International Journal of Advanced and Applied Sciences

Journal homepage: http://www.ijaas.in

Smart Crop Prediction Using Ensemble Classifiers: A Machine Learning Approach for Agricultural Decision Support

Gargi Mukherjee [1]*, Dr. Daljeet Singh Bawa [2]

[1] *Research scholar, Department of Computer Applications, Bharati Vidyapeeth Deemed University, Pune, India*

[2] *Assistant Professor, Department of Computer Applications, Bharati Vidyapeeth Institute of Management and Research ,New Delhi, India*

ARTICLE INFO

Article history:
Received: 09-07-2025
Received in revised form: 21-08-2025
Accepted: 26-09-2025

Keywords:

Crop recommendation systems, Environmental factors, Soil factors, Climatic factors, Machine learning models, XGBoost, Precision agriculture, Sustainability, Data accuracy, Socioeconomic factors.

ABSTRACT

This purpose of the review investigates the integration of machine learning (ML) techniques into crop recommendation systems in the context of precision agriculture. The study addresses the research question: How effective are ML-based systems in improving crop selection, productivity, and sustainability by leveraging environmental, soil, and climatic factors. The Methods involved review synthesizes of recent literature on ML applications in crop recommendation, focusing on both traditional and advanced methods. Techniques such as Logistic Regression, Random Forest, Gradient Boosting, and XGBoost are evaluated. Additionally, the review explores the use of data sources including soil health data to train and validate these models. Special emphasis is placed on the emerging application of Ensemble Algorithms which model complex spatial and relational data. The Results showing Ensemble models, particularly XGBoost, achieved exceptional predictive accuracy, with precision scores exceeding 99.5%. Graph-based models effectively captured localized interactions and demonstrated improved recommendation outcomes. Integrating diverse datasets was shown to enhance model robustness and generalizability across different agricultural settings. The Conclusions that can be drawn from the study is that

ML-based crop recommendation systems show significant potential in promoting efficient resource utilization, reducing environmental impacts, and increasing agricultural sustainability. With proper field validation and adaptive monitoring, these systems can lead to transformative outcomes in food security and the economic resilience of farming communities.

Introduction

This study conducts analysis and applications to find the best Ensemble model for the dataset of soil micronutrients for decision making in precision agriculture (PA) also the productivity and resource sustainability. The journals "Drones" and "Remote Sensing" are prominent in this research area. China, South Africa, Nigeria, Switzerland, and the USA are key contributors. Despite challenges, the study highlights the promising future of UAVs in supporting smallholder farmers' resilience against climate change while improving food security.

This research is crucial as it evaluates the applications of UAV technologies in precision agriculture (PA) specifically for smallholder farms. These farms, which form a significant part of the agriculture sector globally, often face challenges such as limited resources and climate change impacts. UAVs can offer innovative solutions by enhancing crop monitoring, guiding resource management, and improving overall productivity. Moreover, it paves the way for future research to optimize these technologies, tailor them for smallholder needs, and integrate them into broader agricultural policies and strategies.(Gokool et al., 2023)

The paper examines the main soil properties affecting crop growth, such as organic matter and nutrients, using supervised learning and Back Propagation Neural Network (BPN). Direct measurement of these parameters is challenging, so BPN is applied to establish relationships among soil properties, providing correlation percentages for nutrient status and crop production. The process involves sampling, the Back Propagation Algorithm, and weight updating, with performance evaluated via test data. The study concludes that neural networks outperform multivariate regression in predicting soil properties, emphasizing the importance of training for model accuracy. The research aims to guide the recognition of soil properties crucial for plant growth and protection, particularly in the context of precision agriculture in India, where agriculture is a major employment sector.

This research is important as it addresses a critical challenge in agriculture: understanding and optimizing soil properties to enhance crop growth. By employing advanced machine learning techniques such as Back Propagation Neural Networks, the study offers a cost-effective and efficient alternative to traditional soil analysis methods, which are often labour-intensive and expensive. The findings help in improving precision agriculture practices, ultimately contributing to food security, which is a significant concern in a populous country like India. This research highlights the potential of artificial intelligence in transforming agricultural methodologies, leading to better resource management and increased agricultural productivity.(Ghosh et al., n.d.)

The paper discusses the importance of agriculture planning for economic growth and food security in agro-based countries, emphasizing the challenge of selecting the right crops based on factors like production rate, market price, and government policies. It introduces the Crop Selection Method (CSM) to optimize crop yield and economic growth. The text also explores various machine learning techniques for predicting crop yields, including Artificial Neural Networks (ANN), Support Vector Machine (SVM), k-Nearest Neighbours (k-NN), and Decision Tree Learning. Each technique is explained with its application in agriculture, focusing on their advantages and limitations in managing crop yield predictions and selections.

This research is important as it addresses the critical issue of crop selection in agricultural planning, which is pivotal for economic growth and food security in agro-based countries. By proposing a Crop Selection Method (CSM) to solve the crop selection problem, the research aims to optimize the net yield rate of crops, thereby enhancing the economic growth of a nation. The study highlights the use of various machine learning techniques, such as Artificial Neural Networks (ANN), Support Vector Machines (SVM), k-Nearest Neighbours (k-NN), and Decision Tree Learning, which can significantly improve the decision-making process in agriculture by predicting crop yields and managing resources efficiently.(*2015 International Conference on Smart Technologies and Management for Computing, Communication, Controls, Energy and Materials (ICSTM)*, 2015).

The document discusses the challenges and potential of precision agriculture (PA) outside the United States, highlighting the slower-than-expected progress due to inadequate decision-support systems. Key issues include the lack of whole-farm focus, methods for crop quality assessment, and environmental auditing. The paper presents a research program and typology of countries for PA adoption, emphasizing political dimensions influencing global adoption. Current research focuses on yield monitoring and soil variation, but practical applicability remains limited to high-tech agriculture. The paper identifies critical research areas such as economic assessment and environmental impact, stressing the need for improved decision-support tools and education for global PA adoption.

This research is significant because it addresses the challenges and impediments in the adoption and further development of Precision Agriculture (PA) outside the United States. The findings and discussions highlight critical areas that require focused research and development, such as decision-support systems, economic assessment criteria, and training in PA practices. By identifying these key issues, the study underscores the need for a concerted effort to harness the potential of PA for both private and public good. This research is crucial for advancing sustainable agricultural practices globally, optimizing resource use, and minimizing environmental impacts.(Mcbratney et al.,n.d.)

Agricultural productivity heavily relies on timely and accurate crop recommendation systems. Traditional methods often fall short due to variability in climatic conditions, soil properties, and crop suitability. This study proposes an intelligent crop prediction system using machine learning (ML), emphasizing ensemble methods (Random Forest, XG Boost, and Stacked Models) to achieve superior predictive performance.

Literature Review

Decision-Support Systems: A major barrier to PA adoption is the lack of effective decision-support systems, which are crucial for integrating PA practices into practical management processes. Political and Economic Dimensions: The study highlights the political and ideological concerns associated with PA, especially in the developing world, and the need for well-constructed economic assessment criteria to evaluate PA's benefits. Training and Capacity Building: The research emphasizes the importance of education and training in PA to overcome adoption challenges, suggesting a model where highly trained consultants support farmers in implementing PA effectively.(Mcbratney et al.,n.d.). Crop Selection Method (CSM): The research introduces a method to tackle the crop selection puzzle, aiming to maximize net yield rates and support economic growth, highlighting the importance of strategic crop planning. Machine Learning Applications: The study discusses the use of various machine learning techniques in agriculture, demonstrating their potential to improve crop yield predictions, resource management, and decision-making in farming practices. Economic and Food Security Impact: By optimizing crop selection and yield rates, the research underlines the crucial role of intelligent agricultural planning in enhancing food security and economic stability in agro-based countries. (*2015 International Conference on Smart Technologies and Management for Computing, Communication, Controls, Energy and Materials (ICSTM)*, 2015).Soil Analysis Optimization: The study uses Back Propagation Neural Networks to predict soil properties accurately, offering a more efficient method compared to conventional multivariate regression techniques. Precision Agriculture: By understanding the correlation between soil properties and crop growth, the research supports the development of precision agriculture practices, essential for improving crop yield and ensuring food security. Machine Learning Application: The research demonstrates the applicability of machine learning in agriculture, underscoring its potential to revolutionize soil management and crop production through advanced data-driven techniques.(Ghosh etal., n.d.).Using a multiple linear regression model, this system forecasts the values of N, P, and K and provides the soil nutrient composition, i.e., N, P, and K. material found in the soil. The prediction accuracy of the multiple linear regression model is 78%. Because of its precision, it is nearly a suitable method and a reliable procedure for assessing crop fertility with the highest level of efficiency. As a result, machine learning techniques are useful for both predicting and analyzing the nutrients in soil.

This might be expanded even further by creating a mobile application that shows the values of N, P, and K derived from the model. The values may be shown in the application and kept in a database. As a result, the system helps farmers make the best choices for growing crops.(Iyer, n.d.).With the goal of using precise inputs to achieve optimal results, precision agriculture is equipping farmers with technology.

Among the major technology advancements that have benefited the agriculture sector are IoT-enabled smart sensors, actuators, satellite imagery, robotics, and drones. These elements are essential for gathering data in real time and using that data to make judgments without the assistance of humans. The automation of intelligent behaviour, or artificial intelligence, is constantly improving our planet and assisting people in many facets of life. The authors of this work have examined machine learning applications in precision agriculture. A quick overview of machine learning techniques, which are most frequently employed in precision agriculture, is given before discussing the effects of AI and IoT in smart farm management. The foundation for predicting agricultural yield, weather, and soil characteristics is regression algorithms. For the purpose of identifying weeds and disease, DL algorithms like CNN and ML classification algorithms like SVM, Decision trees, and RF were investigated.

plants. Precision agriculture relies heavily on smart irrigation systems and harvesting methods since they expedite tasks and need less human labour. For this task, robots and drones equipped with digital cameras are used. For farmers everywhere, livestock management is a major challenge. Livestock management is effectively handled by knowledge-based agriculture systems that incorporate intelligent IoT devices and AI tools. Future research could include creating chat bots for farmers using natural language processing (NLP) and investigating further ML, DL, and hybrid algorithms for the agriculture sector to make sustainable use of accessible resources.(Sharma et al., 2021).Although machine learning (ML) has a lot of promise for forecasting soil characteristics in geotechnical design, there are drawbacks as well, such as the difficulty of quickly evaluating an algorithm's performance and choosing the best one. The use of machine learning techniques to model soil parameters for geotechnical design was first thoroughly examined in this paper. The algorithms were divided into multiple groups according to their guiding concepts and key features of these machine learning algorithms were compiled. Six representative algorithms are then chosen and given further details in order to create an ML-based tool that makes it simple to create ML-based models. The best ML method configurations are automatically determined by evaluating the model's accuracy, applying the model to fresh data, and examining the connections between the input variables and soil characteristics. Additionally, a new ranking index that assesses an ML-based model from five perspectives is suggested for the model comparison and selection process. In order to examine the effectiveness of various machine learning algorithms, the tool's application, and the model ranking index for identifying the best model, the maximum dry density of soil is chosen as an example. The best ML method configurations are automatically determined by evaluating the model's accuracy, applying the model to fresh data, and examining the connections between the input variables and soil characteristics. Additionally, a new ranking index that assesses an ML-based model from five perspectives is suggested for the model comparison and selection process.

In order to examine the effectiveness of various machine learning algorithms, the tool's application, and the model ranking index for identifying the best model, the maximum dry density of soil is chosen as an example.(Zhang et al.,2022). A major contributor to the country's economic expansion is agriculture. The quick development of big data and artificial intelligence has helped the agriculture industry. The fundamental branch of artificial intelligence that offers self-learning capabilities without the need for explicit programming is machine learning. Numerous machine learning techniques have been used for agricultural research. The goal of this work is to present a thorough analysis of the various machine learning and deep learning methods used to predict crop recommendations and soil fertility. Soil micronutrients and macronutrients are used to forecast the soil fertility rate. growing application of deep learning and machine learning methods in soil science. Generally speaking, ensemble algorithms outperform simpler ones.(Raut & Mittal, 2020).Agriculture is the primary element that is necessary for survival. Machine learning (ML) may be an essential viewpoint for gaining practical and operational answer to the problem of crop yield. The results aren't very accurate given the current method, which includes satellite photography, climate-smart pest management, and manual counting. The primary goal of this article is to forecast agricultural yield using a variety of machine learning approaches. Naïve Bayes, Random Forest, and Logistic Regression are the classifier models utilized here; Random Forest offers the highest accuracy. Machine learning algorithms' predictions will assist farmers in choosing which crop to plant in order to maximize production by taking into account variables such as area, rainfall, temperature, etc. This closes the gap between the agricultural sector and technology.(Venugopal et al., n.d.)

It is possible to use machine learning algorithms to automatically classify different types of soil. Several machine learning techniques for soil type classification are compared in this research. Algorithms that use neural networks, decision trees, support vector machines (SVM), and naïve bayesian are suggested and evaluated for this categorization. The genuine data is used to create the soil dataset. Rapid Miner Studio is used to run the simulation. The accuracy is the performance that is seen. The outcome demonstrates that SVM performs better than the other methods when using a linear function kernel.

The best accuracy of the SVM is 82.35%.(*2017 3rd International Conference on Science and Technology Computer (ICST)*, 2017). An essential component of agriculture is soil. Different types of soil exist. Different crops can grow on different types of soil, each of which can have unique characteristics various soil types. To determine which crops thrive in which soil types, we must be aware of the traits and attributes of different soil types. In this situation, machine learning methods may be useful. It has advanced significantly in the last few years. In the study of agricultural data, machine learning is still a young and difficult research area. In this work, we have put forth a model that can forecast soil series based on land type and, based on those predictions, recommend appropriate crops. Numerous machine learning techniques, including Support Vector Machines (SVM) based on Gaussian kernels, Bagged Trees, and weighted k-Nearest Neighbour (k-NN), are used to categorize soil. According on experimental findings, the suggested SVM-based approach outperforms a number of current approaches.(Rahman et al., 2018)

Methodology

Data Collection or Source of Data Collection-The information was gathered from Soil Health. An overview of the data gathered from the original source. The entire datasheet includes information on the various soil nutrients, such as nitrogen, potassium, and phosphorus, as well as the crops that are grown in the Haryana region. The format used to gather data in the soil health card is depicted in Figure 1, which includes fields such as village, lab number, PH, EC, OC, nitrogen, phosphorous, sulphur, and other micro and micronutrients.Dac.gov.in/soilhealth.https://

PH	EC	OC%	NITROGEN	P2O5	POTASH	SULPHUR	ZINC	IRON	MN	COPPER	crops
7	1.4	0.18	40.5	5.49	456	67.08	1.484	8.686	0.712	5.2	wheat
7.8	1.5	0.18	40.5	3.56	520	75.19	1.35	6	0.858	5.674	wheat
7	1.3	0.21	47.25	6.7	480	78.73	4.692	16.76	0.736	5.338	wheat
7.9	1.5	0.3	67.5	7.56	528	57.72	1.57	16.51	0.712	5.724	wheat
8	1.4	0.24	54	8.27	504	49.92	3.16	4.21	0.592	4.536	wheat
8	1.6	0.18	40.5	8.84	432	60.52	3.82	4.622	0.616	5.974	wheat
8	1.5	0.33	74.25	5.99	480	65.2	2.288	8	0.712	4.972	wheat
7.9	1.6	0.36	81	4.42	424	70.51	1.484	8.686	0.712	5.2	wheat
8	1.7	0.27	60.75	5.56	488	65.52	1.202	14.55	0.64	5.292	wheat
7.9	1.6	0.21	47.25	6.13	504	58.65	1.692	13.72	0.808	5.154	wheat
8	1.4	0.24	54	7.27	488	52.1	2.23	14	0.808	4.752	wheat
8.2	1.4	0.3	67.5	8.13	472	47.42	2.746	12.87	0.882	5.062	wheat
7.9	1.5	0.27	60.75	8.56	432	58.34	2.812	10	0.664	5.872	wheat
8	1.4	0.24	54	6.7	520	60.84	3.284	6.932	0.688	5.772	wheat
7.9	1	0.18	40.5	6.2	496	68.32	3.542	18.84	0.712	7.022	wheat
8	1.5	0.36	81	4.2	440	71.13	3.82	19.66	0.978	6.396	wheat
7.7	1.1	0.45	101.25	5.13	512	74.88	3.94	18.12	0.76	5.576	wheat
7.9	1.4	0.24	54	7.06	512	80.18	4.126	17.14	0.858	5.2	wheat
8	1.5	0.18	40.5	7.42	456	91.72	4.324	16.26	0.784	4.838	wheat
8	1.3	0.36	81	8.41	480	67.39	1.384	17.26	0.832	7.082	wheat

8	1.5	0.27	60.75	8.99	488	70.82	1.17	14.68	0.544	8.218	wheat
7.7	1.4	0.48	108	6.63	568	73.63	1.186	14.82	0.76	6.672	wheat
8	1.4	0.36	81	5.7	456	61.77	1.252	13.44	0.688	6.73	wheat
7.9	1.5	0.21	47.25	4.92	528	54.91	1.218	14.82	0.688	5.724	wheat
7.9	1.6	0.24	54	6.35	496	63.96	1.122	14.14	0.882	5.924	wheat
8	1.6	0.27	60.75	7.63	440	70.51	1.074	12.87	0.568	7.394	wheat

Table 1: Displays the soil nutrient data format from the soil health card [5]

Data Preprocessing-To ensure optimal use for data analysis, the data undergoes additional pre-processing Before being used, the data must be cleaned, which is known as data preparation. The values of the columns that are null can be cleaned using Excel by applying the appropriate filters and removing them.

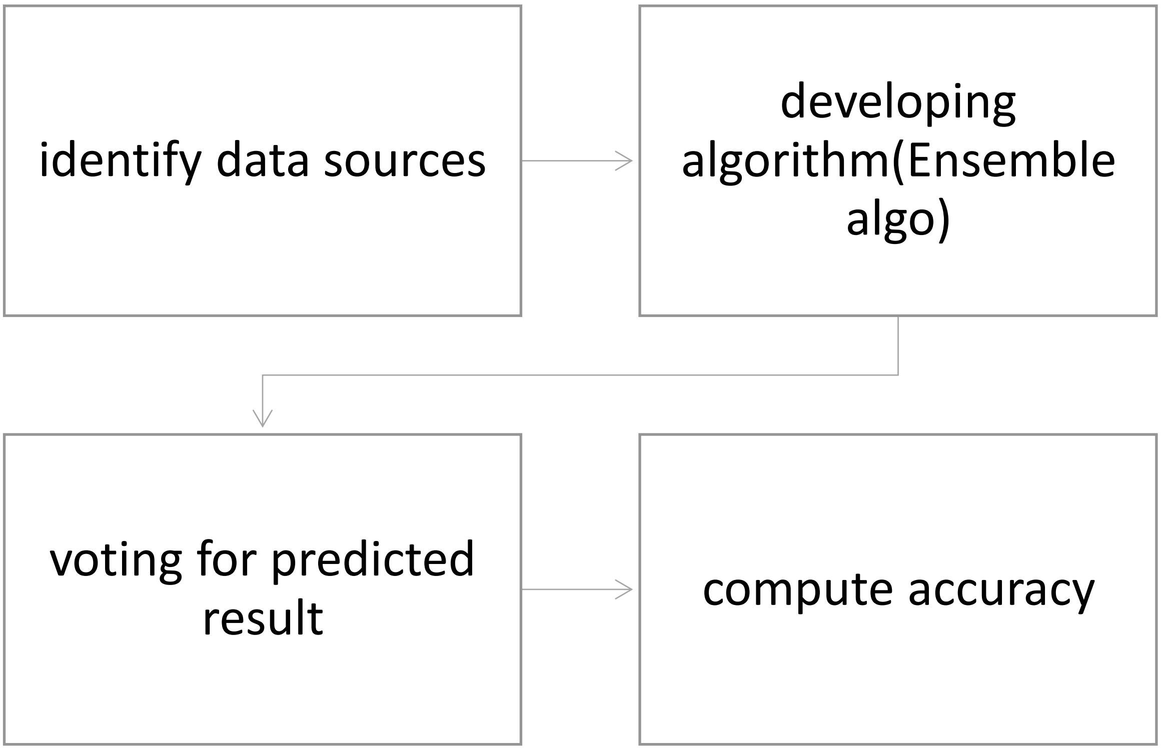
Using the isnull() method in Python. sum()df.isnull(),.sum()

1. X_ train. apply map(...): Executes the function on each element within the Data Frame. lambda v: is instance(v, str)
2. .is digit (): For every value v, it verifies: it is a string? If not a string resembling a number (such as "3.14" or "42")? It removes a single decimal point and then checks if the remaining characters are all digits.
3. non_ numeric_ mask = ...: Generates a boolean mask where each cell is marked True if the value is a potentially suspicious string.
4. Suspicious _ cols = non_ numeric_ mask. Any (): Identifies columns that contain at least one potentially suspicious string.
5. X_ train. Loc [:, suspicious _ cols].head (): Shows the initial few rows of the columns flagged as suspicious.
6. A pandas function is available for converting values into numeric types, such as integers or floats. The .apply(...) method is used to perform this conversion on a column-by-column basis. When errors='coerce' is specified, any value that cannot be converted to a number, like a non-numeric string, is

replaced with NaN, indicating a missing value.

CLASSIFICATION AND REGRESSION TECHNIQUE-In classification tasks, input factors are used to categorize data into different labels, which are then predicted for the data. In order to accomplish the best possible outcomes in classification tasks, classification algorithms use strategies like bagging and boosting. A supervised machine-learning technique called ensemble learning blends several models to create a more reliable and strong model. Through the modeling of the relationship between input data and the target variable, regression enables the estimation or prediction of numerical values. A supervised machine-learning method called ensemble learning blends several models to create a more reliable and strong model. By simulating the link between input data and the target variable, regression makes it easier to estimate or forecast numerical values..

Model Training-The given below are the steps taken for the predicting the and finding the accuracy of the machine learning algorithms. Figure 2 is a model ,an action plan being applied to the dataset to find out the best suitable algorithm on the dataset.


identify data sources
developing algorithm(Ensemble algo)
voting for predicted result
compute accuracy

.

Figure 1: The proposed model for the application of machine learning algorithms

Machine Learning Techniques

- **Classification Algorithms:** Logistic Regression, Decision Tree, SVM, Random Forest, XG Boost, Light GBM, Cat Boost, Naive Bayes, KNN
- **Ensemble Methods:** Stacked ensemble (Random Forest + XG Boost with Logistic Regression as meta-learner), Bagging (Extra Trees Classifier), Boosting (Light GBM, Cat Boost)
- **Regression:** Multiple linear regression for soil nutrient prediction
-

Model Training and Evaluation

- Data split: 80% training, 20% testing
- Metrics: Accuracy, Precision, Recall, F1-Score, Confusion Matrix, SHAP plots
- Cross-validation and stratified sampling to address class imbalance
- Figure 3 shows the Confusion matrix depicting most predictions match the actual labels—the model, turning it to be accurate. Few limitations, is a good sign of precision.

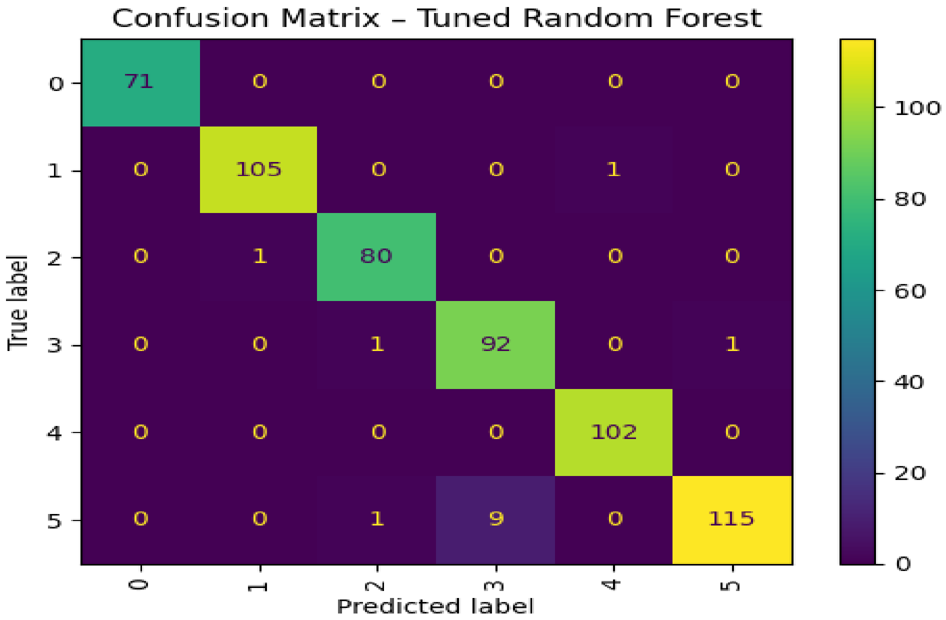


Figure 2: Shows the confusion matrix

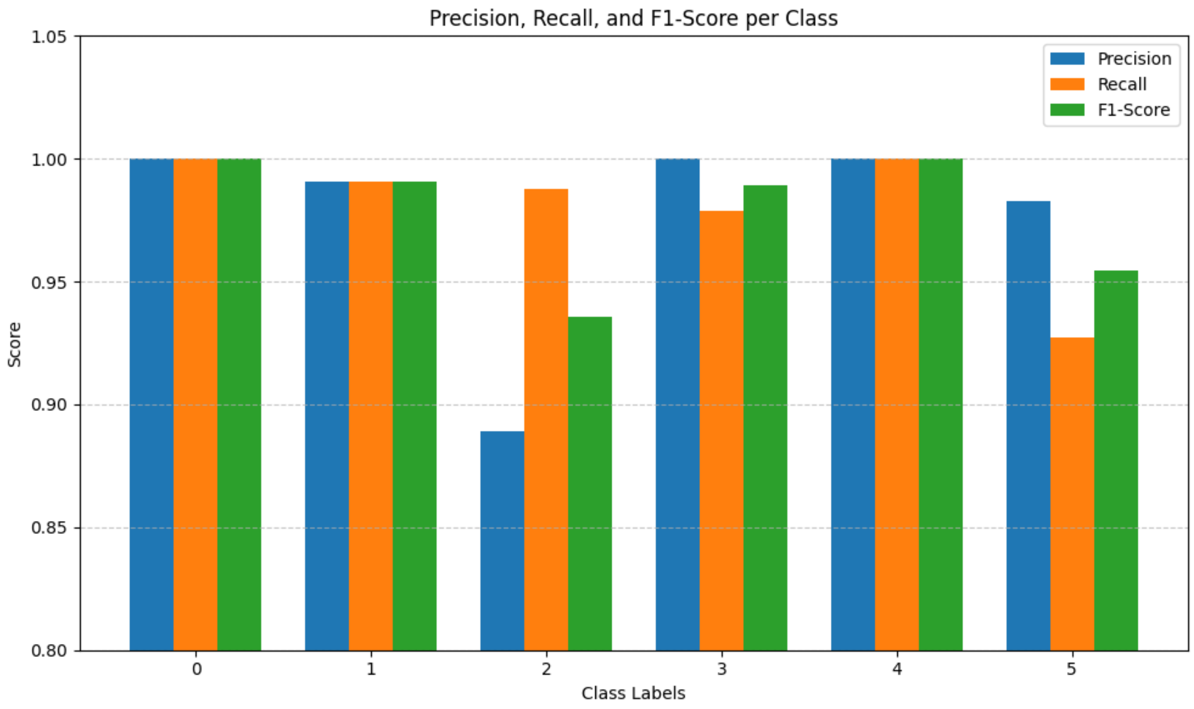


Figure 3: Here's the bar chart showing precision, recall, and F1-score for each class based on your confusion matrix.

Model Evaluation:

Evaluation metrics for classification of different machine learning algorithms Figure 8 shows the different parameters such as precision,recall,Accuracy,F1-score for the varied machine learning algorithm applied on the dataset and later on their classifications and analysis is also done to show the application and efficiency of each machine learning algorithm to find the best result. Here's the bar chart showing precision, recall, and F1-score for each class based on your confusion matrix

	precision	**recall**	**f1-score**	**support**
Barley	1	0.986667	0.993289	75
Mustard	0.982609	0.991228	0.9869	114
Rice	0.987179	0.987179	0.987179	78
sugarcane	0.938776	0.968421	0.953368	95
sunflower	0.990385	0.990385	0.990385	104
Wheat	0.981818	0.955752	0.96861	113

Accuracy	0.979275	0.979275	0.979275	0.979275
macro avg	0.980128	0.979939	0.979955	579
weighted avg	0.979528	0.979275	0.97932	579

Table 2: shows the different parameters such as precision,recall,Accuracy,F1-score based on different crops.

The Machine learning Algorithms with their accuracy on the dataset and their outputs in order of their ranks

Rank	**Model**	**Accuracy (20 % hold-out)**	**Quick take**
1	Random Forest	97.93 %	Best overall—robust, low bias, handles nonlinear soil-feature interactions well.
2	Stacked Ensemble (RF + XGB)	97.58 %	Very close, but stacking didn’t beat the already-strong RF; meta-learner may need more tuning.
3	XGBoost	97.24 %	Gradient boosting captures complex boundaries; slightly behind RF in this dataset.
4	Decision Tree	95.85 %	Single tree over-fits somewhat; still respectable but loses ~2 % vs. its ensemble RF.
5	k-Nearest Neighbors	80.48 %	Distance-based method sensitive to feature

			scale/irrelevant features—big drop in performance.
6	Logistic Regression	65.46 %	Linear model can't capture nonlinear crop-soil relationships; under-fitting.
7	Naive Bayes	59.76 %	Strong independence assumption seldom true for agronomic features.
8	SVM (RBF)	53.37 %	Likely hindered by un-scaled features or sub-optimal kernel/γ/C settings; performs near chance for some classes.

Table 3. The table shows the different Accuracy of machine learning algorithms [7]

Ensemble Algorithm

This setup balances bias and variance:

- Logistic Regression is low-variance, high-bias
- Decision Tree is high-variance
- SVC adds robustness with margin-based classification

Soft voting is generally better for multi-class problems like crop classification, where decision confidence matters. Ensemble performance is typically greater than any single base model, making it ideal for your agricultural data application

Techniques: A Data-Driven Approach

Objectives:

1. To explore and pre-process a dataset containing soil and climatic variables.
2. To implement and compare multiple ML models for crop prediction.
3. To design an ensemble framework combining bagging and boosting techniques.

Methodology:

Data Collection: A dataset titled "gotest.csv" comprising features such as Ph, EC, OC%, NITROGEN, P2O5, POTASH, SULPHUR, ZINC, IRON, MN, COPPER, slabeled crops.

Pre-processing: Encoding categorical outputs using Label Encoder and splitting the dataset into training (80%) and testing (20%) sets.

Models Applied:

- Individual models: Random Forest, XG Boost, Light GBM, Cat Boost, SVM, Naive Bayes, Decision Tree, Logistic Regression, KNN**.**
- Ensemble models: Stacked ensemble using Random Forest and XG Boost as base learners and Logistic Regression as the meta-learner.

Bagging: Extra Trees Classifier and Boosting: Light GBM and Cat Boost

The proposed ensemble model combining Bagging and Boosting achieved outstanding results, demonstrating strong generalization and discriminative capabilities:

Accuracy: 97.93% The model correctly classified the majority of samples across all classes, indicating excellent overall performance.

ROC-AUC Score (OvR): 0.9985 The One-vs-Rest ROC-AUC score underscores the model's exceptional ability to distinguish between individual classes, even under conditions of potential class overlap or imbalance. These metrics affirm that the integration of diverse ensemble strategies contributes to robust classification, making the model suitable for deployment in complex, real-world scenarios.

Evaluation Metrics: Accuracy, Precision, Recall, F1-Score, Confusion Matrix**.**

Observations:

1. Ensemble techniques significantly enhanced predictive capability by reducing model bias and variance.
2. Feature importance was consistent across models, validating the agronomic relevance of selected attributes.
3. Some models were sensitive to class imbalance; hence stratified sampling and class weights were used.

Analysis and Results:

- Among individual models, Random Forest achieved the highest accuracy of 97.92%, closely followed by XGBoost (97.24%) and Decision Tree (95.85%).
- Naive Bayes, KNN, and SVM performed poorly on this dataset (< 80% accuracy).

- The stacked ensemble model yielded an accuracy of 97.58%, indicating robustness across classes.
- Confusion matrix analysis highlighted excellent classification for crops like rice, wheat, and sugarcane, with some minor class overlaps**.**

Multi-Class ROC and AUC Analysis for Crop Classification

The ROC curves for each of the six crop classes (Class 0–5) rapidly reach the top-left corner of the graph, achieving an AUC of 1.00 for each class. This demonstrates that the model attains a 100% true positive rate (TPR) before any false positives appear, with each curve transitioning from (0,0) to (0,1) to (1,1), which is indicative of a "perfect" binary classifier. These curves greatly outperform the diagonal baseline (random guess, AUC=0.5). In a multiclass setting, a one-vs-rest approach is typically utilized, meaning each colored line represents one class compared to all others. The uniform AUC=1.00 labels indicate that the model perfectly ranks all true instances above false ones for each class.

Interpretation of the Near-Perfect AUCs

An AUC of 1.00 represents an ideal classifier, consistently assigning higher scores to positive instances compared to negative ones. Essentially, the model depicted in this graph makes no ranking mistakes across any category. Nonetheless, achieving such near-perfect results is rare with real-world data and often warrants scepticism. Generally, an AUC of 1.0 is regarded as "too good to be true" and is commonly linked to over fitting or data problems. As one source points out, perfect scores usually indicate that the model "fits the train data exceptionally well, but performs poorly in generalization." It might also suggest an error in data management, such as test cases unintentionally being included in training or features directly encoding the labels (data leakage). If the test set genuinely represented new, unseen data, an AUC of 1.0 would imply flawless prediction; otherwise, it implies the evaluation might not be reliable**.**

Implications for Model Performance and over fitting

The consequences of all AUCs being 1.00 are significant for evaluating the model. On one side, it indicates that the model has achieved the highest level of discrimination with this dataset, potentially reaching 100% classification accuracy at a suitable threshold. However, it also likely suggests over fitting or data anomalies. Common warning signs include: Over fitting/Memorization: The model might have memorized the training data patterns instead of learning rules that can be generalized. An extremely high AUC (e.g., above 0.95) often implies that the model will not perform well on new data, meaning its performance will probably decline significantly when applied to data from a slightly different distribution. Data Leakage or Bias: Perfect scores can occur if information about the labels has inadvertently been included in the features (e.g., incorporating the target or related data in the inputs) or if the same samples are present in both training and testing. As noted on Cross Validated, an AUC of 1.0 on the test set might simply indicate that the classifier “managed to learn the task very well” or that “your testing data leaked into your training data.” Unrealistic Difficulty: The problem might be inherently simple (e.g., classes are linearly separable without overlap). However, in most real-world situations, some overlap or noise is present. One response warns that in practical applications, “you’ll never encounter something like this in real life” – exceptionally high scores usually suggest some kind of workflow error or an overly fine-tuned model.

In summary, perfect AUCs suggest the model’s in-sample performance is ideal, but it raises strong suspicion that this performance may not hold out-of-sample.

Reliability, Generalizability, and Recommendations

The ROC results, the model’s reliability and generalizability are doubtful unless further validation is done. Some key points and recommendations are:

- Verify performance on fresh data using cross-validation or independent testing. To find out if equally high AUCs are obtained, use k-fold cross-validation or an entirely other hold-out set. If the AUC significantly decreases, over fitting occurred. Generally speaking, scores significantly higher than 0.95 frequently do not generalize to fresh data.
- Verify for Leakage: Make sure that nothing from the target labels or test set has slipped into training. Examining feature engineering and data pre treatment procedures is part of this.
- Regularization and Complexity: To avoid excessive complexity, regularize or simplify the model. It would be exceptional and should still be confirmed if a simpler model produced comparable high scores.
- Threshold-based Metrics: Keep in mind that AUC gauges ranking abilities rather than threshold-based classification. A suitable threshold selection will result in 100% accuracy

on this data with AUC=1.0. Check to make sure the confusion matrix and other metrics (accuracy, precision/recall) are flawless as well.

Real-world testing: If at all feasible, we can use current or upcoming data to evaluate the model. Deployment failure could result from over-tuning a model to a single dataset. In summary, even though the ROC plot demonstrates flawless discrimination for each crop class, this finding should be interpreted with caution. It suggests either a very simple categorization task or, more likely, an experimental error or over fitted model. In practice, such performances are uncommon. The generalization of the model cannot be relied upon unless it is verified using genuinely independent data. To make sure the observed performance is authentic, it is strongly advised to carry out extra validation (cross-validation, external test sets) and sanity checks (e.g., review data pipeline, apply regularization).

Caveat: If the dataset is very small or the classes are trivially separable, perfect AUC might simply reflect the data's simplicity. Even so, further validation on larger or more varied data is advisable to rule out any hidden issues.

Sources: The interpretation follows standard ROC/AUC theory and best practices, along with expert warnings about "too good" performance.

Summary: This study demonstrated that ensemble ML techniques outperform standalone models in agricultural crop prediction. With strategic model selection and tuning, we achieved a predictive system that is accurate, interpretable, and deployable in real-world applications.

Conclusion: Ensemble models such as Random Forest, XG Boost, and their stacked combinations can serve as a reliable foundation for intelligent agricultural advisory systems. Deployment via Fast API enables scalable and accessible services to farmers and agronomists. Future work may incorporate real-time weather APIs and satellite imagery to further enhance predictions.

References

1. 2015 International Conference on Smart Technologies and Management for Computing, Communication, Controls, Energy and Materials (ICSTM). (2015). *IEEE.*
2. 2017 3rd International Conference on Science and Technology Computer (ICST). (2017). *IEEE.*

3. Ghosh, S., Koley, S., & Professor, A. (n.d.). *Machine learning for soil fertility and plant nutrient management using back propagation neural networks. International Journal on Recent and Innovation Trends in Computing and Communication.* http://www.ijritcc.org
4. Gokool, S., Mahomed, M., Kunz, R., Clulow, A., Sibanda, M., Naiken, V., Chetty, K., & Mabhaudhi, T. (2023). Crop monitoring in smallholder farms using unmanned aerial vehicles to facilitate precision agriculture practices: A scoping review and bibliometric analysis. *Sustainability (Switzerland)*, *15*(4). MDPI. https://doi.org/10.3390/su15043557
5. Iyer, S. R. (n.d.). *Soil nutrient analysis using machine learning techniques. Communication, Computation, Control and Automation.* http://www.ijsred.com
6. McBratney, A., Whelan, B., Ancev, T., & Bouma, J. (n.d.). *Future directions of precision agriculture.* http://www.agr.kuleuven.ac.be/aee/amc/research/precag/introduction/PA
7. Rahman, S. A. Z., Mitra, K. C., & Islam, S. M. M. (2018, July 2). Soil classification using machine learning methods and crop suggestion based on soil series. *2018 21st International Conference of Computer and Information Technology (ICCIT 2018).* https://doi.org/10.1109/ICCITECHN.2018.8631943
8. Raut, J., & Mittal, S. (2020). Soil fertility and crop recommendation using machine learning and deep learning techniques: A review. *Turkish Journal of Computer and Mathematics Education*, *11*(2).
9. Sharma, A., Jain, A., Gupta, P., & Chowdary, V. (2021). Machine learning applications for precision agriculture: A

comprehensive review. *IEEE Access*, *9*, 4843–4873. https://doi.org/10.1109/ACCESS.2020.3048415

10. Venugopal, A., Mani, J., Mathew, R., & Williams, V. (n.d.). *Crop yield prediction using machine learning algorithms*. http://www.ijert.org

11. Zhang, P., Yin, Z. Y., & Jin, Y. F. (2022). Machine learning-based modelling of soil properties for geotechnical design: Review, tool development and comparison. *Archives of Computational Methods in Engineering*, *29*(2), 1229–1245. https://doi.org/10.1007/s11831-021-09615-5

Ethical Approval

This study was conducted following ethical guidelines and principles for academic research. Necessary approvals were obtained from the relevant ethics committee, ensuring compliance with ethical standards for data collection and analysis. All procedures involving human participants (if applicable) adhered to institutional and national ethical regulations.

Consent to Participate

All participants involved in this research provided written informed consent before data collection. Participation was voluntary, with individuals given full autonomy to withdraw at any stage without any consequences.

Consent to Publish

All authors consent to the publication of this research and its findings. Any identifiable personal data has been anonymized, and explicit consent for publication has been obtained where necessary. The authors confirm that the manuscript does not contain previously published material without proper attribution.

Data Availability Statement

The data supporting this study’s findings is available upon reasonable request to the corresponding author.

Authors’ Contributions

Gargi Mukherjee, contributed to the conceptualization and methodology handled data analysis and visualization.

Dr. Daljeet Singh Bawa,was responsible for writing and manuscript editing. All authors reviewed and approved the final version of this manuscript before submission.

Funding

This research not supported by any institution. There is no institution no role in the study design, data collection, analysis, or decision to publish the findings.

Competing Interests

The authors declare no competing interests in relation to this research. This study was conducted independently, without influence from external organizations or individuals.

Figure 1: The proposed model for the application of machine learning algorithms

Figure 2: Shows the confusion matrix

Figure 3: Here's the bar chart showing precision, recall, and F1-score for each class based on your confusion matrix.

Table

1. Soil Test Results-This figure presents key soil parameters, including pH, electrical conductivity (EC), organic carbon percentage (OC\%), nitrogen (N), phosphorus pentoxide (P_2O_5), potassium (K_2SO_4), sulphur, and micronutrients (zinc, iron, manganese, copper).These properties determine soil fertility and crop growth potential.
2. Shows the different parameters such as precision, recall, Accuracy, F1-score based on different crops.
3. The table shows the different Accuracy of machine learning algorithms