Journal of Information Systems Engineering and Management
2025, 10(63s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Machine Learning in Precision Agriculture

Aviral Jain[1], Dr. Umang Soni[2]
[1] *Student, The Shri Ram School, Aravali, Gurgaon*
[2] *Assistant Professor, Netaji Subhas University of Technology, New Delhi*

ARTICLE INFO

Received: 30 Dec 2024

Revised: 05 Feb 2025

Accepted: 25 Feb 2025

ABSTRACT

Introduction: Agriculture has been a fundamental aspect of human existence for thousands of years, dating back to around 9000 BC, when humans began transitioning from a nomadic lifestyle to settled farming [1]. Although agriculture is usually considered an "old-fashioned occupation," recent advancements in machine learning have paved the way for transforming agricultural forecasting through intelligent crop and yield prediction systems. Despite this, many farmers still use traditional crops and yield prediction methods that are often manual, data-scarce, and inaccurate.

Objectives: This research investigates the potential benefits of integrating machine learning into agricultural forecasting, which can reduce the risk of crop failure and financial loss associated with traditional methods. The paper aims to make technology more accessible and actionable rather than just theoretical.

Methods: Previous studies have applied multiple machine learning algorithms, such as Random Forest and Support Vector Machine (SVM), for crop prediction. Despite promising results, little work has been done on combining both crop type and yield predictions into an integrative machine learning framework. This paper explores the utilization of ensemble-based models, namely XGBoost, LightGBM, and Random Forest, trained on two datasets to accurately predict crop type. Additionally, it examines the use of regression-based algorithms to predict crop yield accurately, employing feature selection and 5-fold cross-validation.

Results: The ensemble-based models returned an accuracy of 98% for both crop type prediction and yield forecasting, showing the effectiveness of multiple algorithms. The algorithms also achieved an accuracy of 99.8% on the less comprehensive dataset, while individual models such as CatBoost achieved varying accuracies highlighted in Table 1.

Conclusions: The findings obtained in this study can help future farmers reduce effective costs, increase production rates, and enhance crop yield by minimizing waste. This approach contributes towards the realization of AI-driven, data-centric precision agriculture, optimizes resource utilization, and supports intelligent decision-making in modern farming.

Keywords: crop prediction; yield prediction; machine learning; 5-fold cross validation; integrative machine learning framework; AI-driven data-centric precision agriculture.

1. INTRODUCTION

World population continues to grow every year, due to which much more effort and innovation is required to increase agricultural production, improve global food security and decrease food losses and wastage [2]. In many developing countries, such as India and Brazil, agriculture plays an important role in economic growth and development, making improvements in agricultural practices crucial for social and economic stability. As of 2024, almost 18% of India's Gross Domestic Product (GDP) came from agriculture, and it provides livelihood support to about 43% of the nearly 1.5 billion people living in India [3].

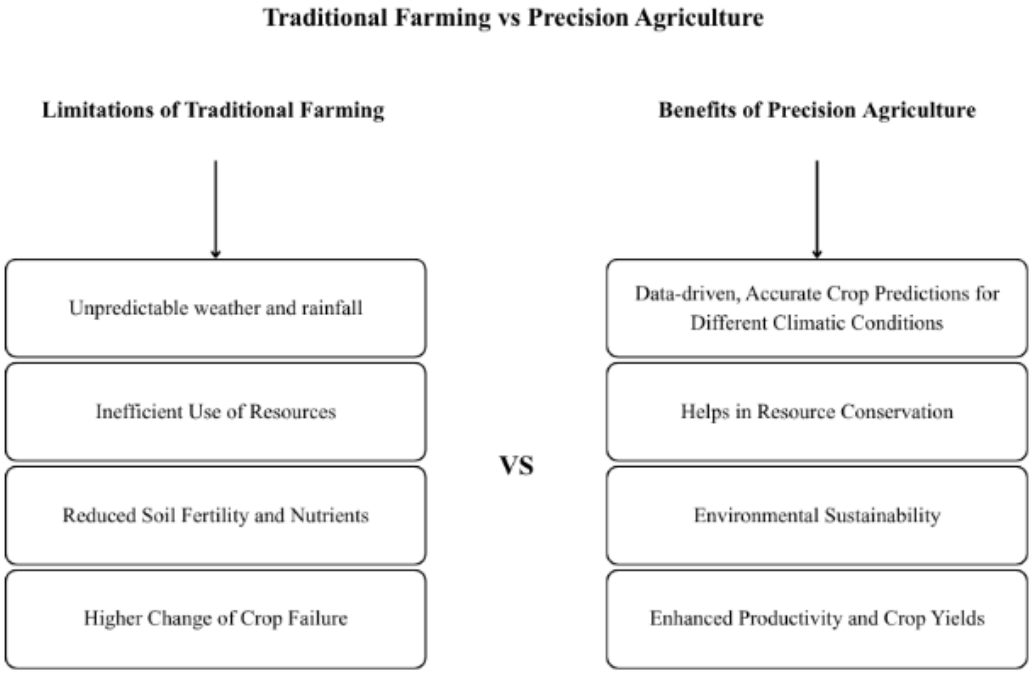
However, agriculture is very vulnerable to soil degradation, pest outbreaks, and unpredictable climate conditions. Farmers usually rely on experience-based judgement to make decisions about crop type selection and expected yield, which can lead to inaccurate results and financial losses. This also contributes to inefficient use of resources such as water and fertilizers.

Journal of Information Systems Engineering and Management
2025, 10(63s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Smart farming, or precision agriculture, is a type of agricultural practice where modern data-driven technologies are used for growing crops. Compared to traditional agriculture, precision agriculture has many benefits: it provides an improved understanding of the landscape, which can be integrated with high-accuracy prediction systems to predict suitable crops, help farmers make decisions on how much fertilizer to spray, when and how much to water, and allows for greater flexibility in adaptation of crops to the given climate conditions. The two main technologies involved in precision agriculture, i.e. Artificial Intelligence and Machine Learning are evolving too, and can be combined to make accurate predictions [4].

Figure 1: Graphical Representation of Traditional Farming vs Precision Agriculture



The crop analysis done in this paper uses datasets that have information from IoT sensors. These sensors can be utilized to measure soil moisture, temperature, the ratio of Nitrogen, Phosphorus, and Potassium in the soil, and other soil parameters. This data can then be used to train machine learning models that can not only predict crop type and yield but also help improve them significantly.

"Machine learning is the subset of Artificial Intelligence focused on algorithms that can 'learn' the patterns of training data and, subsequently, make accurate inferences about new data [5]." Algorithms are blocks of code that can be written and trained on datasets to make predictions. These algorithms can analyze large amounts of data from IoT sensors and help farmers maximize profits while minimizing the risk of failure.

Despite the several benefits of machine learning in agriculture, there are a few limitations as well. For instance, the lack of consistent data, high initial costs for sensors and computing infrastructure, and a risk of technological dependency, where farmers rely too heavily on automated predictions without self-judgement, leading to undesirable results. Furthermore, such advanced tools may be only available to large agricultural farm owners while smaller ones, that populate the majority of India, are left out. However, as more farms implement precision agriculture and gather data, the potential profits of deploying machine learning in agriculture will become more evident. So far, results have been promising, and machine learning will likely continue growing in the agricultural industry [6].

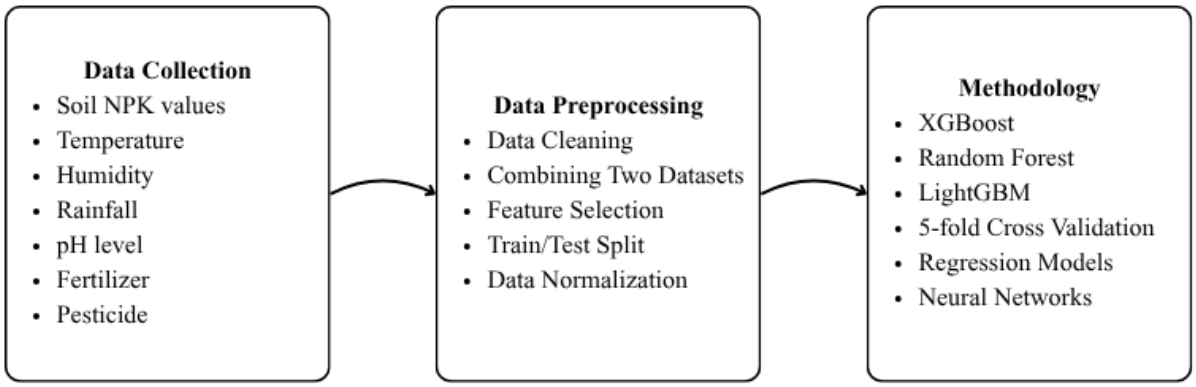
Many researchers have already used machine learning in agriculture. Several machine learning models, such as Random Forest, Convolutional Neural Network, Support Vector Machine, Bayesian Network, and others to predict crops or yield. These studies have excellent results with often an accuracy above 99%, proving that machine learning is very helpful in agriculture. However, most of these studies only focused on one prediction at a time, either crop prediction or yield prediction, not both together. This limits the usage of these models in real farming situations [7].

Journal of Information Systems Engineering and Management
2025, 10(63s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Moreover, most available research papers have been trained on one dataset, and very few include ensemble models that combine multiple algorithms for better accuracy. Hence, there is a need for more "complete" systems that can provide both suggestions and predictions and include multiple algorithms and datasets [7].

Figure 2: IoT and Machine Learning Based Crop Analysis and Prediction Process



The framework built in this paper is an integrative machine learning framework which uses an ensemble model of 3 algorithms, XGBoost, LightGBM, and Random Forest, to accurately predict crop type and yield. It aims to predict the best crop type for the soil, weather conditions, and quantity of fertilizers. The model also predicts crop yield based on similar parameters. Additionally, the model uses 5-fold cross-validation and feature selection to enhance accuracy.

Agriculture serves as a backbone for food security and economic stability. It is also one of the largest sources of employment in many developing nations. The change from traditional farming to intelligent, data-driven precision agriculture has a vast potential to increase productivity and profit, saving money by using the correct amount of water, fertilizer, and other resources, making farming more sustainable, and improving decision-making with accurate predictions. Machine learning can accelerate the shift from traditional agriculture to precision agriculture, marking an important step towards the future.

2. LITERATURE REVIEW

To further understand the current progress and challenges in applying machine learning to agriculture, existing studies on crop prediction and crop yield forecasting must be reviewed. This section reviews literature [8-16] and discusses the methodology of each briefly. This will help identify gaps in previous studies and assist future researchers in successfully fill in those gaps.

In [8], machine learning algorithms have been used for crop prediction. The algorithms used are K-Nearest Neighbor Classifier (KNN Algorithm), Decision Tree Classifier with Entropy and Gini Index, and Random Forest Classifier. The authors analyzed previous works, as well as their proposed work, which achieved an accuracy of 99.32% on a Crop Recommendation Dataset from Kaggle [Dataset 1]. The research offers a clear and concise comparison of supervised Machine Learning Algorithms and uses a commonly used dataset from Kaggle. Additionally, it shows the benefits of using an ensemble-based model. But, the authors have not included multiple datasets, although they wrote about the advantages of using several datasets for crop prediction, and have not mentioned feature selection and cross-validation.

In [9], too, the authors have built machine learning algorithms for crop prediction. The study compares 15 different classifiers on a Crop Recommendation Dataset from Kaggle [Dataset 1], demonstrating that probabilistic models achieve high accuracy in predicting suitable crops. The algorithms they used were the Naïve Bayes classifier, Bayes Net, Hoeffding Tree, Decision Tree with Entropy and Gini criterion, Support Vector Machine, and others, and achieved the highest accuracy with Bayes Net, i.e. 99.59%, with Naïve Bayes and Hoeffding Tree following closely at 99.46%. The paper is well researched and explores feature selection combinations alongside label changes, which affected prediction accuracy. A potential drawback is that the paper uses a regional is dependent on an individual dataset, that might lead to potential region or ranking bias.

Journal of Information Systems Engineering and Management
2025, 10(63s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

In [10], authors have predicted agriculture yields based on machine learning using regression and deep learning. The primary aim of the research "is to predict crop yield utilizing the variables of rainfall, crop, meteorological conditions, area, production, and yield that have posed a serious threat to the long-term viability of agriculture." To forecast crop yield, the study uses Decision Tree, Random Forest, XGBoost Regression, Convolutional Neural Network, and Long-Short Term Memory Network. The dataset has been created using information gathered from publicly available official websites [Dataset 2]. The paper achieved great results, with the accuracy of Random Forest being nearly 99%, and Decision Tree and XGBoost being a little lower at 90% and 86.5% respectively. The paper also uses mean-absolute error, standard deviation, and other parameters have been used to validate results. However, the dataset only contains local parameters and limits field-level applicability, and the paper talks little about the real-world implications of their results.

[11] Talks about crop selection and yield prediction using a machine learning approach. The primary objective of the study is to help farmers plan cultivation by predicting both crop type using classification and crop yield using regression. The paper also talks about improving land utilization and reducing financial risk for farmers. 3 datasets have been used in the paper, all of which focus on Maharashtra, India. The datasets include information about almost 25 crops, including both soil and weather features. The paper uses 4 main algorithms for classification, i.e. Naïve Bayes (99.39%), Random Forest (99.24%), KNN (97.72%), and Logistic Regression (94.69%). Additionally, 3 algorithms for yield prediction, i.e. Random Forest regressor (MAE = 0.64, R^2 = 0.96[1]), and Decision Tree Regressor (R^2 = 0.94). The paper has successfully integrated both crop selection and yield prediction. However, the dataset only includes information from Maharashtra and will struggle to generalize for other areas, and it does not include IoT sensor data either.

Paper [12] is a detailed analysis of crop yield prediction using machine learning. Its primary objective is to predict crop yield using soil and environmental parameters and improve farmer decision-making to avoid crop losses. A dataset from Kaggle has been used that includes area, production, state, crop-wise yield, and soil, climate parameters for accurate yield prediction. 3 machine learning models have been used, i.e. Linear Regression, Decision Tree Regression and Random Forest Regression. The models have been evaluated using RMSE (Root Mean Square Error) and R^2 score. However, no numeric values have been given for the performance of each model, but Random Forest Regression has performed the best with the highest R^2 score. The paper is also not documented well, and no real-time data sources have been used to validate the dataset.

Paper [13] is about using machine learning for crop yield prediction in the past or the future. The study explores how well different machine learning algorithms can predict future crop yields for wheat and sunflowers in Spain. Key parameters used are data size, soil depth, sowing date, N-fertilizer, irrigation, and weather. The study uses Random Forest, Artificial Neural Networks with different hidden-node sizes, and Linear Regression Models such as Ridge & Lasso regularization. Overall, these are unique models and are not often used. The dataset is also fully synthetic. The study found that Random Forest was the best model for predicting future crop yields, while ANN frequently overfitted and performed quite worse. A major drawback of a synthetic dataset is that it does not include real-world uncertainties like pests and climate change anomalies. The study also focused solely on a small geographic region and on two crops only.

Paper [14] aims to develop a model that can recommend the best suitable crop for a given region by integrating Genetic Algorithm (GA) for feature selection, and a Machine Learning classifier for prediction. The datasets have been collected from Kaggle and Government portals that include rainfall, temperature, pH, soil type, and nutrients. A comparison has been made for predictions before and after the use of GA. The paper claims to have improved accuracy after applying GA but no numerical accuracy data has been provided. This is indeed an innovative GA+ML hybrid framework and is a new approach towards precision agriculture. But poor documentation and very few ML models used in the paper can be a drawback.

[15] This paper focuses on crop selection to maximize crop yield rate using ML techniques. The problem it addresses is that farmers often have multiple crop options but limited land, and choosing the right crop at the right time is

[1] MAE = Mean Absolute Error, R^2 = Accuracy for regression models (on a scale from 0.0-1.0, 1.0 being a perfect score)

Journal of Information Systems Engineering and Management
2025, 10(63s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

challenging. The paper proposes a Crop Selection Model (CSM) that can predict the best crop sequence based on plantation duration, and predicted yield rate. A farmer sourced dataset has been used in this paper including parameters such as sowing/harvest time, plantation duration, predicted yield per hectare, and crop type category. The models it uses are Artificial Neural Networks (ANN), SVM, Gradient Boosting Techniques and Decision Trees, Random Forest alongside many others. The paper was accurately able to produce multiple crop sequences however, model improvement is possible. This was one of the first papers that proposed the idea of **crop sequence optimization**. But, the paper does not include detailed ML performance metrics such as R^2 and MAE to check validity and is based on regional data only.

[16] This paper builds a ML model for crop and fertilizer recommendation. The main objective of the paper is to design a system that can recommend the most suitable crop and the most optimal fertilizer for a given piece of agricultural land based on soil and weather parameters. Two datasets have been used for the same from Kaggle, Crop Recommendation Dataset and Fertilizer Recommendation Dataset. 3 algorithms have been used for crop prediction namely, XGBoost, Random Forest and KNN meanwhile fertilizer prediction uses SVM and Random Forest. Additionally, the system can predict yield per acre, required seed amount, market price to support farmer decision making although it has not been validated by a concrete dataset. The system has achieved an accuracy of 98% but it lacks scientific benchmarking. Also, the paper does not explore cross validation or ensemble-based models in detail limiting its scientific reliability in the real world.

Most existing studies focus only on either crop type classification or yield prediction, lack proper validation, or depend on limited regional datasets. To address these limitations, this paper proposes:

1. Developing an integrated machine learning framework that combines crop prediction and yield forecasting
2. Using gradient boosting models such as XGBoost and LightGBM to enhance accuracy of model and 5-fold cross-validation alongside hyperparameter tuning[2]
3. Analyzing two unique crop recommendation datasets that contain several soil and climate parameters such as NPK, pH, rainfall, temperature and humidity obtained from IoT sensors
4. Providing ML evaluation metrics such as MAE, R^2 and confusion matrices to ensure proper scientific benchmarking and for the purpose of future research
5. Emphasizing real-world usability of the information in this paper by focusing on maximizing crop yields, optimizing resource usage and reducing overall costs for farmers.

3. METHODOLOGY

This section describes the methodology used to develop crop type classification and yield prediction models using machine learning.

3.1 Data Collection

3 datasets have been utilized in this research. 2 datasets have been used for crop type prediction and 1 for yield forecasting. These datasets contain information from IoT sensors to capture the key parameters needed for crop and yield forecasting. Crop Recommendation Dataset from Kaggle contains 2200 rows and 22 different crop types. It also has 7 soil and climate parameters based on which accurate crop predictions can be made. Crop Recommendation using Soil Properties and Weather Prediction is another dataset from Mendeley. It contains almost 4000 rows and over 20 different soil and climate parameters. Lastly, Crop yield is a yield prediction dataset from Kaggle that has almost 20000 rows encompassing over 50 different crops and 5 parameters based on which yield can be forecasted.

3.2 Data Preprocessing

In order to make the data ready for prediction, it is necessary to clean the dataset. This also improves model accuracy and reliability. The data is preprocessed by the following steps.

a. Removal of missing values

[2] Hyperparameter tuning: Adjusting the model's external settings (like learning rate or number of trees) to improve performance before training. 5-fold cross-validation: A validation method where the dataset is split into 5 parts, training on 4 and testing on 1 repeatedly, to ensure reliable accuracy.

Journal of Information Systems Engineering and Management
2025, 10(63s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

b. Trimming spaces in categorical entries
c. Label encoding for categorical entries
d. Outlier handling using 1% quantile cutoff[3]
e. Feature scaling using StandardScaler[4]
f. Dataset split into training and testing sets

3.3 Data Modelling

Two different crop datasets containing soil and climate properties were integrated to create a unified source for classification. This required standardizing column names and formats to ensure consistency. Then the relevant agricultural input features were selected and the others were left out, called feature selection. The important parameters were soil nutrients (NPK), temperature, humidity, rainfall, pH, and micronutrients (Zn, S). Sensor readings had to be averaged into single indicators using python libraries such as NumPy and pandas. Then, data preprocessing steps from a-f were followed for the new dataset and it was prepared for training an ensemble of ML models to accurately predict crops and yield.

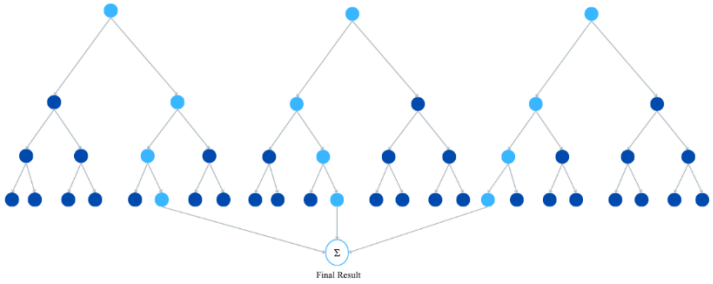
3.4 Machine Learning Models

This paper uses 8 different models for crop prediction and yield forecasting. The constructed model was trained using training data, then the results were evaluated using test data to ensure that predictions were accurate.

The following algorithms were used in our research:

1. *Random Forest* is a popular model in agricultural prediction and research and is a strong baseline decision tree algorithm. It is a supervised ML algorithm. It basically builds an ensemble of decision trees[5] and combines them to attain more accurate predictions. Random Forest can be used in classification as well as regression systems. Random forest does not search for the most important feature while making a decision tree, it searches for the best one that adds additional randomness to the model, helping it generalize better for various situations. In Random Forest regressor, average of the trees' outputs is taken while in classifier a majority vote is taken [18]. A few key benefits of Random Forest are reduced risk of overfitting[6], provides flexibility as one can use its classifier or regressor based on the task, and it's easy to determine feature importance. An important thing to keep in mind is the number of trees. More the number of trees, higher will be the accuracy but higher is the risk of overfitting. The Random Forest algorithm used in the paper was run with 200 trees, 400 trees, 600 trees, 800 trees and 1000 trees. The model achieved its best accuracy around 600-800 trees.

Figure 3: How a Random Forest Algorithm Works. The dark blue circles are part of 3 decision trees and the light blue circles highlight the path taken by the algorithm while decision-making. The model obtains the final result by averaging the result of every individual decision tree.



[3] Outlier handling using a 1% quantile cutoff removes the lowest 1% and highest 1% of numerical values to eliminate unusual extremes that can distort the model.
[4] StandardScaler rescales features to a common range by standardizing them to mean = 0 and standard deviation = 1, which helps machine learning models learn more effectively.
[5] "A decision tree is a supervised machine learning algorithm used for classification or regression tasks [17]."
[6] Overfitting is when a ML model learns the data too well, capturing random fluctuations in the middle causing it to become overly specific on the training piece of the data. This makes it perform poorly on the test set.

Journal of Information Systems Engineering and Management
2025, 10(63s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

2. XGBoost, or eXtreme Gradient Boosting, is an advanced ML model built for efficiency, speed and high performance. XGBoost, like Random Forest, uses decision trees for learning. However, XGBoost combines decision trees sequentially to improve the model's performance and each new tree learns from the previous one's errors to achieve higher accuracy. This is basically an advanced form of Gradient Boosting. It can be viewed as an iterative process, that starts with an initial prediction. After which multiple decision trees are added to reduce errors. Mathematically, the model can be represented as

$$\hat{y}^i = \sum_{k=1}^{K} f_k(x_i)$$

Where:
$\hat{y}^i$ is the final predicted value for the i^{th} data point
K is the number of trees in the ensemble
$f_k(x_i)$ represents the prediction of the K^{th} tree for the i^{th} data point.
A few key advantages of using XGBoost are reduced risk of overfitting, handles complex features effectively, can easily handle missing data, ideal for large datasets such as in our case [19].

3. LightGBM, Light Gradient Boosting Machine, is an open-source high-performance gradient-based classification framework developed by Microsoft. It uses decision trees that grow efficiently by reducing memory usage and running time. Furthermore, it is designed for scalability and high accuracy with large datasets. A unique feature of LightGBM is that its trees grow leaf-wise that helps it to focus on the section where prediction error is highest, resulting in deeper and more informative models. LightGBM also contains a regression model.

   Training using LightGBM involves a few key steps. Cross-validation techniques in the algorithm are usually implemented through it to validate a model's performance. LightGBM also involves hyperparameter tuning that helps it optimize the settings which govern the performance of the model during training.

   A few advantages of LightGBM are that it is faster and more accurate than other gradient boosting algorithms, it uses less memory, and is very effective of large datasets [20].

4. CatBoost, Categorical Boosting, is a ML model that is intended to perform well in regression and classification tasks. It can handle categorical values without the need for manual encoding. This saves user time and effort. The main algorithm implemented in CatBoost is called Ordered Boosting.

   CatBoost also utilizes decision trees to construct a powerful, highly accurate model. It works in a similar way to XGBoost, with every new tree correcting the errors of the previous ones. With CatBoost's ability to handle categorical data well, it makes it one of the most useful algorithms for real-world datasets where time can be saved from label encoding and data preprocessing [21].

   "One of the core innovations of CatBoost is its ordered boosting mechanism [22]." CatBoost avoids prediction shift[7] as it introduces a random permutation of the dataset in each iteration that can help reduce overfitting and increases model reliability.

5. Support Vector Machine (SVM) is a supervise ML algorithm used for regression and classification tasks. SVM works by finding the best boundary, called hyperplane, which separates different classes in the data. The primary goal of this model is to maximize the margin between the two classes separated by the hyperplane. The more the margin, higher will be the model's accuracy on new and untrained data. The best hyperplane is known as the "hard margin" and maximizes distance between the nearest datapoints from both classes and the hyperplane.

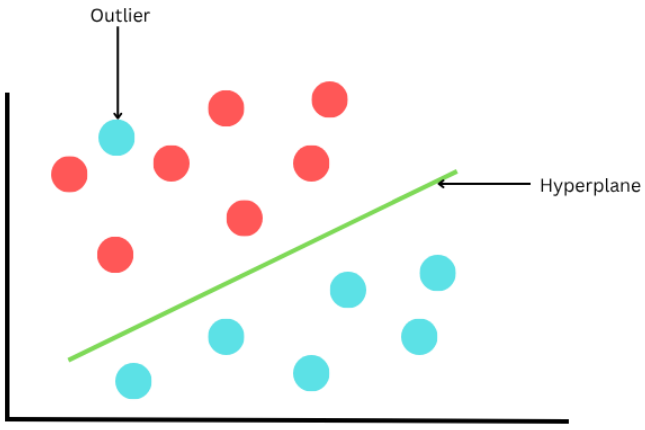
   In Figure 4, one of the blue balls has been mixed with the red ones. This blue ball is referred to as an outlier and a unique property of SVM is to be able to ignore that outlier and work on the remaining data.

Figure 4: Here one of the blue balls has been mixed with the red ones[8]

[7] Prediction shift is a phenomenon where a model's predictions on test data are different from its predictions on training data.
[8] Figure inspiration from GeeksforGeeks [23]

Journal of Information Systems Engineering and Management
2025, 10(63s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article



If data is not linearly separable, SVM can use a technique called kernels that can map the data into a higher-dimensional space where it becomes separable. Kernels are mathematical functions that allow the algorithm to perform linear classification on non-linear data. This paper will not be going into types of kernels, but that information can be found at reference [23-24].

SVM excels in high-dimensional spaces making it suitable for image analysis. It is also compatible with non-linear data and outlier resistance makes it suitable for large data. However, SVM is quite slow it is difficult to perform hyperparameter tuning through this algorithm.

6. Naïve Bayes is an ML classifier that predicts the category of data based on probability. The main concept behind this algorithm is the Bayes' Theorem. Bayes' Theorem is a mathematical formula used to determine the probability of an event happening based on prior knowledge and evidence.

   A great example of Bayes' Theorem is the Monty Hall Problem. Here, assume you are a contestant on a reality show and there is a host. The host has 3 doors, 1 with a prize behind it and 2 with goats. The host knows what's behind each door and you don't. You are asked to choose a door. Now, the host opens one of the doors and there is a goat behind it. The question is that would you stay with your current choice or switch. Initially, every door had a roughly 33% chance of having the prize. When the host opens a door, he reveals "new information." This is where Bayes' Theorem comes in. You have to update the probability given that now only 2 doors are left. Once only 2 doors are left, your choice still has 33% chance of having the prize, but since one door has been eliminated now, the other door takes the entire 66% chance of having the prize. Therefore, it's best for you to switch.

   The expression behind Bayes' Theorem is:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

   Where:

   P(A) and P(B) are the probabilities of events A and B, assuming P(B) != 0

   P(A|B) is the probability of event A happening when event B happens and P(B|A) is the vice-versa.

   Naïve Bayes is built on very few parameters that can predict at faster speeds than other classification algorithms. It is also a probabilistic classifier, that means each feature contributes to the predictions with no relation to each other.

   There are 3 types of Naïve Bayes models, Gaussian NB (the one this paper utilizes), Multinomial NB, and Bernoulli NB. The working of each of these models is beyond the scope of this paper.

   A few advantages of using the Naïve Bayes classifier are it is easy to implement and efficient; it is effective when there is a large number of parameters; it performs well even with limited training data; and it can utilize

Journal of Information Systems Engineering and Management
2025, 10(63s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

categorical features well. A drawback of Naïve Bayes is that it may not generalize well as it assigns zero probability to unseen events, decreasing accuracy in a few cases [25-26].

7. Regression is a technique in ML when a continuous numerical value is predicted based on one or more independent features. It is able to find relations between variables to make predictions. The two types of variables involved here are dependent variables, the variables we are trying to predict, and independent variables, the input variables which influence dependent variables. There are different types of regression models. This paper utilizes non-linear regression models such as XGBoost, CatBoost, and LightGBM (regression version of each of these three). This paper does not go into detail about the other types of regression types and their advantages, but that can be found at reference [27].

4. RESULTS

This section of the paper will talk about the results achieved by the various models deployed. It will also include various performance metrics used to analyze model performance and the working behind them. Classification models for crop prediction have been validated through 4 metrics, accuracy, F-1 score, Kappa, and Standard Deviation.

1. Accuracy measures how well the algorithm performs on the test data after being trained on the training data. It is on a scale from 0-100% where 100% is the highest and means that the model is perfect.
2. F-1 score is a performance metric used to evaluate how well a classification model performs on an imbalanced dataset, where one class appears more frequently than others. Before understanding F-1 score, we need to understand 3 key terms [29] –
   a. Harmonic mean – "It is defined as reciprocal of the arithmetic mean of the reciprocals of the given set of values [28]." Mathematically, if x_1, x_2, x_3,...x_n are n non-zero positive observations,

$$Harmonic\ mean\ formula = \frac{n}{\frac{1}{x_1}+\frac{1}{x_2}+\frac{1}{x_3}+\cdots+\frac{1}{x_n}}$$

   Note: Harmonic mean is always less than the Arithmetic Mean or Geometric Mean. [28].

   b. Precision – It is a measure of accuracy of the positive predictions. For example, if there are 5 positive cases and 5 negative cases. The model can identify 5 positive cases, but out of those 5 identified cases, only 3 are positive and 2 are negative. Then precision is 60%.
   c. Recall – it is the ratio of predicted positive cases to the actual number of total positive cases. Taking the previous example, recall will be 3/5 or 60%.

   F-1 score combines precision and recall using harmonic mean –

$$F_1\ score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

3. Kappa, also known as Cohen's Kappa, "is a statistical metric that measures the reliability of two raters who are evaluating the same thing, accounting for the possibility that they might agree by chance [30]. It always ranges from 0-1, where 0 is no agreement between the two raters, and 1 is perfect agreement between the two raters. Mathematically, it is:

$$k = \frac{(p_o - p_e)}{(1 - p_e)}$$

   Where p_o is the proportion of agreement between the two raters and p_e is the expected proportion of agreement by chance.

   This paper does not go into much detail about Cohen's Kappa, but more information about it can be found at reference [30-31].

4. Standard deviation is a statistical measure that tells the variation in data points. It helps understand how the values are spread out from the mean. A lower standard deviation means that the values are relatively consistent and a higher standard deviation means that there's more variability in the data.

Journal of Information Systems Engineering and Management
2025, 10(63s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

4 other performance metrics have been utilized for crop yield prediction using regression models, namely R^2, Mean Absolute Error, Root Mean Square Error and Relative Absolute Error.

1. R^2 is a statistical measure that represents how well a regression model fits into the data. Its value is between 0 and 1, where 1 represents a perfect fit and there's no difference between the predicted and actual value, and 0 represents no value predicted by the model [32].
2. Mean Absolute Error (MAE) calculates the average difference between the calculated values and the actual values [33]. A lower MAE indicates a more accurate model as the difference is low, and a higher MAE indicates large deviation between predictions and actual values.
3. Root Mean Square Error (RMSE) is basically just MAE, but it penalizes large errors more as it squares the errors. It is also more sensitive to outliers and only a few large errors, can drastically increase its value [34].
4. Relative Absolute Error (RAE) is another metric used to analyze accuracy of a regression model. It first calculates sum of the absolute differences between predicted and actual values for all data points. Then it predicts the average of actual values for all data points called baseline error. This can also be the total absolute error of a naive model[9]. Next, it divides the model's total absolute error by the baseline error to get the RAE [36].

Table 1 shows the accuracies of various models on both datasets used in this study. We can see that all models performed better on Dataset 1 than Dataset 2. This can be as Dataset 2 includes nearly 20 soil and weather parameters, which can be harder to predict and contains almost double the number of rows as compared to Dataset 1. These models have also achieved exceptional F-1 scores and Cohen's Kappa, further validating their performance. Overall, the ensemble model of XGBoost + LightGBM + Random Forest performed the best on both datasets. To achieve higher accuracy, both datasets were combined, and the ensemble model was trained on both, that returned an accuracy of 97.99%, considering almost 6000 rows, 60 different crop types, and nearly 30 soil and weather conditions.

Table 1: Accuracies and Performance Metrics of Different ML models for Crop Prediction

	Dataset 1[10]				Dataset 2[11]			
Model name	Accuracy	F-1 score	Cohen's Kappa	Standard Deviation	Accuracy	F-1 score	Cohen's Kappa	Standard Deviation
XGB+LGBM+RF	99.55%	0.9954	0.9952	0.19%	95.37%	0.9530	0.9495	0.34%
XGBoost	99.50%	0.9954	0.9952	0.27%	94.84%	0.9473	0.9437	0.27%
Random Forest	99.33%	0.9929	0.9928	0.32%	50.30%	0.4552	0.3393	0.20%
CatBoost	97.95%	0.9795	0.9786	NA	50.78%	0.4560	0.3441	NA
LGBM	99.22%	0.9922	0.9919	NA	86.28%	0.8591	0.8503	NA
SVM	98.85%	0.9883	0.9881	NA	66.87%	0.6477	0.6385	NA
Naïve Bayes	99.45%	0.9945	0.9943	0.29%	31.14%	0.2809	0.2509	0.62%

Table 2 shows the values of various performance metrics analyzed from Regression models used for crop yield prediction. It is evident that XGBoost Regressor and the ensemble model comprising of XGBoost, LightGBM, CatBoost and Random Forest performed the best. Low MAE and RMSE indicate minimal prediction error and the high R^2 indicates that the model can accurately map soil and climate factors to predict crop yield.

Table 2: Accuracies and Performance Metrics of Different Regression Models for Crop Yield Prediction

[9] A naive model is a simple, low-effort forecasting approach that predicts values based on a single historical value [35].
[10] Dataset 1: Crop Recommendation Dataset from Kaggle [1]
[11] Dataset 2: Crop Recommendation using Soil Properties and Weather [2]

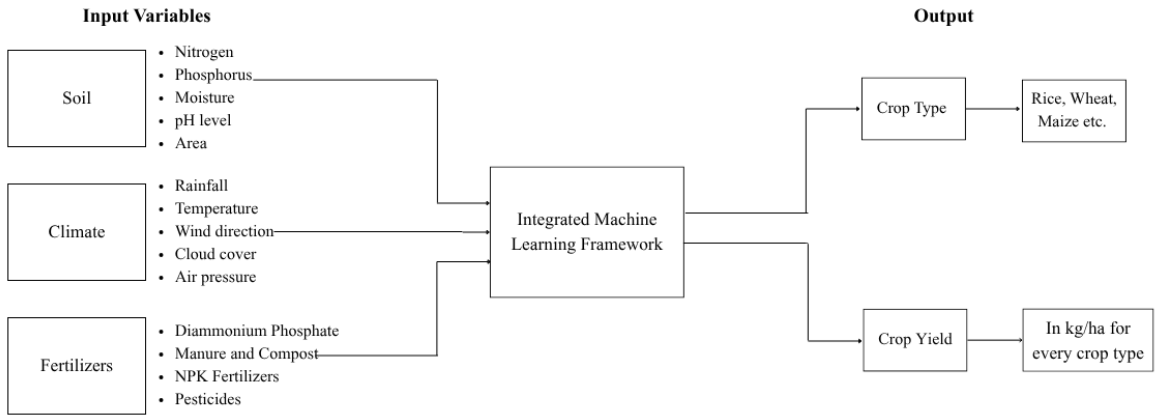
Journal of Information Systems Engineering and Management
2025, 10(63s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Regression Model	R square	MAE	RMSE	RAE	Standard Deviation
XGB Reg	0.9800	0.3644	1.1367	0.0853	NA
LGBM Reg	0.9796	0.4102	1.1577	0.0960	NA
CAT Reg	0.9757	0.4130	1.2630	0.0967	NA
RF Reg	0.9735	0.3882	1.3190	0.0909	NA
XGB+LGBM+CAT+RF	0.9800	0.3719	1.1438	0.0870	0.0028

To build the integrative machine learning framework, the classification ensemble based model and the regression ensemble model are combined. Then the model analyzes the input parameters and first predicts crop type, then predicts crop yield. Through this, the earnings of the farmer can be calculated if prices for various crops are known. Unfortunately, these prices differ drastically based on geographical location and no concrete dataset could be found for the same. A diagrammatic explanation of the framework has been explained through Figure 3.

Figure 5: Methodology of the Integrative Machine Learning Framework



For example, a farmer's soil has Nitrogen: 90, Phosphorus: 42, Potassium: 38, pH: 6.5, and the area receives moderate rainfall. The ML model processes these values and recommends "Rice" as the optimal crop. Since the model accuracy is 99.8%, out of 100 farms, rice will successfully grow in nearly all cases, with only about one farm possibly failing due to uncontrollable environmental factors.

Similarly, suppose a farmer has around 2000 hectares of land for cotton cultivation. The area receives 2000 cm of annual rainfall, and the farmer applies 500 kg of fertilizer. The model predicts that the farmer will obtain approximately 4000 kg of cotton per hectare.

5. CONCLUSION

By applying advanced machine learning in agriculture, this research demonstrates how machine learning can help the transition to precision agriculture. Integrating soil and climate conditions, feature scaling techniques and ensemble-based ML algorithms, the proposed system was able to achieve accurate results while predicting crop type and forecasting crop yield. This shows the effectiveness of using multiple datasets, proper data preprocessing and data balancing techniques that can contribute to model accuracy.

Integrating machine learning in agriculture reduces uncertainty by identifying the most suitable crops for specific soil and climate conditions. This enables more informed decisions that optimize the use of resources like water and fertilizer, minimize wastage, and enhanced productivity for higher earnings.

There remains significant potential for future research in this field. Neural networks can be used to analyze crop images for crop disease detection that can reduce crop wastage, and building of accurate crop price datasets through which crop price can be predicted too. Additionally, real time data from sensors can be used for real time predictions and information about crops through mobile applications can truly transform farming. The datasets used in this research are local to India only, but future research can incorporate datasets from various geographic locations and improve model generalization.

This study has attempted to bridge the gap between theoretical AI research and real-world agricultural decision making. Providing farmers with accessible precision-based agriculture solutions that can help them enhance their farming techniques is the next step towards precision agriculture, and machine learning will play a crucial role in it.

REFRENCES

Links

[1] Wikipedia contributors. (2025, October 14). History of agriculture. Retrieved from https://en.wikipedia.org/wiki/History_of_agriculture

[2] Food security and nutrition and sustainable agriculture | Department of Economic and Social Affairs. (n.d.). Retrieved from https://sdgs.un.org/topics/food-security-and-nutrition-and-sustainable-agriculture?utm_source=chatgpt.com

[3] AGRICULTURE SECTOR HAS REGISTERED AN AVERAGE ANNUAL GROWTH RATE OF 4.18 PER CENT OVER THE LAST FIVE YEARS : ECONOMIC SURVEY. (n.d.). Retrieved from https://www.pib.gov.in/PressReleasePage.aspx?PRID=2034943

[4] Precision agriculture. (n.d.). Retrieved from https://climate-adapt.eea.europa.eu/en/metadata/adaptation-options/precision-agriculture

[5] Bergmann, D. (2025, October 23). Machine learning. *Think*. Retrieved from https://www.ibm.com

[6] Koptelov, A. (2023, August 1). Machine Learning in agriculture: challenges and solutions. Retrieved from https://www.iiot-world.com/artificial-intelligence-ml/machine-learning/machine-learning-in-agriculture-challenges-and-solutions/

[7] https://scholar.google.com/

Literature

[8] Madhuri Shripathi Rao et al 2022 J. Phys.: Conf. Ser. 2161 012033

[9] Elbasi, E.; Zaki, C.; Topcu, A.E.; Abdelbaki, W.; Zreikat, A.I.; Cina, E.; Shdefat, A.; Saker, L. Crop Prediction Model Using Machine Learning Algorithms. Appl. Sci. 2023, 13, 9288. https://doi.org/10.3390/ app13169288

[10] P. Sharma, P. Dadheech, N. Aneja and S. Aneja, "Predicting Agriculture Yields Based on Machine Learning Using Regression and Deep Learning," in IEEE Access, vol. 11, pp. 111255-111264, 2023, doi: 10.1109/ACCESS.2023.3321861.

[11] Patil P, Athavale P, Bothara M, Tambolkar S, More A. Crop Selection and Yield Prediction using Machine Learning Approach. Curr Agri Res 2023; 11(3). Doi: http://dx.doi.org/10.12944/CARJ.11.3.26

[12] Patil, P., Kadam, S., Patil, A., & Gaikwad, P. (2023). *Crop yield prediction using machine learning techniques* (Unpublished B.E. project report). Department of Computer Engineering.

[13] Morales A and Villalobos FJ (2023) Using machine learning for crop yield prediction in the past or the future. Front. Plant Sci. 14:1128388. doi: 10.3389/fpls.2023.1128388

[14] T. Mahmud *et al.*, "An Approach for Crop Prediction in Agriculture: Integrating Genetic Algorithms and Machine Learning," in *IEEE Access*, vol. 12, pp. 173583-173598, 2024, doi: 10.1109/ACCESS.2024.3478739.

[15] R. Kumar, M. P. Singh, P. Kumar and J. P. Singh, "Crop Selection Method to maximize crop yield rate using machine learning technique," *2015 International Conference on Smart Technologies and Management for Computing, Communication, Controls, Energy and Materials (ICSTM)*, Avadi, India, 2015, pp. 138-145, doi: 10.1109/ICSTM.2015.7225403.

[16] Kumar, M. D. P., Malyadri, N., Srikanth, M. S., & Ananda Babu, J. (2021). *A machine learning model for crop and fertilizer recommendation*. Natural Volatiles & Essential Oils, 8(6), 4570–4580.

Links continuation

[17] Corbo, A. (2025, June 3). What is a decision tree? Retrieved from https://builtin.com/machine-learning/decision-tree
[18] Donges, N. (2024, November 26). Random Forest: A complete guide for machine learning. Retrieved from https://builtin.com/data-science/random-forest-algorithm
[19] GeeksforGeeks. (2025, October 24). XGBoost. Retrieved from https://www.geeksforgeeks.org/machine-learning/xgboost/
[20] GeeksforGeeks. (2025a, July 15). LightGBM (Light Gradient Boosting Machine). Retrieved from https://www.geeksforgeeks.org/machine-learning/lightgbm-light-gradient-boosting-machine/
[21] GeeksforGeeks. (2025a, April 28). How CatBoost algorithm works. Retrieved from https://www.geeksforgeeks.org/machine-learning/catboost-algorithms/
[22] Kolli, A. (2024, February 13). Understanding CatBoost: The Gradient Boosting Algorithm for Categorical Data. Retrieved from https://aravindkolli.medium.com/understanding-catboost-the-gradient-boosting-algorithm-for-categorical-data-73ddb200895d
[23] GeeksforGeeks. (2025c, October 24). Support Vector Machine (SVM) algorithm. Retrieved from https://www.geeksforgeeks.org/machine-learning/support-vector-machine-algorithm/
[24] GeeksforGeeks. (2025a, February 7). Major kernel functions in support Vector Machine (SVM). Retrieved from https://www.geeksforgeeks.org/machine-learning/major-kernel-functions-in-support-vector-machine-svm/
[25] GeeksforGeeks. (2025d, October 3). Bayes' theorem. Retrieved from https://www.geeksforgeeks.org/maths/bayes-theorem/
[26] GeeksforGeeks. (2025d, August 25). Naive Bayes classifiers. Retrieved from https://www.geeksforgeeks.org/machine-learning/naive-bayes-classifiers/
[27] GeeksforGeeks. (2025a, January 13). Regression in machine learning. Retrieved from https://www.geeksforgeeks.org/machine-learning/regression-in-machine-learning/
[28] GeeksforGeeks. (2025g, October 9). Harmonic Mean. Retrieved from https://www.geeksforgeeks.org/maths/harmonic-mean/
[29] GeeksforGeeks. (2025e, July 23). F1 score in machine learning. Retrieved from https://www.geeksforgeeks.org/machine-learning/f1-score-in-machine-learning/
[30] Pykes, K. (2025, April 3). Cohen's Kappa explained. Retrieved from https://builtin.com/data-science/cohens-kappa
[31] Bobbitt, Z. (2022, May 10). Cohen's Kappa Statistic: Definition & Example. Retrieved from https://www.statology.org/cohens-kappa-statistic/
[32] GeeksforGeeks. (2025d, July 11). Rsquared in regression analysis in machine learning. Retrieved from https://www.geeksforgeeks.org/machine-learning/ml-r-squared-in-regression-analysis/
[33] GeeksforGeeks. (2025d, May 27). How to calculate mean absolute error in Python? Retrieved from https://www.geeksforgeeks.org/python/how-to-calculate-mean-absolute-error-in-python/
[34] C3.ai. (2024, June 11). Root Mean square Error (RMSE). Retrieved from https://c3.ai/glossary/data-science/root-mean-square-error-rmse/
[35] Naïve model | IBF. (n.d.). Retrieved from https://ibf.org/knowledge/glossary/naive-model-197
[36] Eland, M. (2022, May 20). Understanding regression metrics. Retrieved from https://accessibleai.dev/post/regression_metrics/

Additional Links

[37] Meleshko, M. (2025, August 12). Importance of Machine Learning in agriculture: main applications. Retrieved from https://indatalabs.com/blog/ml-in-agriculture
[38] All figures have been made using https://canva.com

Journal of Information Systems Engineering and Management
2025, 10(63s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Datasets

1. Niteshhalai. (2020, December 30). Crop Recommendation Dataset. Retrieved from https://www.kaggle.com/code/niteshhalai/crop-recommendation-dataset
2. Alemu, S. (2024). Crop Recommendation using Soil Properties and Weather Prediction Dataset. *Mendeley Data*. https://doi.org/10.17632/8v757rr4st.1
3. Agricultural crop yield in Indian States dataset. (2023, July 17). Retrieved from https://www.kaggle.com/datasets/akshatgupta7/crop-yield-in-indian-states-dataset