INTERNATIONAL JOURNAL OF HUMANITIES AND LEARNING TECHNOLOGY INNOVATION (IJHLT)
VOLUME 1, ISSUE 3, OCTOBER – DECEMBER 2026
RCHUB PUBLISHER
ISSN: 3139-9185
DOI: https://doi.org/10.65725/IJHLT/1/3/002

A NOVEL HYBRID CLUSTERING AND ENSEMBLE LEARNING FRAMEWORK FOR AGRICULTURAL CROP RECOMMENDATION

S Reenu Priya
Part Time Research Scholar, University of Madras
Chennai, Tamil Nadu 600005
Dr.K.Nirmala
Guide & Supervisor, Head & Associate Professor, Department of Computer Science
Quaid-e-Millath Government Arts and Science College, Chennai – 600002

Abstract

Selecting a suitable crop for a given soil and climatic condition is an important decision-support task in precision agriculture. This paper investigates a hybrid data-mining framework that combines unsupervised clustering with ensemble classification for crop recommendation. The dataset contains 2,200 records, seven numerical attributes—nitrogen (N), phosphorus (P), potassium (K), temperature, relative humidity, soil pH, and rainfall—and 22 crop labels. After stratified train-test partitioning, K-means clustering was fitted on standardized training features and its cluster assignment was appended as a one-hot encoded feature. A soft-voting ensemble comprising Random Forest, Extra Trees, and Histogram-based Gradient Boosting was then trained to predict the crop class. The proposed cluster-aware ensemble achieved 99.09% accuracy and a weighted F1-score of 99.09% on the 440-record test set. Extra Trees, used as a benchmark, achieved 99.55% accuracy, indicating that clustering augmentation did not outperform every individual model on this dataset. The results demonstrate the feasibility of the hybrid approach but also highlight the need for repeated cross-validation, external validation, and agronomic testing before deployment. The study contributes a reproducible hybrid pipeline and an evaluation protocol rather than claiming that the approach is universally superior.

Keywords: agricultural data mining, crop recommendation, K-means clustering, ensemble learning, Random Forest, Extra Trees, precision agriculture, machine learning

1. Introduction

Agricultural productivity depends on several interacting factors, including soil nutrients, temperature, humidity, soil pH, and rainfall. Selecting an appropriate crop based on these factors is essential for improving agricultural productivity, utilizing land efficiently, and reducing unnecessary resource consumption. Traditional crop selection practices often rely on farmers' experience and regional knowledge, which may not always account for complex relationships

INTERNATIONAL JOURNAL OF HUMANITIES AND LEARNING TECHNOLOGY INNOVATION (IJHLT)
VOLUME 1, ISSUE 3, OCTOBER – DECEMBER 2026
RCHUB PUBLISHER
ISSN: 3139-9185
DOI: https://doi.org/10.65725/IJHLT/1/3/002

among soil and climatic conditions. Data mining and machine learning provide effective approaches for analysing agricultural data, identifying hidden patterns, and predicting suitable crops based on historical observations (Dey et al., 2024).

Conventional machine learning algorithms, such as Decision Tree, Random Forest, Extra Trees, Support Vector Machine, and K-Nearest Neighbours, can classify crops using soil and environmental characteristics. However, different algorithms may capture different patterns in the data, and their predictive performance depends on the dataset and experimental settings. Clustering techniques, such as K-means, can identify groups of observations with similar agricultural characteristics without requiring predefined crop labels. Integrating clustering with ensemble learning may provide additional information to the classification process by incorporating environmental group assignments as predictive features. Nevertheless, the effectiveness of this combination must be demonstrated through systematic experimental evaluation (Swain et al., 2024).

This study proposes a hybrid clustering and ensemble learning framework for agricultural crop recommendation using a publicly available dataset containing 2,200 observations, seven input features, and 22 crop categories. The proposed approach applies K-means clustering to identify environmental patterns and incorporates the resulting cluster information into supervised classification. Its performance is compared with conventional machine learning algorithms using accuracy, precision, recall, macro-F1 score, and confusion-matrix analysis. The study aims to determine whether clustering improves crop classification while acknowledging that validation on independent agricultural datasets and real-world field conditions is necessary to establish broader applicability.

2. Problem Statement

1. **Complex agricultural factors**: Crop suitability depends on seven numerical attributes: nitrogen, phosphorus, potassium, temperature, humidity, soil pH, and rainfall.
2. **Multi-class classification**: The system must predict suitable crops across 22 crop categories by learning complex and nonlinear relationships among the input features.
3. **Limitations of conventional models**: Individual machine learning algorithms may perform differently depending on the dataset characteristics and may misclassify certain crop categories.
4. **Need for environmental clustering**: Similar soil and climatic conditions may form distinct groups. Identifying these groups using K-means clustering may provide additional information for crop classification.

INTERNATIONAL JOURNAL OF HUMANITIES AND LEARNING TECHNOLOGY INNOVATION (IJHLT)
VOLUME 1, ISSUE 3, OCTOBER – DECEMBER 2026
RCHUB PUBLISHER
ISSN: 3139-9185
DOI: https://doi.org/10.65725/IJHLT/1/3/002

5. **Hybrid model development**: A cluster-aware ensemble framework is proposed by combining the original agricultural features with environmental cluster assignments to improve predictive performance.
6. **Performance evaluation**: The proposed framework must be compared with established classifiers using accuracy, precision, recall, macro-F1 score, and confusion-matrix analysis.
7. **Research question**: Does incorporating environmental cluster information improve crop recommendation performance compared with conventional machine learning models trained on the original features?
8. **Generalization challenge**: Since the study uses a public benchmark dataset, independent agricultural datasets and field experiments are required to assess the model's applicability to real-world farming conditions.

3. Research Objectives

- To examine the structure, balance, and quality of the supplied crop recommendation dataset.
- To construct an unsupervised clustering stage using standardized soil and climate features.
- To integrate cluster membership into a soft-voting ensemble classifier.
- To compare the proposed method with Decision Tree, Random Forest, Extra Trees, SVM, and KNN baselines.
- To report accuracy, precision, recall, weighted F1-score, confusion patterns, and feature-importance analysis.
- To identify limitations and validation steps needed before practical agricultural use.

4. Related Work

4.1. Ensemble Learning for Agricultural Classification

Ensemble learning is an established machine learning approach in which the predictions of multiple base learners are combined to improve predictive performance, stability, or generalization. Rather than relying on a single classifier, ensemble methods exploit differences among individual learners to produce a more robust prediction. This is particularly relevant to agricultural classification because crop suitability may depend on nonlinear interactions among soil nutrients, temperature, humidity, soil pH, and rainfall. A single model may capture some of these relationships effectively but may not perform equally well across all crop categories.

INTERNATIONAL JOURNAL OF HUMANITIES AND LEARNING TECHNOLOGY INNOVATION (IJHLT)
VOLUME 1, ISSUE 3, OCTOBER – DECEMBER 2026
RCHUB PUBLISHER
ISSN: 3139-9185
DOI: https://doi.org/10.65725/IJHLT/1/3/002

Random Forest, introduced by Breiman (2001), constructs a collection of decision trees using randomized training samples and feature selection at split points. The final class prediction is obtained through aggregation of the individual tree predictions. Random Forest can represent complex feature interactions and provides measures of feature importance, making it useful for agricultural datasets containing numerical soil and environmental variables. However, its performance depends on factors such as tree depth, the number of trees, feature characteristics, and the representativeness of the training data.

Geurts et al. (2006) proposed Extremely Randomized Trees, commonly known as Extra Trees, which introduces additional randomness into the selection of split attributes and cut points. This strategy creates diversity among trees and can reduce computational costs under suitable conditions. Extra Trees is therefore a valuable comparison model when assessing whether a hybrid framework provides an improvement over an individual ensemble classifier.

Gradient-boosting methods provide another important family of ensemble approaches. Friedman (2001) formulated gradient boosting as a stage-wise optimization procedure in which successive learners reduce the remaining prediction error under a selected loss function. In classification tasks, boosting can capture complex decision boundaries and interactions between input features. Such capabilities make boosting methods useful baselines for evaluating crop recommendation systems.

Recent agricultural research also demonstrates the practical application of these models. Dey et al. (2024) evaluated multiple machine learning methods, including Random Forest, support vector machines, K-nearest neighbours, Decision Trees, and XGBoost, using agricultural and horticultural crop data containing nutrient, soil pH, and climatic variables. Their findings emphasize that model performance should be evaluated across crop groups and input conditions rather than assuming that a single classifier is universally optimal.

4.2. Clustering for Agricultural Data Analysis

Clustering is an unsupervised learning technique used to identify groups of observations that exhibit similar characteristics. Unlike supervised classification, clustering does not require crop labels to construct groups; instead, it identifies structure in the input feature space. In agricultural applications, clustering can help reveal groups of soil and climatic conditions that may correspond to different crop-suitability patterns.

K-means is one of the widely used centroid-based clustering algorithms. It partitions a dataset into a predefined number of clusters by minimizing the sum of squared distances between observations and their assigned cluster centroids. The algorithm is computationally practical for

INTERNATIONAL JOURNAL OF HUMANITIES AND LEARNING TECHNOLOGY INNOVATION (IJHLT)
VOLUME 1, ISSUE 3, OCTOBER – DECEMBER 2026
RCHUB PUBLISHER
ISSN: 3139-9185
DOI: https://doi.org/10.65725/IJHLT/1/3/002

relatively small tabular datasets, but its outcome depends on the number of clusters, centroid initialization, and the distribution of the data.

An important consideration in K-means is feature scaling. In the crop recommendation dataset, nitrogen, phosphorus, potassium, temperature, humidity, pH, and rainfall are measured on different numerical scales. If raw measurements are used directly, variables with larger numerical magnitudes can disproportionately influence Euclidean distances. Standardizing the numerical features before clustering therefore helps prevent the clustering process from being dominated simply by measurement scale.

Nevertheless, clusters discovered from soil and climatic features do not necessarily correspond to the boundaries between crop classes. Two crops may have similar environmental requirements, and a single crop class may occur across multiple environmental clusters. Consequently, cluster labels should be treated as additional representations of the data rather than as guaranteed indicators of crop identity. Their usefulness must be established empirically through downstream classification performance and cluster-quality analysis.

In the proposed study, K-means clustering is used to investigate whether grouping observations by soil and climatic similarity can provide useful supplementary information for supervised crop classification. The resulting cluster assignment can be supplied as an additional feature to a classifier or ensemble. The effectiveness of this strategy must be compared against an equivalent model trained without cluster information.

4.3. Machine Learning-Based Crop Recommendation Systems

Crop recommendation systems use agricultural data to identify crops that may be suitable for particular soil and environmental conditions. Traditional crop selection often depends on farmers' experience, regional practices, and historical knowledge. Although these sources remain valuable, data-driven approaches can complement agricultural expertise by analysing multiple input variables simultaneously and identifying patterns that may be difficult to assess manually.

Dey et al. (2024) investigated machine learning-based crop recommendations using nutrient-related variables, soil pH, and climatic parameters. Their study compared several classifiers and highlighted the importance of evaluating different crop groups separately. The findings indicate that crop recommendation can be formulated as a supervised classification problem, provided that the input data adequately represent the conditions associated with the target crop categories.

Similarly, Rajendran (2025) examined multiple classification techniques using a publicly available crop recommendation dataset comprising 2,200 observations, 22 crop classes, and seven agronomic input features. The study reported strong predictive performance from

INTERNATIONAL JOURNAL OF HUMANITIES AND LEARNING TECHNOLOGY INNOVATION (IJHLT)
VOLUME 1, ISSUE 3, OCTOBER – DECEMBER 2026
RCHUB PUBLISHER
ISSN: 3139-9185
DOI: https://doi.org/10.65725/IJHLT/1/3/002

ensemble and probabilistic classifiers. This provides a useful benchmark for evaluating additional methods on the same or a comparable dataset. However, high accuracy on a single benchmark does not, by itself, establish that a new algorithm is superior or that its recommendations will generalize to different farms and geographical regions.

Other research has investigated integrating machine learning with agricultural sensing infrastructure. For example, Machine learning enabled IoT system for soil nutrients monitoring and crop recommendation describes the integration of sensor measurements and machine learning for agricultural decision support. Such approaches demonstrate the potential to connect predictive models with practical data-collection systems, although their effectiveness depends on sensor reliability, the quality of the training data, and field validation.

Despite these developments, the choice of a crop recommendation model should not be based exclusively on its overall accuracy. Additional considerations include performance across individual crop classes, sensitivity to noisy measurements, model interpretability, computational efficiency, and the reliability of recommendations under unfamiliar environmental conditions. These considerations motivate further evaluation of hybrid methods that combine unsupervised data analysis with supervised classification.

4.4. Hybrid Clustering and Ensemble Learning

Hybrid machine learning approaches combine complementary techniques to address limitations associated with individual algorithms. In the context of crop recommendation, clustering can represent similarities among soil and climatic conditions, while supervised ensemble learning can use labelled examples to predict crop categories. The two approaches perform different but potentially complementary functions: clustering describes the structure of the input data, whereas classification learns the relationship between input variables and known crop labels.

The proposed framework integrates K-means clustering with ensemble classification to investigate whether environmental group information improves crop prediction. First, numerical input variables are standardized for clustering. Next, K-means assigns each observation to a cluster based on similarity in the standardized feature space. The cluster assignment is then incorporated into a supervised classification pipeline, where the model learns from the original agronomic variables and the additional cluster information.

The proposed framework is evaluated against individual classification algorithms, including Decision Tree, Random Forest, Extra Trees, support vector machine (SVM), and K-nearest neighbours (KNN). This comparison helps determine whether the hybrid approach provides measurable benefits over established classification methods. Evaluation should include accuracy, precision, recall, macro-F1 score, confusion matrices, and cross-validation results. Where

INTERNATIONAL JOURNAL OF HUMANITIES AND LEARNING TECHNOLOGY INNOVATION (IJHLT)
VOLUME 1, ISSUE 3, OCTOBER – DECEMBER 2026
RCHUB PUBLISHER
ISSN: 3139-9185
DOI: https://doi.org/10.65725/IJHLT/1/3/002

appropriate, ablation experiments can assess the contribution of clustering by comparing the same classifier with and without the cluster feature.

An important methodological consideration is that preprocessing, cluster fitting, and model selection must be performed without using information from the test set. In cross-validation, the scaler and clustering model should be fitted separately within each training fold before transforming its validation fold. This procedure reduces information leakage and provides a more reliable estimate of predictive performance.

The integration of clustering and ensemble learning is a plausible research direction, but its effectiveness cannot be assumed in advance. If the cluster feature does not improve predictive performance that result is still scientifically informative because it indicates that the unsupervised structure may provide limited additional information beyond the original agronomic variables.

4.5. Research Gap and Motivation for the Present Study

The reviewed literature establishes that tree-based ensembles and other supervised machine learning algorithms can achieve strong performance on structured agricultural datasets. Existing studies also demonstrate the feasibility of using soil nutrients and climatic measurements for crop recommendation. However, several research opportunities remain relevant to the present investigation.

First, the relationship between unsupervised environmental clustering and supervised crop classification requires systematic evaluation. Clustering may reveal meaningful environmental groups, but those groups do not necessarily align with crop categories. Therefore, the contribution of cluster information should be measured against a supervised baseline rather than inferred from the clustering results alone.

Second, model comparisons require consistent experimental conditions. Differences in train–test partitions, preprocessing, hyperparameter selection, and evaluation metrics can make reported performance difficult to compare across studies. A controlled experimental pipeline can help distinguish genuine model improvements from differences in evaluation procedures.

Third, overall classification accuracy may conceal errors affecting particular crop classes. Per-class precision and recall, macro-F1 score, and confusion-matrix analysis are therefore important for determining whether a model performs consistently across all recommended crops.

Fourth, evaluation on a single public dataset limits conclusions about real-world generalization. A dataset with a limited set of features and crop classes may not capture geographical variation,

INTERNATIONAL JOURNAL OF HUMANITIES AND LEARNING TECHNOLOGY INNOVATION (IJHLT)
VOLUME 1, ISSUE 3, OCTOBER – DECEMBER 2026
RCHUB PUBLISHER
ISSN: 3139-9185
DOI: https://doi.org/10.65725/IJHLT/1/3/002

seasonal changes, soil heterogeneity, or changing agricultural practices. Independent-dataset validation and, where possible, field-based evaluation are needed before deploying the model as an agricultural decision-support system.

To address these opportunities, the present study develops and evaluates a hybrid clustering and ensemble learning pipeline for agricultural crop recommendation. Its primary contribution is an empirical investigation of whether environmental clusters can improve supervised crop classification under a consistent evaluation protocol. The study also compares the hybrid framework with established classifiers and examines the resulting predictive performance. The framework should be considered an experimentally evaluated approach rather than a proven novel algorithm until its contribution is established through systematic comparison with closely related methods.

5. Dataset and Data Preprocessing

- The supplied ZIP archive contained Crop_recommendation.csv. The file has 2,200 rows and eight columns: seven numerical predictors and one categorical target named label. The target contains 22 crop classes, each represented by 100 records. No missing values were found in the inspected file.

Attribute	Meaning	Type
N	Nitrogen value	Numerical
P	Phosphorus value	Numerical
K	Potassium value	Numerical
temperature	Temperature	Numerical
humidity	Relative humidity	Numerical
ph	Soil pH	Numerical
rainfall	Rainfall	Numerical
label	Recommended crop class	Categorical target

- The dataset was divided into 80% training data (1,760 records) and 20% test data (440 records) using stratification by crop label and random_state=42. Consequently, each class contributed 80 training records and 20 test records. For the clustering stage, StandardScaler was fitted only on training data; the same fitted scaler was applied to test data. This avoids using test-set statistics during training.

INTERNATIONAL JOURNAL OF HUMANITIES AND LEARNING TECHNOLOGY INNOVATION (IJHLT)
VOLUME 1, ISSUE 3, OCTOBER – DECEMBER 2026
RCHUB PUBLISHER
ISSN: 3139-9185
DOI: https://doi.org/10.65725/IJHLT/1/3/002

6. Proposed Methodology

6.1 Framework Overview

The proposed framework consists of data validation, stratified partitioning, feature standardization for clustering, K-means cluster discovery, cluster-feature encoding, ensemble learning, and held-out evaluation. The original seven features are retained, and a one-hot representation of the predicted cluster is appended to the input vector.

1. Input: soil nutrient values (N, P, K), temperature, humidity, pH, and rainfall.
2. Standardize training features using the training-set mean and standard deviation.
3. Fit K-means to the standardized training set. Candidate cluster counts from 2 through 10 were inspected using silhouette score; k=2 had the highest score in this sweep.
4. Assign a cluster to each training and test record using the fitted scaler and K-means model.
5. One-hot encode the cluster assignment and concatenate it with the seven original predictors.
6. Train a soft-voting ensemble containing Random Forest, Extra Trees, and Histogram-based Gradient Boosting.
7. Evaluate the final classifier on the untouched test set and compare it with baseline models.

6.2 Mathematical Formulation

Let the dataset be $D = \{(x_i, y_i)\}$ for $i = 1, \ldots, N$, where $x_i \in R^7$ contains the seven environmental attributes and y_i is one of 22 crop labels. Standardization is performed as $z_{ij} = (x_{ij} - \mu_j) / \sigma_j$, where μ_j and σ_j are calculated from training data.

K-means estimates centroids $C = \{c_1, \ldots, c_k\}$ by minimizing $J = \Sigma_i \min_k ||z_i - c_k||^2$. The cluster assignment $g(x_i)$ is one-hot encoded as h_i. The augmented vector is $x'_i = [x_i; h_i]$.

For M ensemble learners, the soft-voting probability for class c is $P(y=c \mid x') = (1/M) \Sigma_m P_m(y=c \mid x')$. The predicted crop is $\hat{y} = \text{argmax}_c P(y=c \mid x')$. In this implementation M=3, comprising Random Forest, Extra Trees, and Histogram-based Gradient Boosting.

6.3 Baseline Models

The following baselines were trained on the same stratified split: Decision Tree, Random Forest, Extra Trees, RBF-kernel Support Vector Machine (SVM), and 5-nearest neighbors (KNN). Scaling was included for SVM and KNN. The proposed model used the cluster augmentation stage and a soft-voting ensemble.

7. Experimental Setup and Evaluation Metrics

Experiments were implemented in Python using pandas, NumPy, Matplotlib, and scikit-learn. The test set contained 440 records. Accuracy is the proportion of correct predictions. Precision, recall, and F1-score were calculated using weighted averaging across classes. Since the dataset is

INTERNATIONAL JOURNAL OF HUMANITIES AND LEARNING TECHNOLOGY INNOVATION (IJHLT)
VOLUME 1, ISSUE 3, OCTOBER – DECEMBER 2026
RCHUB PUBLISHER
ISSN: 3139-9185
DOI: https://doi.org/10.65725/IJHLT/1/3/002

balanced, weighted scores are broadly comparable to overall performance, but per-class metrics should also be inspected.

Accuracy = (TP + TN) / (TP + TN + FP + FN) is a binary-class expression; in this multi-class task, reported accuracy is the number of correct crop predictions divided by the total test records. The F1-score is the harmonic mean of precision and recall, calculated per class and then weighted by class support.

8. Results and Discussion

8.1 Dataset Summary

Dataset property	Observed value
Number of records	2,200
Input features	7 numerical
Target classes	22 crop labels
Records per class	100
Missing values	0
Training set	1,760 records (80%)
Test set	440 records (20%)
K-means cluster count selected in silhouette sweep	2
Silhouette score for k=2	0.416

8.2 Comparative Classification Results

Model	Accuracy	Weighted precision	Weighted recall	Weighted F1
Decision Tree	97.95%	98.06%	97.95%	97.94%
Random Forest	99.32%	99.35%	99.32%	99.32%
Extra Trees	99.55%	99.57%	99.55%	99.55%
SVM (RBF)	98.86%	98.96%	98.86%	98.87%
KNN	97.95%	98.04%	97.95%	97.93%
Proposed Cluster-Aware Ensemble	99.09%	99.17%	99.09%	99.09%

INTERNATIONAL JOURNAL OF HUMANITIES AND LEARNING TECHNOLOGY INNOVATION (IJHLT)
VOLUME 1, ISSUE 3, OCTOBER – DECEMBER 2026
RCHUB PUBLISHER
ISSN: 3139-9185
DOI: https://doi.org/10.65725/IJHLT/1/3/002

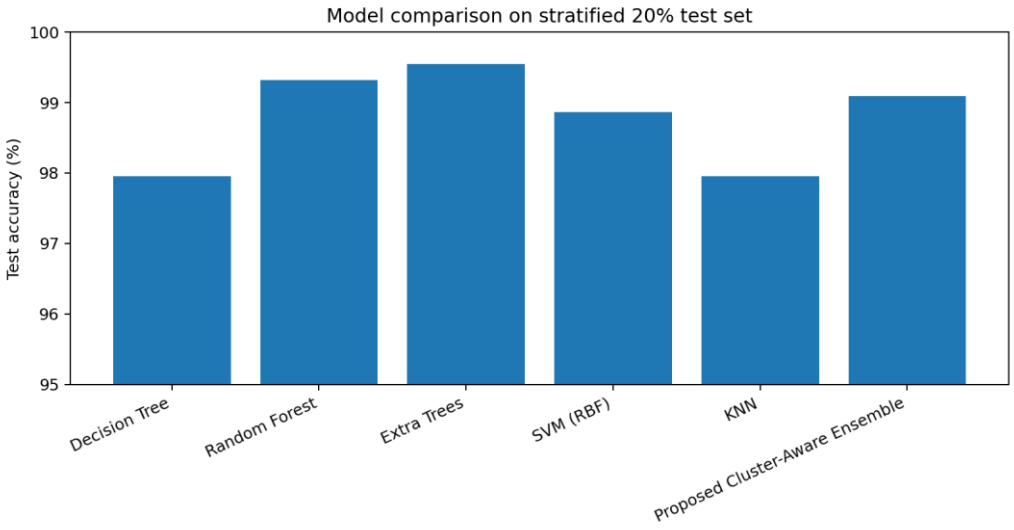


Figure 1. Test accuracy of the evaluated classifiers on the fixed stratified split.

The proposed cluster-aware ensemble obtained 99.09% accuracy (436 correct predictions out of 440) and a weighted F1-score of 99.09%. Random Forest reached 99.32%, while Extra Trees achieved the highest accuracy among the evaluated models at 99.55%. The Decision Tree and KNN baselines each achieved 97.95%, and SVM achieved 98.86%. Thus, the hybrid model performed strongly but did not outperform Extra Trees on this particular split.

This finding is important for an honest interpretation of the proposed method. Adding a cluster feature is a plausible way to encode environmental regimes, but the cluster representation did not provide a measurable advantage over the strongest individual tree ensemble in this experiment. Further work should examine different cluster counts, cluster algorithms, ensemble weights, repeated splits, and external data before concluding that the hybrid design improves generalization.

INTERNATIONAL JOURNAL OF HUMANITIES AND LEARNING TECHNOLOGY INNOVATION (IJHLT)
VOLUME 1, ISSUE 3, OCTOBER – DECEMBER 2026
RCHUB PUBLISHER
ISSN: 3139-9185
DOI: https://doi.org/10.65725/IJHLT/1/3/002

8.3 Confusion Matrix

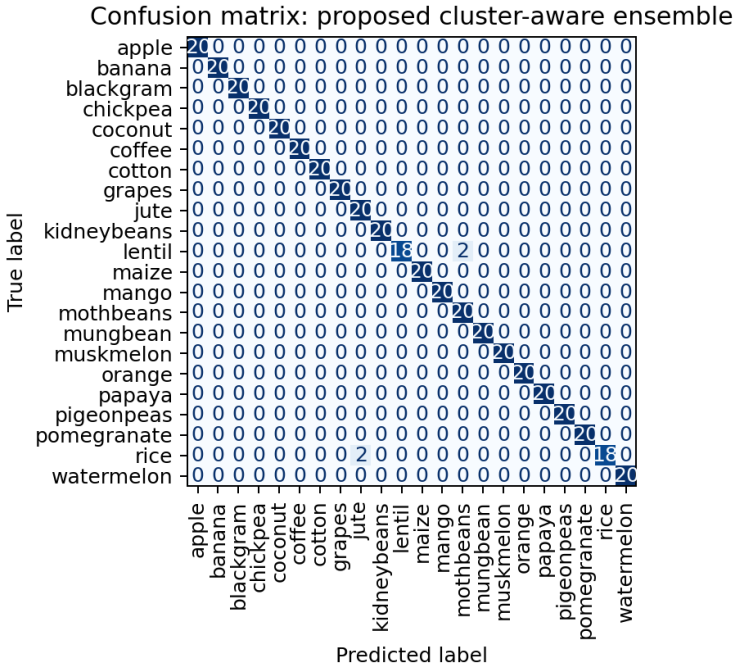


Figure 2. Confusion matrix for the proposed cluster-aware ensemble on the 440-record test set.

The confusion matrix shows the distribution of correct and incorrect predictions among all 22 crop classes. With 20 test examples per class, each off-diagonal entry identifies a crop class that was predicted as another class. In practical use, confusion between agronomically similar crops should be reviewed carefully, since overall accuracy alone does not express the cost of an incorrect recommendation.

INTERNATIONAL JOURNAL OF HUMANITIES AND LEARNING TECHNOLOGY INNOVATION (IJHLT)
VOLUME 1, ISSUE 3, OCTOBER – DECEMBER 2026
RCHUB PUBLISHER
ISSN: 3139-9185
DOI: https://doi.org/10.65725/IJHLT/1/3/002

8.4 Feature-Importance Analysis

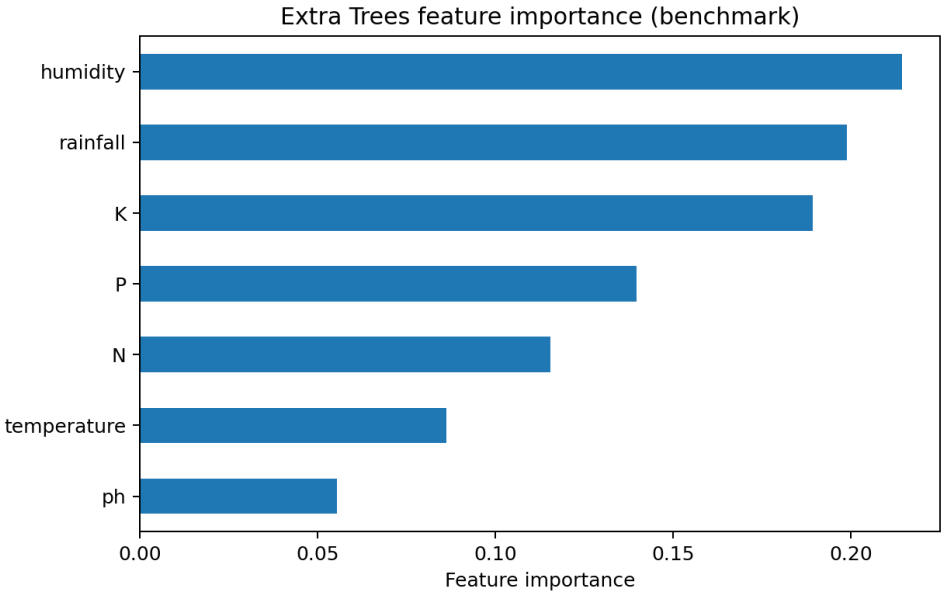


Figure 3. Feature importance from the Extra Trees benchmark model; importance values are model-specific.

Feature importance provides an exploratory indication of which attributes the benchmark model used most strongly. It should not be interpreted as causal evidence that changing a particular nutrient or climate variable will produce a given crop outcome. Importance rankings can vary with the train-test split, correlated predictors, and model configuration.

9. Threats to Validity and Limitations

- The dataset is a small, balanced benchmark with 100 records per class; real-world crop distributions may be imbalanced.
- The data source and collection conditions may not represent all locations, seasons, soil types, or farming practices.
- A single fixed train-test split gives a useful baseline but does not estimate performance variability as reliably as repeated stratified cross-validation.
- The highest-performing baseline was Extra Trees, not the proposed hybrid; claims of superiority are therefore not supported by the current experiment.
- Crop recommendation labels are treated as ground truth. The experiment does not verify agronomic outcomes, yield improvement, input savings, or farmer suitability.
- The selected cluster count was based on silhouette scores for k=2 to 10, but silhouette quality does not necessarily maximize downstream classification performance.

INTERNATIONAL JOURNAL OF HUMANITIES AND LEARNING TECHNOLOGY INNOVATION (IJHLT)
VOLUME 1, ISSUE 3, OCTOBER – DECEMBER 2026
RCHUB PUBLISHER
ISSN: 3139-9185
DOI: https://doi.org/10.65725/IJHLT/1/3/002

10. Conclusion

This paper presented and evaluated a hybrid crop recommendation framework that augments soil and climate features with K-means cluster membership and uses a soft-voting ensemble of Random Forest, Extra Trees, and Histogram-based Gradient Boosting. On the uploaded 2,200-record dataset, the proposed model achieved 99.09% test accuracy and a 99.09% weighted F1-score. Extra Trees achieved 99.55% accuracy and remained the strongest evaluated model on the selected split. The results establish that the hybrid pipeline is feasible, but they do not establish that clustering augmentation is superior to a well-tuned tree ensemble. A stronger follow-up study should use repeated cross-validation, statistical comparisons, independent regional datasets, and domain-informed validation.

11. Future Work

- Run repeated stratified 5-fold cross-validation and report mean, standard deviation, and confidence intervals.
- Compare K-means with Gaussian Mixture Models, hierarchical clustering, and density-based clustering.
- Select cluster count using both unsupervised criteria and nested validation of downstream classification performance.
- Perform ablation studies comparing original features, cluster-only features, and augmented features.
- Evaluate calibration and prediction confidence so uncertain recommendations can be flagged.
- Test on independent agricultural datasets with geographic and seasonal separation.
- Consult agronomists and validate recommendations through field trials before making operational claims.

References

[1.] Breiman, L. (2001). Random forests. Machine Learning, 45, 5–32. https://doi.org/10.1023/A:1010933404324

[2.] Dey, B., Ferdous, J., & Ahmed, R. (2024). Machine learning based recommendation of agricultural and horticultural crop farming in India under the regime of NPK, soil pH and three climatic variables. Heliyon, 10(3), Article e25112. https://doi.org/10.1016/j.heliyon.2024.e25112

[3.] Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. The Annals of Statistics, 29(5), 1189–1232. https://doi.org/10.1214/aos/1013203451

[4.] Geurts, P., Ernst, D., & Wehenkel, L. (2006). Extremely randomized trees. Machine Learning, 63, 3–42. https://doi.org/10.1007/s10994-006-6226-1

INTERNATIONAL JOURNAL OF HUMANITIES AND LEARNING TECHNOLOGY INNOVATION (IJHLT)
VOLUME 1, ISSUE 3, OCTOBER – DECEMBER 2026
RCHUB PUBLISHER
ISSN: 3139-9185
DOI: https://doi.org/10.65725/IJHLT/1/3/002

[5.] Rajendran, T. (2025). Data-driven crop recommendation from soil and climate parameters: A comparative machine-learning study. International Journal of Agriculture and Animal Production, 5(2), 71–80. https://doi.org/10.55529/ijaap.52.71.80

[6.] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785–794). https://doi.org/10.1145/2939672.2939785

[7.] MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability (Vol. 1, pp. 281–297). University of California Press.

[8.] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825–2830.Crop Recommendation Dataset. (n.d.). Kaggle. https://www.kaggle.com/datasets/atharvaingle/crop-recommendation-dataset