architecture Stages:

1. Data Acquisition (soil and climate data)
2. Data Preprocessing and Feature Scaling
3. Classification Model Training
4. Model Evaluation
5. Crop Prediction and Decision Support

Figure 3. Working architecture of Proposed Methodology

4.2 Experimental Setup

The experiments were conducted using the dataset described in Section 3.1, consisting of 2200 instances and seven input features. The dataset was split into 70% training and 30% testing samples, as discussed in Section 3.2. All models were trained and tested under identical conditions to ensure fair comparison. Performance was measured using accuracy, precision, recall, and F1-score.

4.3 Classification Performance Results

The classification accuracy results are summarized in Table 2.

Table 2. classification algorithms – Performance Results

Algorithm	Accuracy (%)	Precision	Recall	F1-Score
Logistic Regression	97.8	0.97	0.97	0.97
KNN	97.6	0.97	0.97	0.97
Decision Tree	98.4	0.98	0.98	0.98
Naïve Bayes	99.2	0.99	0.99	0.99
Random Forest	99.4	0.99	0.99	0.99
SVM	98.9	0.98	0.98	0.98
XGBoost Classifier	**99.6**	**0.996**	**0.996**	**0.996**

4.3.1 Visualization of the results

The bar chart demonstrates that ensemble-based classifiers significantly outperform traditional models. Logistic Regression and KNN show strong baseline performance but fall slightly behind ensemble approaches. XGBoost achieves the highest accuracy due to its boosting mechanism and regularization capabilities. The confusion matrix of the XGBoost classifier shows strong diagonal dominance, indicating that most crop classes are correctly predicted. Misclassification is minimal and primarily occurs among crops with similar soil and climate requirements. Precision, recall, and F1-score trends closely follow accuracy results. Ensemble models maintain balanced precision and recall, indicating stable generalization across crop classes. Traditional classifiers exhibit slightly lower recall for minority crop classes shown in Figure 4.

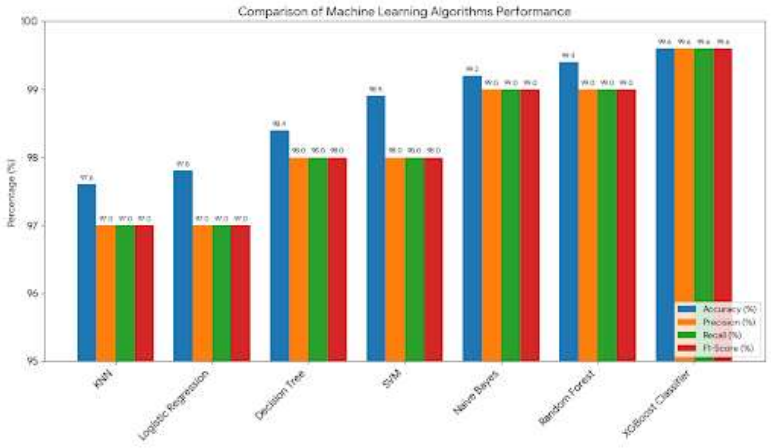

Figure 4: Comparison of Machine Learning Algorithms Performance

4.4.2 Discussion

The experimental results clearly demonstrate the effectiveness of supervised machine learning classifiers for crop prediction. Baseline models such as Logistic Regression and KNN provide satisfactory performance, confirming that soil and climatic attributes are strong predictors of crop suitability. However, these models are limited in capturing complex nonlinear relationships.

Decision Tree models offer interpretability but are prone to overfitting, which impacts their generalization ability. Naïve Bayes performs remarkably well due to the structured nature of the dataset, despite its independence assumptions.

Random Forest improves classification robustness by aggregating multiple decision trees, reducing variance and improving accuracy. Support Vector Machine demonstrates strong generalization, particularly for high-dimensional data, but requires careful parameter tuning.

Among all evaluated models, XGBoost Classifier achieves the best overall performance. Its boosting framework effectively minimizes classification error by iteratively correcting misclassified samples. Regularization and efficient optimization further enhance its robustness, making it highly suitable for large-scale smart farming applications [8], [15].

5. CONCLUSION

This study presented a systematic evaluation of supervised machine learning classification algorithms for crop prediction using soil nutrient and climatic features within a smart farming framework. The experimental results demonstrate that machine learning models can effectively capture the relationship between environmental conditions and crop suitability. Traditional classifiers such as Logistic Regression, K-Nearest Neighbors, Decision Tree, and Naïve Bayes provided strong baseline performance, while Support Vector Machine showed reliable generalization across multiple crop classes. The consistent performance across models confirms the effectiveness of data-driven approaches for intelligent crop prediction.

Among all evaluated methods, ensemble-based classifiers achieved superior results, with the XGBoost Classifier delivering the highest accuracy and most balanced performance across evaluation metrics. Its boosting mechanism and regularization capabilities enabled robust learning and error minimization, making it well suited for complex agricultural datasets. The findings highlight the potential of advanced ensemble learning techniques for deployment in IoT-enabled smart agriculture systems, supporting informed decision-making, optimized resource utilization, and sustainable crop management.

REFERENCES

[1]. T. van Klompenburg, A. Kassahun, and C. Catal, “Crop yield prediction using machine learning: A systematic literature review,” Computers and Electronics in Agriculture, vol. 177, p. 105709, 2020.
[2]. L. Li et al., “Developing machine learning models with multi-source environmental data to predict wheat yield,” Computers and Electronics in Agriculture, vol. 194, p. 106790, 2022.
[3]. M. Kuradusenge et al., “Crop yield prediction using machine learning models,” Agriculture, vol. 13, no. 1, p. 225, 2023.

[4]. A. Gehlot et al., “Crop production prediction using machine learning algorithms,” in Proc. IEEE International Conference on Innovative Computing, Intelligent Communication and Smart Electrical Systems (ICSES), 2022, pp. 1–5.
[5]. R. Ranjan, R. Garg, and J. K. Rai, “Artificial intelligence applications in soil and crop management,” in Proc. IEEE IATMSI, 2022, pp. 1–6.
[6]. E. Elbasi et al., “Crop prediction model using machine learning algorithms,” Applied Sciences, vol. 13, no. 16, p. 9288, 2023.
[7]. S. Vashisht, P. Kumar, and M. C. Trivedi, “Support vector machine–based crop yield prediction,” in Proc. IEEE ICIEM, 2022, pp. 754–757.
[8]. T. Chen and C. Guestrin, “XGBoost: A scalable tree boosting system,” in Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, 2016, pp. 785–794.
[9]. W. Xu et al., “Smart farm systems based on machine learning,” in Proc. IEEE ICET, 2021, pp. 417–421.
[10]. M. Senthil Kumar and D. S. Mary, “Smart farming using machine learning techniques,” Decision Analytics Journal, vol. 3, p. 100041, 2022.
[11]. S. Wolfert, L. Ge, C. Verdouw, and M.-J. Bogaardt, “Big data in smart farming – A review,” Agricultural Systems, vol. 153, pp. 69–80, 2017.
[12]. J. Kamilaris, A. Kartakoullis, and F. Prenafeta-Boldú, “A review on the practice of big data analysis in agriculture,” Computers and Electronics in Agriculture, vol. 143, pp. 23–37, 2017.
[13]. A. Khanna and S. Kaur, “Evolution of Internet of Things (IoT) and its significant impact in the field of precision agriculture,” Computers and Electronics in Agriculture, vol. 157, pp. 218–231, 2019.
[14]. M. Chlingaryan, S. Sukkarieh, and B. Whelan, “Machine learning approaches for crop yield prediction and nitrogen status estimation,” Computers and Electronics in Agriculture, vol. 151, pp. 61–69, 2018.
[15]. G. Liakos, P. Busato, D. Moshou, S. Pearson, and D. Bochtis, “Machine learning in agriculture: A review,” Sensors, vol. 18, no. 8, p. 2674, 2018.
[16]. I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning, MIT Press, 2016.