

Machine Learning for Crop Recommendation and Yield Forecasting: A Practical Framework for Sustainable Agriculture

Payal A. Thete¹, Prof. S. V. Jaybhay², Prof. K. D. Jamkar³, Dr. H. B. Jadhav⁴

Student¹, Guide², Co-Guide³, HOD⁴,

Department of Computer Engineering¹⁻⁴

Adsul's Technical Campus, Chas, Ahilyanagar, India

Abstract: Agriculture is increasingly influenced by changing weather, soil conditions, water availability, and input costs. Farmers often have to make crop-selection and production decisions with incomplete or delayed information. This paper presents a practical machine-learning framework that connects basic soil and weather observations with crop recommendation and yield forecasting. The proposed workflow includes data preparation, feature selection, supervised learning, model validation, and explainability. For crop recommendation, we discuss classification methods such as Random Forest, Extra Trees, Logistic Regression, Decision Tree, K-Nearest Neighbour, Gaussian Naive Bayes, and Support Vector Machine.

In contrast, regression methods such as Decision Tree Regressor, Random Forest Regressor, KNN Regressor, Support Vector Regressor, and XGBoost Regressor are considered for yield estimation. The supplied reference study reported very high benchmark performance on its datasets, including 99.7% classification accuracy for Random Forest and an R^2 value of 99.99% for SVR. We treat these numbers as reported benchmark results rather than new experimental results. The main contribution of this rewritten paper is a clearer, application-oriented framework that explains how machine learning can support agricultural decisions while recognizing data quality, geographic transferability, interpretability, and real-world deployment limitations

Keywords: Crop Recommendation, Yield Forecasting, Machine Learning, Precision Agriculture, Explainable AI, Sustainable Farming

I. INTRODUCTION

Agricultural decisions are closely tied to local conditions. The same crop may perform differently when rainfall, temperature, soil nutrients, or soil acidity changes. Because of this, a fixed rule such as planting the same crop every season cannot always provide a dependable decision. Data-driven systems offer another way: instead of relying only on experience, they can combine several measurable factors and learn patterns from previous observations.

Machine learning is useful in this setting because agricultural datasets contain relationships that are difficult to describe with a single rule. A classification model can map soil and environmental conditions to a suitable crop category, while a regression model can estimate a numerical yield value. The supplied research paper demonstrates both tasks using crop-recommendation and wheat-yield datasets, with preprocessing, feature engineering, cross-validation, and explainability methods.

This paper does not reproduce the supplied study. Instead, it reorganizes the core technical idea into an application-focused framework. The emphasis is on how a student or developer could structure such a system, what information should be collected, how models should be evaluated, and what limitations should be reported before a model is used for practical farming decisions.



II. RELATED WORK

Earlier agricultural machine-learning research has explored several directions, including crop selection, yield estimation, disease identification, pest detection, irrigation support, and remote-sensing-based prediction. The supplied paper summarizes studies using linear regression, Random Forest, Support Vector Regression, neural networks, CNNs, and LSTM models for different agricultural prediction tasks.

A common lesson across this area is that no single algorithm is guaranteed to work equally well for every crop, location, or dataset. Tree-based models can capture non-linear interactions, distance-based methods can be useful when similar observations have similar outcomes, and kernel-based regression can model complex relationships. Model comparison is therefore more meaningful than selecting an algorithm only because it is popular.

Another important theme is interpretability. In farming, a recommendation is more useful when the user can understand why it was generated. The supplied study uses feature importance and LIME to explain model outputs. This paper retains that idea but frames explainability as a practical communication layer between the prediction system and the farmer.

III. PROPOSED FRAMEWORK

The proposed framework follows a simple data-to-decision pipeline. It begins with collecting soil and environmental measurements, continues with cleaning and transforming the data, and then applies classification or regression models according to the required output. The final stage presents the result together with an explanation of the important input factors.

Proposed Data-to-Decision Pipeline



Fig. 1. Proposed data-to-decision pipeline for agricultural machine learning.

A. DATA INPUTS

For crop recommendation, useful inputs can include soil pH, nitrogen, phosphorus, potassium, temperature, humidity, and rainfall. The supplied crop dataset contains these variables and 2,200 observations associated with 22 crop labels. For yield forecasting, additional time-related and production variables may be needed. The supplied wheat dataset uses annual production-related information together with rainfall, pesticide usage, and temperature.

In a real deployment, the input layer should also record the location, measurement date, unit of measurement, and source of each value. This helps prevent a common problem in agricultural analytics: a model may appear accurate because the training data is clean and consistent, while field measurements may be incomplete or collected using different procedures.

B. DATA PREPROCESSING

Preprocessing converts raw observations into a form that a model can use reliably. Typical operations include removing duplicate records, handling missing values, checking unrealistic measurements, scaling numerical variables when required, and separating input features from the target label. For a classification task, the target is the recommended crop. For a regression task, the target is the numerical yield.

The source study used feature engineering and MinMax scaling for the crop recommendation task and MICE-based imputation for missing values in the yield-prediction workflow. In a new implementation, the preprocessing rules should be fitted only on training data and then applied to validation or test data. This avoids leaking information from the test set into model training.



C. MODEL SELECTION

For crop recommendation, the framework can compare Logistic Regression, Decision Tree, Random Forest, Extra Trees, KNN, Gaussian Naive Bayes, and SVM. For yield estimation, Decision Tree Regressor, Random Forest Regressor, KNN Regressor, SVR, and XGBoost Regressor can be compared. These algorithms represent different learning strategies, so the comparison provides a broader view of how the data behaves.

TABLE I. MODEL GROUPS USED IN THE PROPOSED FRAMEWORK

Task	Candidate models	Typical output
Crop recommendation	RF, Extra Trees, DT, LR, KNN, GNB, SVM	Crop class
Yield forecasting	DTR, RFR, KNNR, SVR, XGBoostR	Predicted yield
Explanation	Feature importance, LIME	Reason for prediction

D. VALIDATION AND EVALUATION

A model should not be judged only by its training accuracy. For classification, accuracy can be accompanied by precision, recall, F1-score, and a confusion matrix. For regression, MAE, MSE, RMSE, and R^2 provide complementary information. Cross-validation is useful because it checks whether the model remains consistent across different training and validation splits.

The supplied study used a 70% training and 30% testing split and also applied K-fold validation. For a future independent implementation, repeated cross-validation and a genuinely separate field dataset would provide stronger evidence of generalization.

E. EXPLAINABILITY

Explainable AI can make a prediction easier to inspect. Feature importance can indicate which variables contributed strongly to a model's decisions, while LIME can provide a local explanation for an individual prediction. In the supplied study, humidity and rainfall were among the influential variables identified for some crop-classification models.

For a farmer-facing application, the explanation should be expressed in simple terms. Instead of displaying only a probability value, the interface could say that the recommendation was influenced by soil pH, rainfall, and nutrient levels. Such an explanation does not prove that the recommendation is agronomically correct, but it makes the model's reasoning easier to inspect.

IV. BENCHMARK RESULTS AND DISCUSSION

The following results are included only as a benchmark summary of the supplied research paper; they are not claimed as new experiments performed for this rewritten manuscript. The source study reported 99.7% accuracy for its Random Forest crop-classification model and reported 99.4% and 99.2% for Extra Trees and Gaussian Naive Bayes, respectively. It also reported 98.9% for SVM, 98.1% for KNN, and 95.9% for Logistic Regression.



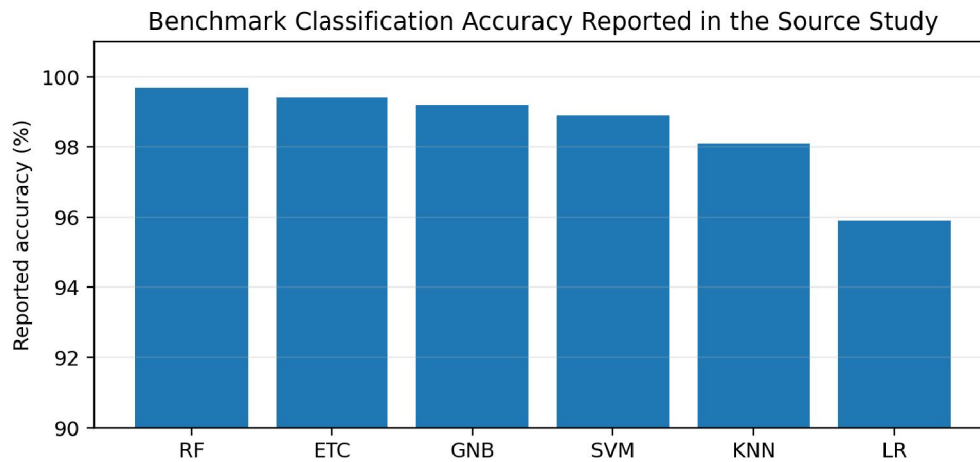


Fig. 2. Reported classification accuracy values from the supplied reference study.

TABLE II. REPORTED CLASSIFICATION BENCHMARKS FROM THE SUPPLIED STUDY

Model	Reported classification accuracy
Random Forest	99.7%
Extra Trees	99.4%
Gaussian Naive Bayes	99.2%
Support Vector Machine	98.9%
K-Nearest Neighbour	98.1%
Logistic Regression	95.9%

For the yield-prediction experiment, the supplied study reported that SVR produced MAE of 0.1614, MSE of 0.1047, and RMSE of 0.3236, with an R^2 value reported as 99.99%. These values should be interpreted within the dataset, preprocessing choices, time period, and experimental design of that study. They should not be assumed to represent performance on a different region or on current field data.

One practical observation is that high benchmark accuracy does not automatically mean that a model is ready for field deployment. Agricultural conditions vary by geography and season, and a dataset may not contain unusual weather, new crop varieties, soil changes, or measurement errors. A stronger deployment study would therefore test the model on data collected from locations and seasons that were not represented during training.

V. PRACTICAL APPLICATION DESIGN

A farmer-oriented system can be designed as a lightweight web or mobile application. The user enters or uploads soil and weather values, the server performs the same preprocessing used during training, and the selected model returns a crop suggestion or expected yield. The application can then display the important factors and a confidence indicator.

The system should also include input validation. For example, impossible pH values, negative rainfall, or missing nutrient measurements should trigger a warning rather than silently producing a prediction. A timestamp and location should be stored with each prediction so that future evaluation can determine whether the model remains reliable.

TABLE III. SUGGESTED APPLICATION MODULES

Module	Main responsibility	User-visible output
Input layer	Collect soil/weather values	Validated measurements
Preprocessing	Clean, scale, and transform data	Model-ready features



Prediction engine	Run trained classifier/regressor	Crop or yield estimate
Explanation layer	Identify influential inputs	Simple reason for prediction
Monitoring	Track later outcomes	Accuracy and drift reports

VI. LIMITATIONS AND FUTURE SCOPE

The framework has several limitations. First, model quality depends heavily on the quality and geographic coverage of the training data. Second, a model trained on one region may not transfer directly to another region because soil, climate, crop practices, and irrigation patterns can differ. Third, synthetic or imputed values can reduce uncertainty in a dataset but cannot fully replace real measurements. The supplied study itself notes the need for broader data and additional evaluation across different areas.

Future work can move toward real-time agriculture by combining machine-learning models with IoT soil sensors, weather services, satellite imagery, and farm-management records. A mobile application could make the predictions accessible to small farms. Models could also be retrained periodically to detect changes in climate and farming practices. Another useful direction is uncertainty estimation, where the system reports not only a predicted yield but also how confident the prediction is.

For student-level implementation, an achievable next step is to build a prototype using a public dataset, compare three to five algorithms, add a simple explainability module, and then test the final model on a small independently collected dataset. This would provide a clearer demonstration of the difference between a laboratory benchmark and real-world agricultural performance.

VII. CONCLUSION

Machine learning can support agricultural decision-making by turning soil, weather, and historical production data into understandable predictions. A useful system should do more than return a crop name or yield number: it should validate inputs, compare models fairly, explain important factors, and make its limitations clear. The supplied research demonstrates that conventional machine-learning methods can achieve strong benchmark results on structured agricultural datasets, while also showing the value of feature engineering, cross-validation, and explainability.

The framework presented in this paper reorganizes those ideas into a practical workflow suitable for further development. The next stage should focus on independent experiments, broader regional data, real-time measurements, and field validation before any model is used as a high-stakes farming recommendation tool.

ACKNOWLEDGMENT

The authors acknowledge the publicly available datasets and research literature that support work in machine-learning-based agricultural analytics. The supplied reference paper was used as a technical source for the benchmark context; the wording and organization of this manuscript were independently written for this submission.

REFERENCES

- [1]. A. Badshah, B. Y. Alkazemi, F. Din, K. Z. Zamli, and M. Haris, "Crop Classification and Yield Prediction Using Robust Machine Learning Models for Agricultural Sustainability," IEEE Access, vol. 12, 2024, doi: 10.1109/ACCESS.2024.3486653.
- [2]. A. Ingle, "Crop Recommendation Dataset," Kaggle, accessed 2024.
- [3]. Food and Agriculture Organization and World Bank, "Crop Yield Prediction Dataset," Kaggle, accessed 2024.
- [4]. R. Sharma, P. K. Tiwari, and K. Gupta, "Temporal crop yield prediction using LSTM networks," Agricultural and Forest Meteorology, vol. 318, 2024.
- [5]. Y. Chen, X. Zhang, H. Li, and M. Hu, "Random forest regression for predicting maize yield with climate and soil data," Field Crops Research, vol. 300, 2023.



- [6]. T. T. Nguyen, H. T. Pham, D. P. Hoang, and N. H. Vu, "Predicting rice yield using convolutional neural networks with satellite imagery," *Remote Sensing of Environment*, vol. 265, 2023.
- [7]. A. Banerjee, S. Ray, and M. Sarkar, "Machine learning in precision agriculture: A review of techniques and applications," *Computers and Electronics in Agriculture*, vol. 199, 2023.
- [8]. X. Gao, Y. Liu, L. Wang, and M. Sun, "Predicting crop yield using a linear regression model based on climate and soil data," *Computers and Electronics in Agriculture*, vol. 193, 2022.
- [9]. L. Zhang, Y. Liu, Q. Chen, and Z. Wu, "Support vector regression for crop yield prediction using meteorological and soil variables," *Computers and Electronics in Agriculture*, vol. 195, 2022.
- [10]. Y. Liu, M. Gao, Z. Huang, and S. Wang, "Deep learning models for wheat yield prediction using climate and soil data," *Computers and Electronics in Agriculture*, vol. 192, 2022.
- [11]. M. Al-Yaari, F. Kogan, and M. A. Wahid, "Automated irrigation systems using machine learning algorithms: A review," *Irrigation Science*, vol. 40, 2022.

