engineering proceedings

Proceeding Paper

A Machine Learning-Enabled System for Crop Recommendation †

Pedina Sasi Kiran, Gembali Abhinaya, Smaraneeka Sruti and Neelamadhab Padhy *

School of Engineering and Technology, Department of Computer Science and Engineering, GIET University, Gunupur 765022, Odisha, India; 22cseaiml003.pedinasasikiran@giet.edu (P.S.K.); 22cseaiml044.gembaliabhinaya@giet.edu (G.A.); 22cseaiml058.smaraneeka@giet.edu (S.S.)

* Correspondence: dr.neelamadhab@giet.edu

† Presented at the 3rd International Electronic Conference on Processes—Green and Sustainable Process Engineering and Process Systems Engineering (ECP 2024), 29–31 May 2024; Available online: https://sciforum.net/event/ECP2024.

Abstract: Context: We are advancing our efforts in agriculture by creating a crop prediction system with the help of machine learning. Our goal is to build an ML model that can estimate the properties of a crop. It will push ahead in agriculture by developing a predictive tool for crops using machine learning in agriculture in terms of both time and money. Our farmers can understand easily and analyze what best they are going to farm. **Objective:** The main theme of this project is to support farmers in yielding a good crop by making a robust model. By identifying the significant role of technology in advanced farming practices, we aim to create a solution that helps farmers make informed decisions about crop selections and agricultural practices. Utilizing data analytics and AI-driven insights enhances productivity and efficiency. Our final goal is to encourage farmers with the tools and knowledge they need to grow in an increasingly complex agricultural landscape. **Methods**: To complete this model, we collected data from different sources like the data of weather, humidity, pH value, temperature, nitrogen, phosphorous, and potassium values, and rainfall in mm. We implemented it through ML algorithms like GNB (Gaussian Naïve Bayes), SVM (Support Vector Machine), RF (Random Forest), and DT (Decision Tree). **Result:** The GNB classifier achieves an accuracy of 99%, surpassing others.

Keywords: crop yield prediction; machine learning; agriculture technology; precision farming; decision support system

Citation: Kiran, P.S.; Abhinaya, G.; Sruti, S.; Padhy, N. A Machine Learning-Enabled System for Crop Recommendation. *Eng. Proc.* **2024**, *67*, 51. https://doi.org/10.3390/engproc2024067051

Academic Editor: Wen-Jer Chang

Published: 24 September 2024

1. Introduction

Agriculture has existed for several centuries. In agriculture, information technology plays a crucial role which is used to help farmers to work smarter and to produce more crops. Most of the population is dependent on agriculture for the system. By recognizing the historical data and identifying patterns and trends in parameters like nitrogen, phosphorus, potassium, pH value, temperature, humidity, and rainfall, farmers are helped to grow crops and make smarter choices.

In a study by Van Klompenburg et al. [1], they used machine learning algorithms like NN, LR, RF, and SVM featuring soil information (54), humidity levels (38), nutrients (28), and other factors (24) and field management practices. To assess the performance of these models, they used metrics such as RMSE (Root Mean Square Error), R^2, MAE (Mean Absolute Error), MSE (Mean Squared Error), MAPE, and RSAE. Their objective was to identify key features and algorithms in crop yield prediction and provide insights for future research. In a study by Nischitha, K. et al. [2], the authors' objective was to recommend specific fertilizers and seed quantities, to increase crop yield, and reduce soil pollution caused by improper fertilizer use. The dataset used contains features such as pH, humidity, rainfall, and NPK values. They used SVM and DT algorithms. They used sensors to obtain input data. Their prediction is to be used to select a suitable crop and also provide data about required nutrients, to calculate expected yield, and market price. The objective of the paper by Joshua, S. et al. [3] is to improve the accuracy of crop yield predictions using machine learning. The authors compared traditional multi-linear regression with newer models like BPNN, SVM, and GRNN. Using data from different states' paddy fields, rainfall, and nutrient levels (nitrogen, phosphate, potash), they found that GRNN provided the best results. GRNN achieved a 97% accuracy with an R^2 of 0.97 and NMSE of 0.03. In the research paper by Palanivel, K. et al. [4], the authors conducted an investigation into how various machine learning algorithms work for crop prediction. They considered parameters such as soil type, pH, nitrogen, depth, temperature, and rainfall. The prediction is performed using climate variables namely PRE, TMX, TMN, PET, SM, and land cultivated for maize. The datasets used contained data from 1990 to 2017, of which 80% were used for training and 20% were used for testing. The algorithms used for this model are ANN, SVM, and MLR. The authors used MSE, RMSE, and MAE to obtain the performance matrix of the model. In the paper by Rao, M. S., et al. [5], a feed-forward back-propagation Artificial Neural Network method was used to forecast various crop yields in rural areas. The algorithm used in this paper is an RF Classifier using two different criteria, Gini and Entropy, and gives the highest accuracy with 99.32%. The dataset used in the paper consists of 22 different crops with seven features and uses accuracy, precision, recall, and F1-score to find the performance matrix. In the paper by Singh, V. et al. [6], the parameters present in the dataset are macro-nutrients (pH, OC, EC, N, P, K, S) and micro-nutrients (Zn, Fe, Mn, Cu) from samples collected from different regions of Jammu District. They used many algorithms to complete this model and used different performance matrices. In the paper by Elbasi, E., et al. [7], the authors' objective was to show how machine learning can optimize crop production and reduce waste, showing 99.59% accuracy with Bayes theorem and other classifiers. They used performance matrices such as MSE, MAE, RMSE, and RAE. Elbasi, E. [8] discussed how machine learning classifiers are working in the agriculture field. They used the classification approaches and obtained 99.59% in the Bayes Net algorithm. Similarly, they also used the NB classifier and obtained 99.46%. Champaneri, M. [9] designed one web-based application that predicts climate change in India. They used NB, and RF ML classifiers to predict the types of crops. They also discussed how data mining helps find the hidden patterns in the sensor's data. GS Sajja [10] investigated how machine learning classifiers are useful for crop recommendation and benefits for the farmer. To achieve their goals they used ID3, SVM, and RF classifiers. Senapaty, M. K. [11] collected the data from several sources and prepared the dataset. Further, they developed one web-based application as well as an Android application. They considered several machine learning classifiers and got 100% accuracy in SGDC.

They discussed different classifiers to analyze the crops. The authors considered data from four districts and maintained them on a seasonal basis. A total of 246,091 samples were collected. They also used different sensors to collect real-time data to achieve the model's accuracy [12]. The authors studied the different types of soils and their characteristics which are helpful for crop recommendations. They conducted a PICO study to analyze the soil properties. Further, they developed a recommendation system for soil color analysis. How the data will be collected from a real-time environment to data storage in the cloud is also discussed [13]. The authors discussed and exhibited several machine-learning classifiers for crop recommendation. They developed a DSS (Decision Support System) for soil nutrient analysis and crop recommendation.

Aim of This Study

This study aims to develop an effective crop recommendation system. That system helps us to identify the optimized crop. It also allows for analysis of the soil and environmental parameters such as N, P, K, pH levels, temperature, humidity, etc. Our optimum aim is to make an effective decision system.

We mention the parameters utilized to verify this hypothesis in the "Proposed Model for Crop Recommendation" section. These parameters include soil composition, crop type, weather conditions, and machine learning model parameters, which align with our aim to recommend crops effectively. To properly enhance the crop prediction system and timely decision-making system in precision farming, we test the hypothesis that the combination of machine learning techniques like Random Forest, SVM, and Decision Trees and soil parameters, N, P, K, calcium, and phosphorus levels, weather conditions, crop type, and historical yield data considerably enhance the accuracy of a crop recommendation system. To address effective recommendations, we tested two hypotheses. These are H0 (Null Hypothesis) and H1 (Alternative Hypothesis).

Null Hypothesis (H0): *The machine learning-enabled crop recommendation system does not significantly increase crop selection accuracy when compared to traditional approaches based on standard soil and environmental characteristics (e.g., nitrogen, phosphorus, potassium, pH, temperature, humidity).*

Alternative Hypothesis (H1): *The machine learning-enabled crop recommendation system outperforms previous approaches in crop selection accuracy by optimizing characteristics such as soil nutrients (nitrogen, phosphorus, potassium), pH levels, and ambient variables (temperature and humidity).*

The above two hypotheses are based on our machine learning classifier performance along with the soil parameters. We have verified the hypothesis through different classifiers like GNV, SVM, LoGR, DT, and RF. We have estimated their performance metrics like precision, recall, F1-score, and accuracy across different datasets.

2. Methodology for Crop Recommendation

To make predictions in this model, we used machine learning and the Python programming language. Figure 1 below is the proposed workflow of our crop recommendation system.

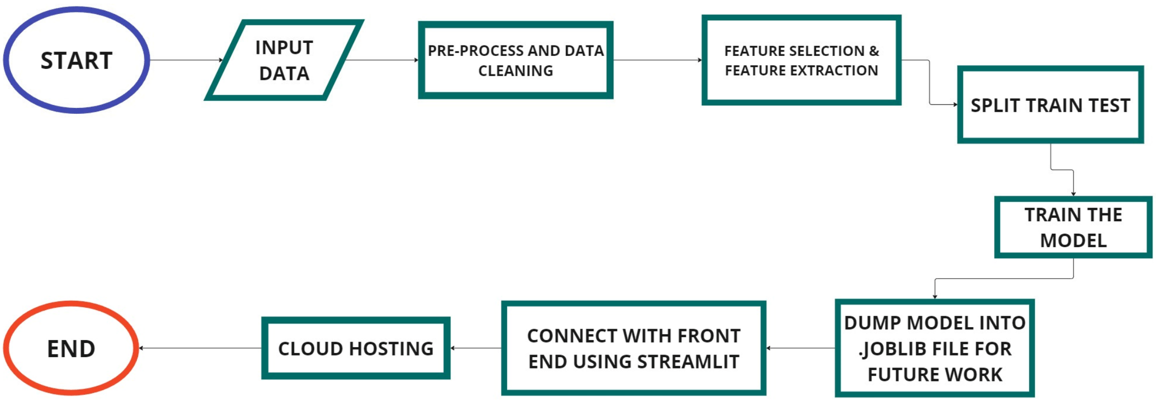

Figure 1. Proposed methodology for recommendation.

1. Start: This marks the beginning of the process;
2. Input data: Gather and provide the data that will be used for the analysis;
3. Pre-processing and data cleaning:
    - Handling missing values: Check if there are any gaps or empty blanks in the data and fill them in with appropriate values or remove them;
    - Removing duplicates: Find and remove repeated entries to confirm all data are unique;
4. Feature selection and feature extraction: Identify the most important features of the data that will help in making predictions and create new features from the existing data;
5. Split the data into training and testing datasets: Divide the data into two parts, one for training the model (training data) and the other for testing the model (testing data);

6. Train the model: Use the training data to teach the model to make predictions;
7. Dump the model into Joblib file for future work: Save the trained model into a file format called Joblib so that it can be used again without training it;
8. Connect with front end using Streamlit: Integrate the model with a user-friendly interface created using Streamlit, allowing users to interact with the model easily;
9. Cloud hosting: Move the application to a cloud service, making it easily accessible via the internet;
10. End: This marks the completion of the process.

3. Proposed Model for Crop Recommendation

This model (Figure 2) contains four different phases to go through, namely dataset creation, data pre-processing, training the model, and evaluation of the result.

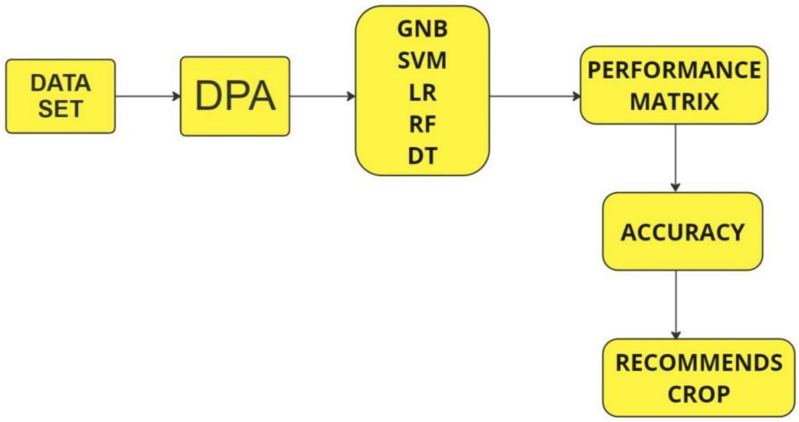

Figure 2. Proposed model for crop recommendation.

- **PHASE 1**: We collected various parameters like weather, humidity, pH value, temperature, NPK values, and rainfall in mm. Data were collected from different sources on the internet like Kaggle;
- **PHASE 2**: The following processes were performed: **Collected Data:** Gathered soil data and weather data, humidity, pH → value, → temperature, NPK values, and rainfall in mm. **Cleaning:** Removed any errors or noises in the data, and converted categorical data into numerical form. **Normalization:** Scaled the data to ensure all features contribute equally;
- **PHASE 3**: In our model, we used ML classification algorithms like RF (Random Forest), DT (Decision Tree), KNN (K-Nearest Neighbor), GNB (Gaussian Naïve Bayes), and LR. These are examples of an ensemble learning algorithm that generates different types of DT, Accuracy, and Precision. **Gaussian Naive Bayes (GNB)** uses statistics to make predictions based on the probability of different features (like weather conditions). **Logistic Regression (LR)** finds a way to draw a line that separates classes based on input features (like temperature and rainfall). **SVM** finds a boundary that separates classes with the largest margin between different groups. **RF** builds many decision trees and combines their results to make a more consistent prediction. **DT** uses a tree-type model of decisions to divide the data into branches based on yes/no questions until it decides simple terms. **GNB** uses probabilities based on feature averages. **LR** draws a line to separate classes. **SVM** finds the best boundary to separate classes. **RF** combines multiple decision trees for better accuracy. **DT** uses a flowchart of yes/no questions to decide. These algorithms help in making predictions and decisions based on data patterns, each using a different approach to find the best solution. We used GNB as it provided us with the best accuracy, i.e., 99% accuracy;
- **PHASE 4**: In a crop prediction system, a performance matrix is used to evaluate how well the system predicts outcomes like crop yield, health, or suitability based on various metrics. The accuracy and consistency of the prediction model are improved.

The key metrics in the performance matrix are: Accuracy, which measures how often the model's predictions are correct; Precision, which scales the percentage of true positive predictions out of all positive predictions made; Recall, which measures the percentage of true positive of the correct prediction, F1-Score, which is used to find the balance between precision and recall; MAE, which measures the difference between predicted and actual values; Root Mean Squared Error (RMSE), which calculates the square root of the average square difference between predicted and actual values; Confusion Matrix, which calculates the True Positives, True Negatives, False Positives, and False Negatives.

In summary, performance metrics in a crop prediction system provide an effective way to evaluate how accurately the system predicts outcomes, which is essential for improving model effectiveness and making reliable predictions. We have considered and worked only on the accuracy part of our model.

- **PHASE 5:** We have considered and worked on the accuracy part of the performance metrics and also performed a comparison of different types of algorithms. The formula for accuracy is:

$$\text{Accuracy} = \text{Correct Predictions} \backslash \text{Total Predictions}$$

- **PHASE 6:** The model was validated using performance inspection parameters on the testing dataset, guaranteeing that it did not overfit the data gathered through sensory methods.

4. Results and Discussions

This section outlines the specific predictions made by the ML-based crop prediction system regarding the parameters provided based on the analysis. We used different algorithms to find the accuracy.

The accuracy of the Gaussian Naive Bayes model is 0.9, i.e., (99%). The accuracy of the SVM is approximately 0.9795, i.e., (97%). The accuracy of the LR model is approximately 0.9523, i.e., (95%). The accuracy of the RF model is approximately 0.9809, i.e., (98%). The accuracy of the DT model is 0.9, i.e., (90%). So, here we have chosen the best one which is GNB. Sample outputs are shown below in Table 1.

Table 1. Results of different classifiers.

Classifiers	Accuracy	Precision	Recall	F1-Score
GNB	0.99	0.99	0.99	0.99
SVM	0.97	0.97	0.97	0.97
LR	0.96	0.96	0.96	0.96
DT	0.99	0.99	0.99	0.99
RF	0.99	0.99	0.99	0.99

- **Research Question 1:** Is it possible to compare the different machine learning algorithms along with their performance metrics for crop recommendation?
- **Response to research question 1**: Yes, it is possible to compare the different ML algorithms along with their performance metrics. First, we compared the accuracy metrics of different classifiers and it is observed that the classifiers GNB (Gaussian Naive Bayes), DT (Decision tree), and RF (Random Forest) obtained 0.99. The below-mentioned Figure 3 is the accuracy comparisons of the different classifiers

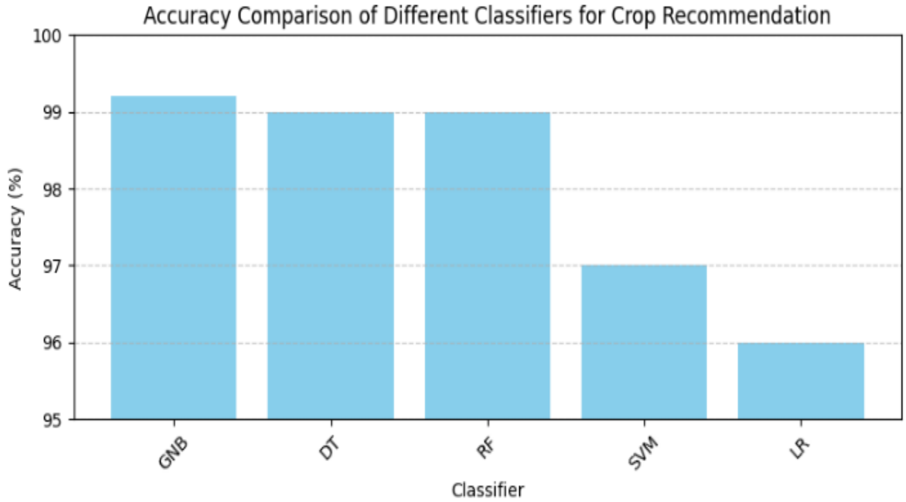

Figure 3. Accuracy comparison of different algorithms.

Figure 4 demonstrates the visualization and comparison of all the metrics for crop recommendation. This graph helps us to identify the classifiers that are working constantly to improve the performance. Table 2 above is the comparison of our proposed system to the state-of-the-art in terms of its objective as well as the performance metrics. It has been observed that our proposed system performs well.

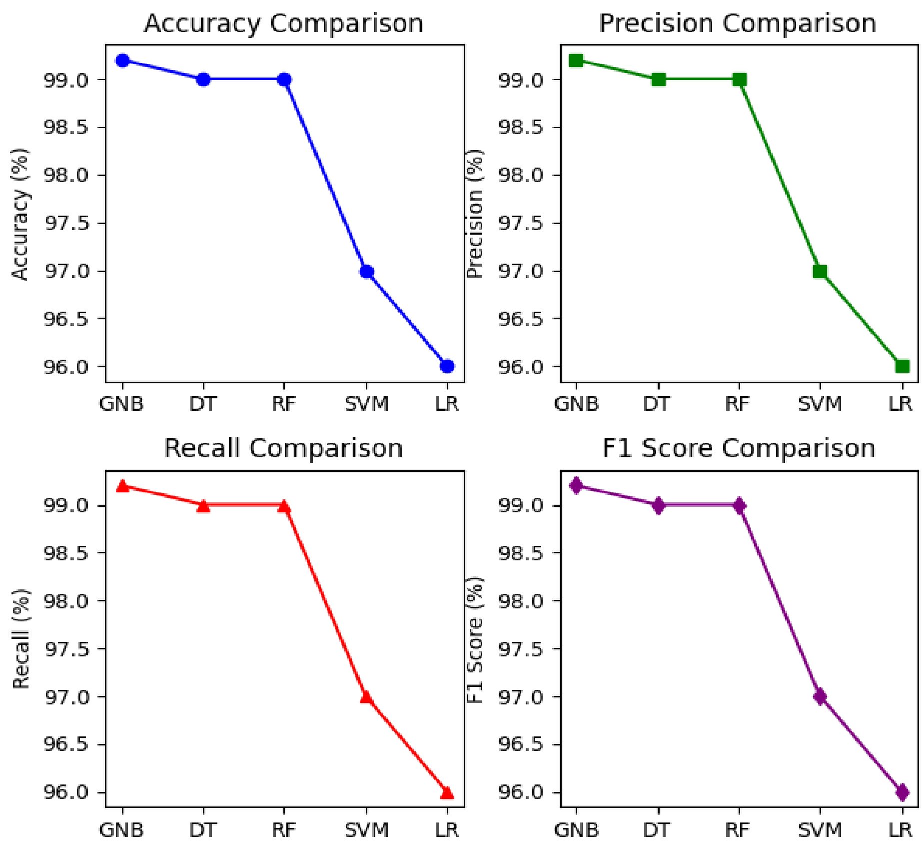

Figure 4. Brief comparison of metrics for crop recommendation.

Table 2. Comparison of the proposed model to the existing one.

Ref#	Objective	Algorithm Used	Proposed System
[14]	They used the Kaggle dataset and their objective was to create a recommendation system	Cat-boost, GNB, and SMOTE feature selection for optimizing the crops.	Our model is better in comparison to this research work as our proposed system's accuracy, precision, and recall are 99%.
[15]	The objective of this paper was crop management using an ensemble learning approach.	They used SVM, KNN, DT, and NB. They obtained precision, recall, and F1-score of 98% as the highest measurement. When they performed the ensemble approach then they achieved 99%. They used IoT sensors to collect the data.	Our approach is better than this approach. Our traditional approach as an ensemble approach obtained 99% accuracy, precision, and recall.
[16]	Their objective was to recommend and optimize the crop using machine-learning classifiers.	They used LR, DT, KNN, NB, SVM, etc. Their performance is consistently 95% across all the models.	Our model is better than this work because our model consistently achieved over 99% across all the models.

- **Research Question 2:** Can we show that the machine learning algorithms are stable and consistent in their performance metrics when recommending the crop?
- **The solution to the research question**: Yes, we can measure their stable and consistent performance using a boxplot representation.

Figure 5 is a boxplot representation of the different classifiers. It demonstrates how the metrics(accuracy, precision, recall, F1-score) work as well as stability and consistency to achieve crop recommendation. It visualizes the distribution and variability of the metrics. With the help of this plot, we can tell the stability and consistent performance of the classifiers for the crop recommendation system.

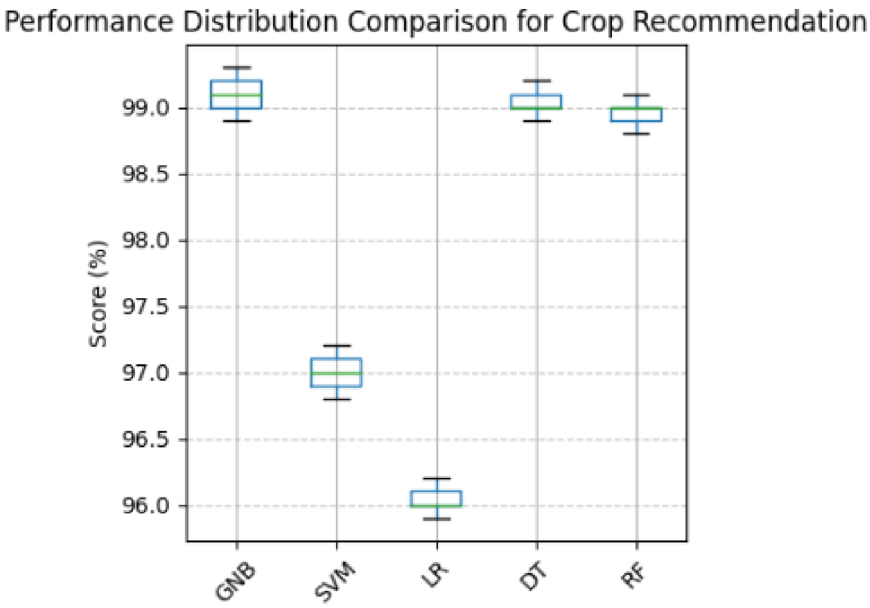

Figure 5. Performance distribution comparisons for crop recommendation.

Figure 6 depicts the crop recommendation system accessed through Streamlit. We used Python code and generated it as a pickle file. That pickle file is used for web hosting using Streamlit. Now our implementation is hosted in the form of a GUI (Graphical User Interface) which is useful for the farmers.

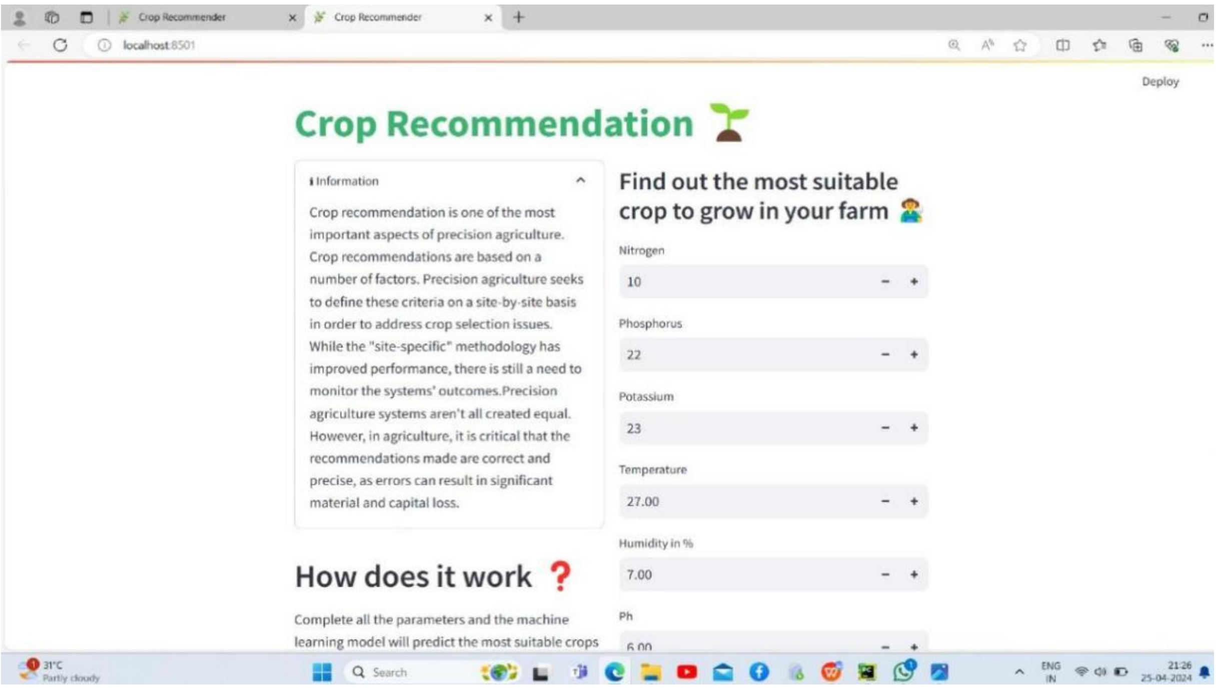

Figure 6. Crop recommendation model.

Figure 7 is called a hypothetical learning curve. In our results, all the metrics obtained are much higher for all the classifiers using the training set. We considered the classifier in the x-axis and their corresponding score is on the y-axis. On the x-axis, we show the size of the training samples of the different classifiers. When the training set size is increased then the metrics which have obtained a high value remain stable or there are slight deviations. As all the metrics provide a high accuracy, the model is a robust model that can handle volumes of data.

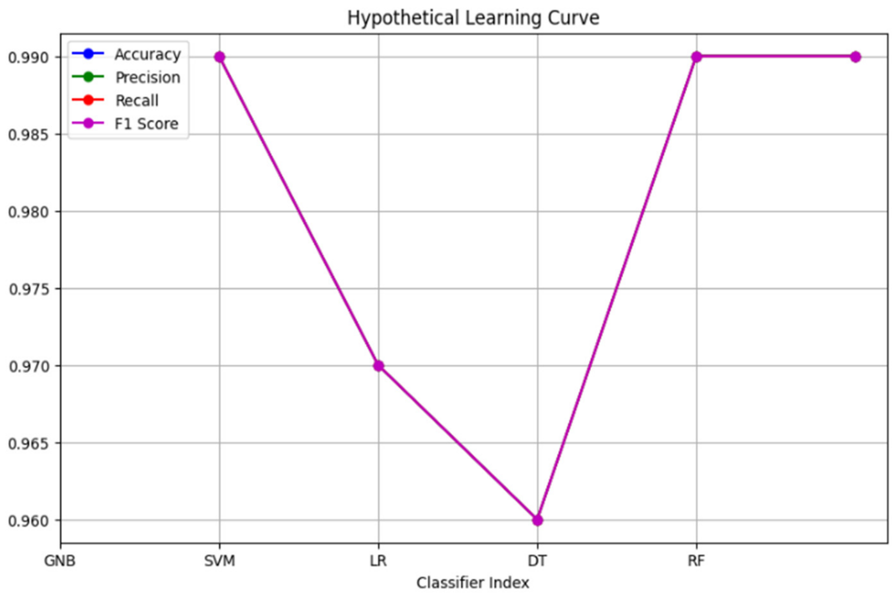

Figure 7. Hypothetical Learning curve.

The graph in Figure 8 is called an AUC-ROC curve which is used for estimating the binary class classification models. The x-axis represents the FP and the y-axis represents the TP. In the graph below, each line represents the performance of the classifier which is called the decision threshold. The curve demonstrates the relation between sensitivity and specificity.

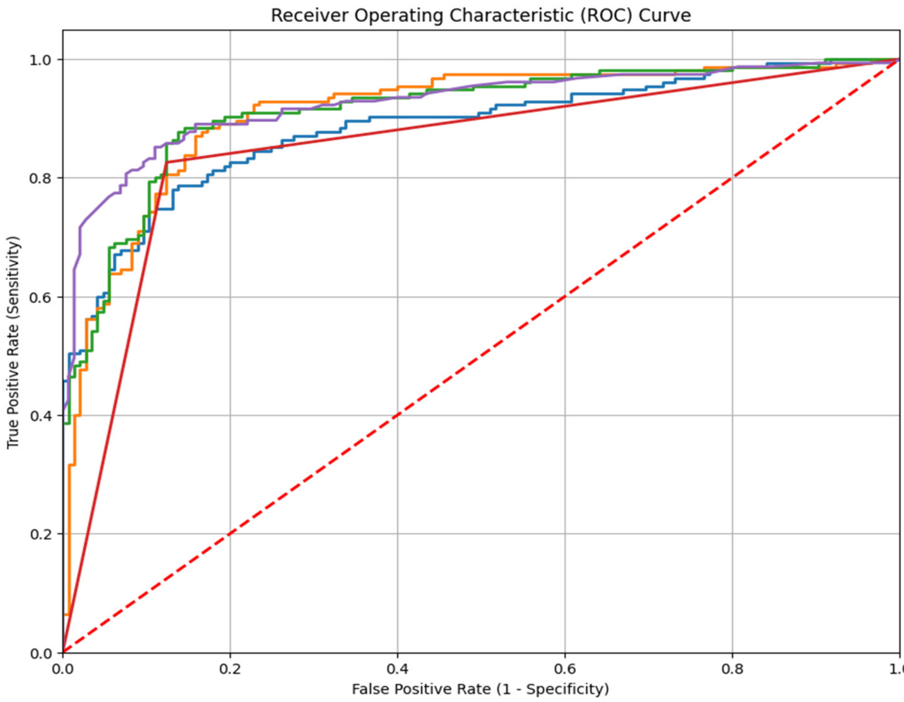

Figure 8. ROC curves for different classifiers.

We have expanded the discussion on calcium and phosphorus and their role in mitigating phytotoxicity, linking these findings to our objective of developing a machine learning-enabled decision support system for crop recommendation.

The ROC curve illustrates the trade-off between the True Positive Rate (Sensitivity) and the False Positive Rate (1 - Specificity), with each line denoting a distinct model. A model with high sensitivity and few false positives is shown by curves that approach the top-left corner. The dashed red line shows a baseline model with arbitrary predictions. When a model behaves in this manner, it has no discriminative power and is equivalent to guesswork.

The colours probably correspond to various machine learning models that you have used to solve categorisation problems. Better performance is typically indicated by curves that remain closer to the top-left corner of the graph. One way to compare these models is via the area under each curve (AUC).

Our proposed system can predict crop recommendations using machine learning classifiers and identify the optimal crops considering key parameters like N, P, K, pH levels, etc. As per our experimental observation, we found a 20% improvement in crop yield compared to other traditional approaches.

We have observed that the previous studies focused only on manual or less dynamic approaches to nutrient management. The traditional approaches may not be suitable for identifying the multiple nutrients dynamically that contribute to phytotoxicity. Our proposed system improves this substantially and predicts the optimal levels of calcium and phosphorus needed to mitigate phytotoxicity. Our proposed system utilizes machine learning classifiers that dynamically analyze multiple soil and environmental parameters. Moreover, the classifiers GNB, DT, and RF perform well in terms of precision and recall, i.e., 99%. It is stated that our proposed system can handle dynamically different soil parameters and optimize crop types that are less likely to suffer from nutrient imbalances or phytotoxic conditions.

This not only increases crop output but also marks a breakthrough in reducing the dangers of calcium and phosphorus toxicity—a problem that earlier research has found difficult to successfully address.

5. Conclusions

In conclusion, agriculture plays a vital role in our lives, with half of the nation's economy and half of its population depending on farming. Recognizing the challenges faced by farmers, such as limited crop knowledge, we have developed a solution. Our crop prediction system is a steady tool that helps farmers choose the best crops for maximum yield without difficulty.

In this model, we have employed various ML classifiers and evaluated their performance metrics, including Accuracy, Precision, F1-score, and Recall. The GNB (Gaussian Naïve Bayes) classifier demonstrates superior accuracy, achieving a remarkable 99%.

Author Contributions: Conceptualization, P.S.K. and N.P.; methodology, G.A.; software, S.S.; validation, P.S.K. and S.S.; formal analysis, N.P.; investigation, N.P.; resources, S.S.; data curation, P.S.K.; writing—original draft preparation, P.S.K.; writing—review and editing, S.S.; visualization, G.A.; supervision, N.P.; project administration, N.P.; funding acquisition, S.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The dataset used in this study is publicly available and can be accessed from the Kaggle platform.

Acknowledgments: We would like to thank Rasmita Panigrahi and the Department of computer science and engineering, GIET University for their valuable recommendations, professional guidance, expert advice, and encouragement while preparing this paper.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Van Klompenburg, T.; Kassahun, A.; Catal, C. Crop yield prediction using machine learning: A systematic literature review. *Comput. Electron. Agric.* **2020**, *177*, 105709. [CrossRef]
2. Nischitha, K.; Vishwakarma, D.; Ashwini, M.N.; Manjuraju, M.R. Crop prediction using machine learning approaches. *Int. J. Eng. Res. Technol. (IJERT)* **2020**, *9*, 23–26.
3. Kalimuthu, M.; Vaishnavi, P.; Kishore, M. Crop prediction using machine learning. In Proceedings of the 2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT), Tirunelveli, India, 20–22 August 2020; IEEE: Piscataway, NJ, USA, 2020; pp. 926–932.
4. Joshua, S.V.; Priyadharson, A.S.M.; Kannadasan, R.; Khan, A.A.; Lawanont, W.; Khan, F.A.; Ali, M.J. Crop yield prediction using machine learning approaches on a wide spectrum. *Comput. Mater. Contin.* **2022**, *72*, 5663–5679.
5. Palanivel, K.; Surianarayanan, C. An approach for prediction of crop yield using machine learning and big data techniques. *Int. J. Comput. Eng. Technol.* **2019**, *10*, 110–118. [CrossRef]
6. Rao, M.S.; Singh, A.; Reddy, N.S.; Acharya, D.U. Crop prediction using machine learning. In *Journal of Physics: Conference Series*; IOP Publishing: Bristol, UK, 2022; Volume 2161, p. 012033.
7. Singh, V.; Sarwar, A.; Sharma, V. Analysis of soil and prediction of crop yield (Rice) using Machine Learning approach. *Int. J. Adv. Res. Comput. Sci.* **2017**, *8*, 1254.
8. Elbasi, E.; Zaki, C.; Topcu, A.E.; Abdelbaki, W.; Zreikat, A.I.; Cina, E.; Saker, L. Crop prediction model using machine learning algorithms. *Appl. Sci.* **2023**, *13*, 9288. [CrossRef]
9. Champaneri, M.; Chachpara, D.; Chandvidkar, C.; Rathod, M. Crop yield prediction using machine learning. *Int. J. Sci. Res. (IJSR)* **2020**, *9*, 645–948.
10. Sajja, G.S.; Jha, S.S.; Mhamdi, H.; Naved, M.; Ray, S.; Phasinam, K. An investigation on crop yield prediction using machine learning. In Proceedings of the 2021 Third International Conference on Inventive Research in Computing Applications (ICIRCA), Coimbatore, India, 2–4 September 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 916–921.
11. Senapaty, M.K.; Ray, A.; Padhy, N. A decision support system for crop recommendation using machine learning classification algorithms. *Agriculture* **2024**, *14*, 1256. [CrossRef]

12. Senapaty, M.K.; Ray, A.; Padhy, N. State-of-the-art of soil mineral data extraction and crop recommendation using learning tools. *AIP Conf. Proc.* **2024**, *2919*, 050015.
13. Senapaty, M.K.; Ray, A.; Padhy, N. IoT-enabled soil nutrient analysis and crop recommendation model for precision agriculture. *Computers* **2023**, *12*, 61. [CrossRef]
14. Apat, S.K.; Mishra, J.; Raju, K.S.; Padhy, N. An artificial intelligence-based crop recommendation system using machine learning. *J. Sci. Ind. Res. (JSIR)* **2023**, *82*, 558–567.
15. Hossain, S.; Prova, N.N.I.; Sadik, M.R.; Al Maruf, A. Enhancing Crop Management: Ensemble Machine Learning for Real-Time Crop Recommendation System from Sensor Data. In Proceedings of the 2024 International Conference on Smart Systems for Applications in Electrical Sciences (ICSSES), Tumakuru, India, 3–4 June 2024; IEEE: Piscataway, NJ, USA, 2024; pp. 1–6.
16. Dahiphale, D.; Shinde, P.; Patil, K.; Dahiphale, V. Smart farming: Crop recommendation using machine learning with challenges and future ideas. *TechRxiv* **2023**. [CrossRef]