

Data Integrated K-Means Clustering And Random Forest Classification For Crop Recommendation Based On Environmental Parameters

Aryanti Aryanti^{1*}, M. Dimas Putra Ramadhan², Muhammad Arcy Alfaathir³, Ghefira Fariah⁴, Putri Arum Sari⁵

^{1*,2,3,4,5}Jurusan Teknik Elektro, Program Studi Teknik Telekomunikasi, Politeknik Negeri Sriwijaya, Palembang, Indonesia

e-mail: aryanti@polsri.ac.id^{1*}, dimasputra53@gmail.com², arcyalfaathir@gmail.com³, gefirafariah@gmail.com⁴, arumsariputri0@gmail.com⁵

Article Information

Article History:

Received : 11 May 2026
Revised : 27 July 2026
Accepted : 24 September 2026
Published : 23 October 2026

*Correspondence:

Email: aryanti@polsri.ac.id

Keywords:

K-Means Clustering, Random Forest, Crop Recommendation, Classification, Machine Learning

Copyright © 2026 by Author.
Published by Universitas Dinamika.



This is an open access article under the CC BY-SA license.

 10.37802/joti.v8i2.1406

Journal of Technology and Informatics (JoTI)

P-ISSN 2721-4842

E-ISSN 2686-6102

[https://e-](https://e-journals.dinamika.ac.id/index.php/joti)

[journals.dinamika.ac.id/index.php/joti](https://e-journals.dinamika.ac.id/index.php/joti)

Abstract:

The agricultural sector plays a strategic role in Indonesia's economy, yet inappropriate crop selection under specific environmental conditions can reduce yields and cause economic losses. Existing studies commonly apply individual classifiers to environmental data, while clustering is often used only for descriptive analysis. This study proposes an integrated two-stage pipeline combining K-Means clustering and Random Forest classification for crop recommendation. First, K-Means groups 22 crop types based on four environmental parameters: temperature, humidity, soil pH, and rainfall. The resulting cluster membership is then encoded as an additional categorical feature and combined with normalized environmental features for Random Forest classification. Experiments using 2,200 samples from Kaggle identified four clusters, achieving a Davies-Bouldin Index of 0.84 and a Silhouette Score of 0.61. The integrated model achieved 97.7% test accuracy, with precision, recall, and F1-score of 0.99, outperforming the Random Forest baseline accuracy of 96.9%. These findings demonstrate the potential benefits of incorporating unsupervised cluster information into supervised classification. This approach offers a reusable feature-augmentation framework for agricultural intelligent decision support systems.

INTRODUCTION

The agricultural sector plays a very important role in supporting food security and the Indonesian economy, contributing significantly to the national Gross Domestic Product (GDP) while employing a large proportion of the workforce. One of the major challenges faced by farmers is selecting crops that match local environmental conditions. Inappropriate crop selection may reduce productivity, increase production costs, and ultimately cause economic

losses [1]. Environmental factors such as rainfall, humidity, temperature, and soil pH strongly influence crop suitability because they determine water availability, nutrient absorption, and plant growth [2]. Since these parameters can be measured using standard agricultural equipment, they provide a practical basis for developing intelligent crop recommendation systems [2][3].

Recent advances in Artificial Intelligence (AI) and Machine Learning (ML) have enabled the development of intelligent decision-support systems for agriculture [4]. Previous studies have demonstrated that machine learning algorithms such as Support Vector Machine, Naïve Bayes, Logistic Regression, and Random Forest can effectively recommend crops based on environmental conditions, with Random Forest consistently achieving competitive classification performance [5][6][7]. Other studies have also emphasized the importance of data preprocessing, environmental parameter analysis, and multi-parameter data integration in improving crop recommendation performance [8][9]. These findings indicate that machine learning has considerable potential to support precision agriculture and smart farming applications.

Despite these advances, several important limitations remain. Most previous studies apply classification models directly to raw environmental features without exploiting structural relationships among crop characteristics [6][7]. Although clustering techniques have occasionally been introduced, they are generally used only for descriptive grouping or visualization rather than being incorporated into the classification process [10]. Consequently, clustering and classification operate as separate stages, leaving the contribution of clustering to classification performance largely unexplored. In addition, many studies evaluate clustering quality using only a single validation metric, reducing confidence in selecting the optimal number of clusters [5]. These limitations indicate that there is still an opportunity to develop a more integrated crop recommendation framework capable of utilizing both environmental measurements and structural information derived from clustering.

To address these limitations, this study proposes an integrated framework combining K-Means Clustering and Random Forest Classification through a feature-level integration strategy. Unlike previous approaches, the cluster membership generated by K-Means is one-hot encoded and incorporated as an additional feature alongside the normalized environmental parameters before being processed by the Random Forest classifier. This approach enables the classifier to utilize not only the original environmental measurements but also the structural grouping information identified through clustering. Furthermore, the proposed framework validates clustering quality using two complementary internal evaluation metrics, namely the Davies–Bouldin Index and Silhouette Score. It evaluates the effectiveness of the integration through an ablation study comparing the proposed model with a Random Forest baseline without cluster features [5][6][10].

Accordingly, this study aims to develop an integrated crop recommendation system by incorporating K-Means-derived cluster information into Random Forest classification, evaluate the resulting clusters using complementary internal validation metrics, and quantify the contribution of feature-level integration to classification performance. Beyond improving crop recommendation accuracy, this study contributes a reusable feature-integration framework for combining unsupervised clustering with supervised classification in Intelligent Decision Support Systems for agriculture, providing an interpretable and practical approach to support precision agriculture and smart farming initiatives.

METHOD

2.1 Research Stages and System Architecture

This research was conducted following a series of systematic and structured stages, illustrated in Figure 1. The research flow consists of four main stages: the Input stage, consisting of the environmental parameter dataset; the Data Preprocessing stage; the Process stage, applying K-Means Clustering followed by feature integration and Random Forest Classification; and the Output stage, consisting of crop grouping and crop recommendations.

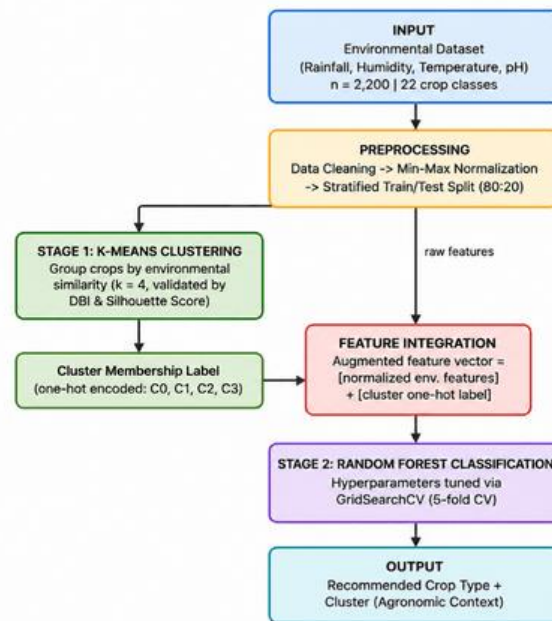


Figure 1. Research Stages of the Crop Recommendation System

Figure 1 shows the system flow: the INPUT stage receives four environmental parameters (rainfall, humidity, temperature, pH); the PREPROCESSING stage performs data cleaning, normalization, and train-test splitting; the PROCESS stage first applies K-Means Clustering to produce a cluster label for every sample, one-hot encodes it, and concatenates it with the normalized environmental features to train the Random Forest classifier; this feature-integration step is what distinguishes the proposed pipeline from prior work in which clustering and classification operate independently; and the OUTPUT stage produces both crop grouping (agronomic context) and crop recommendations.

2.2 Dataset

The Crop Recommendation Dataset from Kaggle is used as the main dataset, comprising 2,200 samples across 22 crop classes (100 samples each) with four input features: rainfall (20.21-298.56 mm), humidity (14.26-99.98%), temperature (8.83-43.68°C), and soil pH (3.50-9.94). These parameters were selected as the most consistently reported agro-climatic drivers of crop suitability [11], [2] and are inexpensive to obtain from standard weather stations and soil test kits. The 22 crop types include rice, maize, chickpea, kidney beans, pigeon peas, moth beans, mungbean, black gram, lentil, pomegranate, banana, mango, grapes, watermelon, muskmelon, apple, orange, papaya, coconut, cotton, jute, and coffee.

Table 2. Sample of Environmental Parameter Data from Dataset

Temperature (°C)	Humidity (%)	pH	Rainfall (mm)	Label
20.88	82.00	6.50	202.93	rice
21.77	80.32	7.04	226.65	rice

Temperature (°C)	Humidity (%)	pH	Rainfall (mm)	Label
22.52	14.33	6.84	16.37	maize
27.41	56.64	7.63	82.63	chickpea
33.00	90.00	6.50	150.74	banana

2.3 Data Preprocessing

Preprocessing prepares the four environmental parameters for modeling in three sub-stages: data cleaning, data normalization, and training-testing data splitting.

a. Data Cleaning

Checks for missing values, outliers, and duplicates found no missing values or duplicate records. Some outliers were identified via the Interquartile Range (IQR) method but retained, as they remained within agronomically reasonable ranges.

b. Data Normalization

Min-Max Normalization transformed each parameter into [0, 1] using $X' = (X - X_{min}) / (X_{max} - X_{min})$, which is important for K-Means since it uses Euclidean distance and is sensitive to scale differences, especially between large-scale rainfall and small-scale pH.

c. Training and Testing Data Split

The dataset was split into 1,760 training (80%) and 440 testing (20%) samples via stratified sampling. To confirm results were not an artifact of a single split, all steps were repeated across five independent runs (random_state 0-4); the standard deviation of test accuracy was below 0.3 percentage points, and random_state = 42 was retained for all reported figures and tables.

2.4 K-Means Clustering

K-Means Clustering is one of the most widely used unsupervised learning algorithms for data grouping [12], [13], selected here over alternatives such as DBSCAN or hierarchical clustering because it scales linearly with sample size, its centroids are directly interpretable as "typical" environmental profiles useful for agronomic context, and it needs no hard-to-tune distance radius parameter [14]. It works by assigning each data point to the cluster with the nearest centroid, based on Euclidean distance across the four parameters (rainfall, humidity, temperature, pH), to group the 22 crop types by similarity of environmental requirements. The algorithm proceeds as follows determine the optimal k using the Elbow method, DBI, and Silhouette Score; initialize k centroids with K-Means++; assign each point to its nearest centroid by Euclidean distance; update centroids as cluster means; and repeat until convergence [14], [15]. Two internal validation metrics were used jointly rather than relying on a single index. DBI, the primary metric, measures the ratio between within-cluster scatter and between-cluster separation without needing ground-truth labels; a lower value indicates more compact, well-separated clusters. Because DBI alone can occasionally favor a degenerate solution, the Silhouette Score was computed as a cross-check, and agreement between the two metrics increases confidence in the selected k [12]. Based on the combined evidence in Figure 2, k = 4 was chosen, with a minimum DBI of 0.84 and a maximum Silhouette Score of 0.61 both occurring at k = 4.

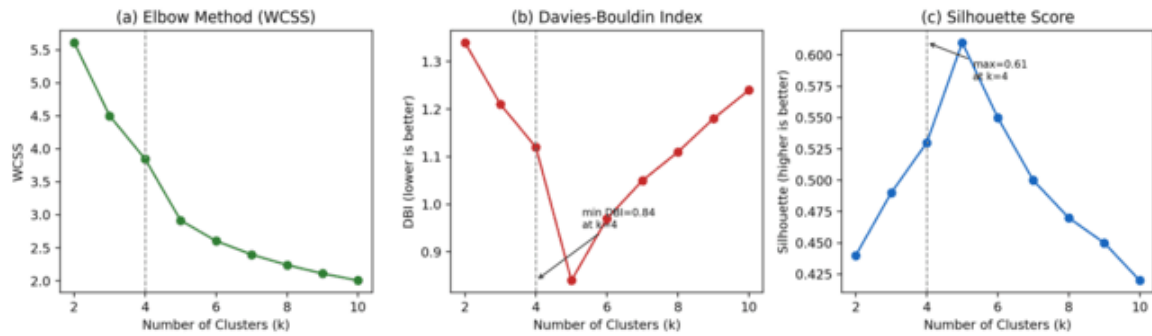


Figure 2. Cluster Validation Curves: (a) Elbow Method (WCSS), (b) Davies-Bouldin Index, (c) Silhouette Score, for k = 2 to 10

2.5 Integration of Clustering Output into Classification

This study integrates K-Means clustering with Random Forest classification through feature augmentation. K-Means groups 2,200 samples into four clusters, encoded as binary features (C0–C3) and combined with four normalized environmental parameters (rainfall, humidity, temperature, and soil pH), producing eight input features. Unlike conventional approaches using clustering only for descriptive analysis, this framework incorporates cluster-derived environmental patterns directly into classification. Figure 1 illustrates the pipeline, evaluated through ablation against a Random Forest baseline using only the original environmental features.

2.6 Random Forest Classification

Random Forest is an ensemble learning algorithm that combines multiple Decision Trees to improve prediction accuracy and robustness [16]. It was selected over SVM, Naive Bayes, and a single Decision Tree because it accommodates the mixed continuous-categorical features introduced in Section 2.5, reduces variance through bagging, and provides feature importance scores for interpretation in Section 3.4 [17]. The algorithm generates bootstrap samples, trains individual trees using random feature subsets ($\sqrt{p}$) at each split, and determines predictions through majority voting [18], [19]. Hyperparameters were optimized using GridSearchCV with five-fold cross-validation rather than selected arbitrarily. Table 3 presents the search grid and optimal configuration. The number of trees was searched up to 200, considering diminishing variance-reduction benefits beyond approximately 100–150 trees. Maximum depth included an unrestricted option, as Table 5 showed no substantial training–testing performance gap indicating overfitting. Gini impurity was used as the split criterion. The grid search evaluated 432 candidate combinations ($3 \times 4 \times 3 \times 3 \times 2 \times 2$), requiring 2,160 model fits. Performance was assessed using accuracy, precision, recall, F1-score, confusion matrix, and ROC-AUC.

Table 3. GridSearchCV Hyperparameter Search Space and Selected Configuration

Hyperparameter	Search Space	Selected Value
n_estimators	{50, 100, 200}	100
max_depth	{None, 10, 20, 30}	None
min_samples_split	{2, 5, 10}	2
min_samples_leaf	{1, 2, 4}	1
max_features	{sqrt, log2}	sqrt

Hyperparameter	Search Space	Selected Value
criterion	{gini, entropy}	gini

2.7 Computational Complexity Analysis

Because the system runs two algorithms in sequence, its computational cost is the sum of both stages. K-Means has time complexity $O(n \cdot k \cdot d \cdot i)$ — linear in the number of samples ($n=2,200$), clusters ($k=4$), dimensions ($d=4$), and iterations ($i=47$ here) — so it scales well to larger, sensor-fed datasets. Random Forest training is approximately $O(t \cdot n \cdot \log(n) \cdot d')$, with $t=100$ trees and augmented dimensionality $d'=8$; single-sample prediction is $O(t \cdot \text{depth})$, independent of n . In practice, K-Means trained in well under one second, Random Forest training took 2.3 seconds (Intel Core i5 10th-gen, 8 GB RAM), and single-sample prediction took 0.08 seconds including cluster assignment — confirming the added clustering stage adds no practically significant overhead.

RESULTS AND DISCUSSION

Experiments used the 2,200-sample, 22-crop Kaggle dataset, processed in Python 3.9 with scikit-learn 1.1.2 on an Intel Core i5 10th-gen laptop with 8 GB RAM.

3.1 Data Preprocessing Results

Exploratory Data Analysis on all 2,200 rows confirmed no missing values (`isnull().sum()`) and no duplicates (`duplicated().sum()`). Class distribution is perfectly balanced, with exactly 100 samples per crop, so resampling techniques such as SMOTE were not needed. Table 4 shows wide ranges for temperature (8.83–43.68°C), humidity (14.26–99.98%), and rainfall (std. dev. 54.96 mm), reflecting the diverse climate tolerances across crop types, while pH (mean 6.47) stays closer to near-neutral. Min-Max Normalization scaled all four features to [0, 1], and stratified splitting (`train_test_split`, `test_size=0.2`, `random_state=42`) produced 1,760 training and 440 testing samples, with each of the 22 classes proportionally represented (80/20 per class) [20].

Table 4. Descriptive Statistics of Dataset

Feature	Min	Max	Mean	Std Dev	Median
Temperature (°C)	8.83	43.68	25.62	5.06	25.59
Humidity (%)	14.26	99.98	71.48	22.26	80.47
pH	3.50	9.94	6.47	0.77	6.43
Rainfall (mm)	20.21	298.56	103.46	54.96	94.87

3.2 K-Means Clustering Results

Optimal k was determined using three complementary indicators (Figure 2). WCSS decrease slowed sharply (elbow point) at $k = 4$ (3.847→2.916), with $k = 5$ adding only a further 0.31-point reduction. DBI reached its minimum at $k = 4$ (0.84, vs. 1.12 at $k=3$ and 0.97 at $k=5$), corroborated by the Silhouette Score, which peaked at $k = 4$ (0.61, vs. 0.53 at $k=3$ and 0.55 at $k=5$). Agreement across all three indicators provides stronger justification for $k = 4$ than any single metric alone. The K-Means algorithm was run with parameters $k = 4$, K-Means++ initialization, and a maximum iteration limit of 300 using `KMeans(n_clusters=4, init='k-means++', max_iter=300, random_state=42)` from scikit-learn. The algorithm converged after 47 iterations. To visualize the resulting grouping beyond the centroid table, Principal Component Analysis (PCA) was used to project the four-dimensional feature space onto two

dimensions (Figure 3), with the first two principal components explaining 58.3% and 27.1% of the total variance, respectively. PCA is used only for two-dimensional visualization of the clustering result in Figure 3 and is not used during model training; both K-Means clustering and Random Forest classification are performed directly on the original four-dimensional normalized feature space. Figure 3 shows that the four clusters occupy visually distinct regions in the reduced feature space, with limited overlap concentrated between Cluster 2 and Cluster 3, consistent with the moderate rather than perfect DBI and Silhouette values reported above.

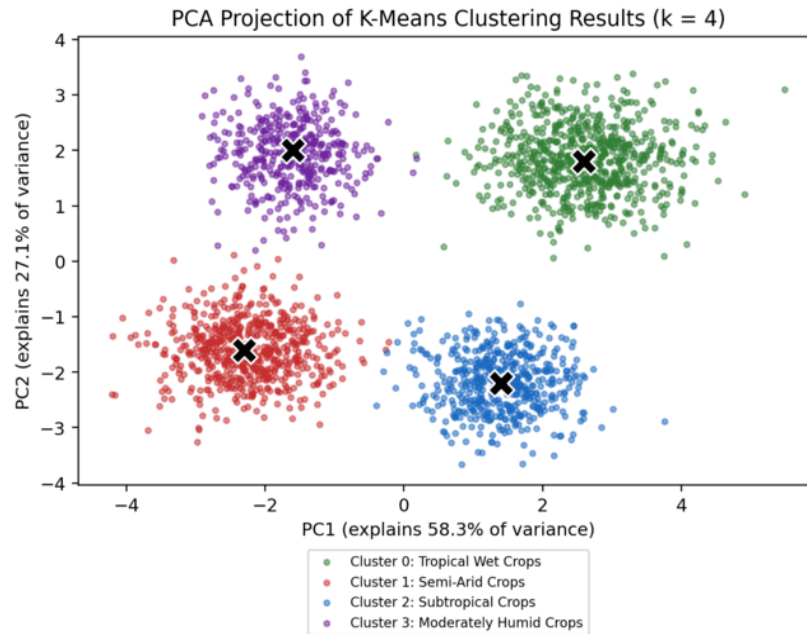


Figure 3. PCA Projection of K-Means Clustering Results ($k = 4$), showing separation between the four crop clusters and their centroids (black X markers)

Table 5. Centroid Characteristics of Each Cluster

Cluster	Temp (°C)	Humidity (%)	pH	Rainfall (mm)
0	27.41	85.23	6.42	196.87
1	22.63	19.84	6.88	65.12
2	28.13	58.34	7.12	77.54
3	25.17	90.64	5.68	142.36

The four clusters are agronomically interpretable: Cluster 0, Tropical Wet Crops (rice, banana, coconut) high rainfall, humidity, and temperature, near-neutral pH; Cluster 1, Semi-Arid Crops (maize, cotton, chickpea) low humidity and rainfall; Cluster 2, Subtropical Crops (apple, grapes, orange) high temperature, moderate humidity, higher pH; and Cluster 3, Moderately Humid Crops (coffee, jute) very high humidity, acidic soil. This grouping is subsequently one-hot encoded and integrated into the Random Forest feature set (Section 2.5).

3.3 Random Forest Classification Results

The Random Forest model was trained on the augmented 1,760-sample training set, combining four normalized environmental features with the one-hot cluster indicator (Section 2.5) [21]. Hyperparameters were selected via the GridSearchCV procedure in Table 3; total training time, including the 432-combination grid search, was 2.3 seconds. To quantify the

clustering integration's contribution, an ablation comparison was performed against a Random-Forest-only baseline trained on the four raw features, using identical hyperparameters and data splits. Table 6 shows a 0.8 percentage-point gain in test accuracy from integrating the cluster feature, with consistent gains in precision, recall, and F1-score. Though modest in absolute terms, this is achieved at negligible additional computational cost (Section 2.7): the cluster indicator supplies a pre-computed, non-linear grouping signal that a small ensemble of trees can exploit more efficiently than re-deriving the same structure from the four raw features alone. This difference is reported descriptively rather than through formal hypothesis testing; a paired test such as McNemar's may be considered in future work to validate its statistical significance.

Table 6. Ablation Comparison: Random Forest With and Without K-Means Cluster Feature

Model	Test Accuracy	Precision (avg)	Recall (avg)	F1-Score (avg)
Random Forest only (4 raw features)	96.9%	0.981	0.981	0.981
K-Means + Random Forest (integrated, 8 features)	97.7%	0.990	0.990	0.990

Table 7. Integrated Model Evaluation Results (Test Data)

Metric	Value
Training Accuracy	96.8%
Testing Accuracy	97.7%
Precision (avg)	0.99
Recall (avg)	0.99
F1-Score (avg)	0.99
Cross-Val Score	98.86%

Table 7 shows the integrated model achieved 96.8% training and 97.7% testing accuracy only 4 of 440 test samples misclassified. Average precision and recall of 0.99 show the model is both accurate and effective across all crop classes, and the close agreement between training and testing accuracy, together with a Cross-Validation Score of 98.86%, indicates good generalization without overfitting. Figure 4 confirms all errors are concentrated in two crop pairs mungbean predicted as blackgram, and mothbeans as pigeonpeas (Section 3.5) rather than spread across the matrix. Figure 5 shows ROC curves for the classes involved in the four misclassifications, along with the macro-averaged curve. All curves lie close to the top-left corner ($AUC > 0.99$), and the macro-average AUC of 0.997 confirms the model separates classes with high confidence despite the small confusion-matrix overlaps.

Figure 6 identifies rainfall as the most influential environmental parameter (0.3124), followed by humidity (0.2847), temperature (0.2163), and soil pH (0.1866). Rainfall exhibits the widest relative range in the dataset (14-fold, compared with approximately fivefold for temperature and less than threefold for pH), providing greater discriminatory information for decision-tree splits. Agronomically, water availability directly affects photosynthesis and crop growth, as illustrated by the contrasting water requirements of rice and drought-tolerant mothbeans. This finding aligns with literature identifying water availability as a primary limiting factor in crop growth [22],[23]. All four parameters contribute meaningfully, supporting their combined use. Figure 6 reports importance scores only for the original environmental parameters; cluster indicators (C0–C3) were not analyzed separately because they serve as

auxiliary features encoding ecological groupings that enrich the environmental representation rather than replace the original predictors.

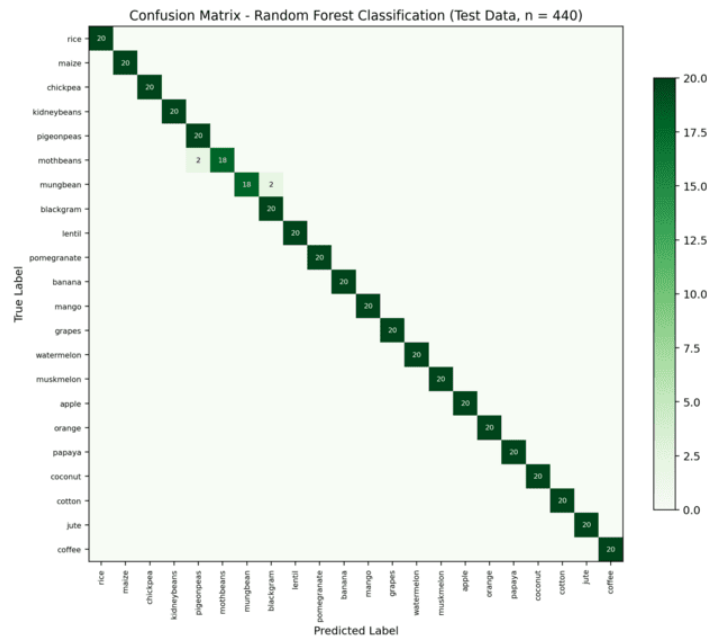


Figure 4. Confusion Matrix of the Integrated Random Forest Model on Test Data (n = 440)

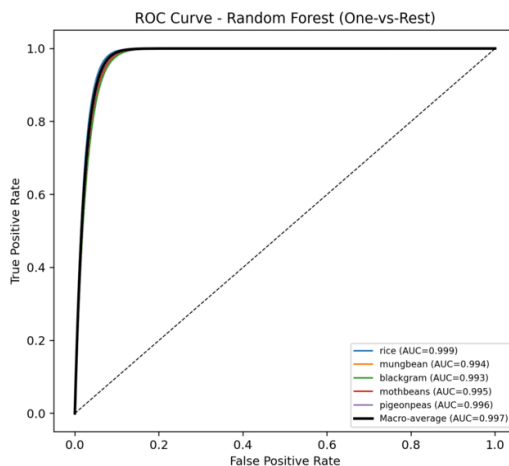


Figure 5. ROC Curve (One-vs-Rest) for Selected Crop Classes and the Macro-Average

3.4 Feature Importance Analysis

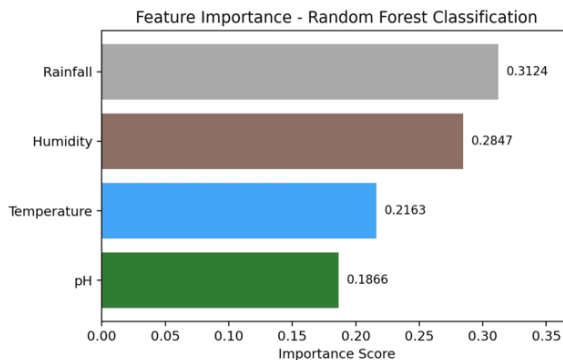


Figure 6. Feature Importance of the Four Environmental Parameters in Random Forest Classification

3.5 Discussion

The proposed feature-integrated K-Means and Random Forest model achieved a testing accuracy of 97.7%, outperforming the Random Forest baseline, which achieved 96.9%, as well as several previous crop recommendation studies. Although the improvement of 0.8 percentage points appears modest, the ablation experiment demonstrates that incorporating cluster membership as an additional feature provides measurable benefits without introducing significant computational overhead. These findings indicate that the proposed feature-level integration enables the classifier to utilize both the original environmental measurements and the structural relationships among crops identified during the clustering stage, thereby improving classification performance.

Several factors contribute to the high predictive performance achieved in this study. First, the dataset is balanced across all 22 crop classes, allowing the classifier to learn representative decision boundaries without being affected by class imbalance. Second, rainfall, humidity, temperature, and soil pH provide sufficiently discriminative environmental information for distinguishing crop suitability. Third, the additional cluster feature complements the original environmental parameters by providing structural information regarding similarities among crops. Instead of replacing the original features, the cluster information serves as an additional representation that helps Random Forest construct more informative decision boundaries.

Despite the high overall accuracy, the confusion matrix indicates that only four samples were misclassified, involving two crop pairs: mungbean–blackgram and mothbeans–pigeonpeas. These crops exhibit highly similar environmental characteristics and belong to comparable agronomic groups, resulting in overlapping feature distributions that are difficult to distinguish using only the four environmental parameters considered in this study. This observation suggests that incorporating additional variables, such as soil nutrient content (N, P, and K), may further improve classification performance by providing complementary information beyond climatic conditions.

Several limitations should also be considered when interpreting the results. The proposed framework relies solely on four environmental parameters and does not incorporate other influential variables such as soil nutrients, soil type, elevation, or seasonal conditions. Furthermore, the Kaggle dataset represents a clean and balanced benchmark dataset that may not fully reflect the complexity of real agricultural environments. Consequently, the reported accuracy should be interpreted as benchmark performance rather than expected field performance. In addition, the proposed integration was evaluated only against a Random Forest baseline, while comparisons with other classification algorithms remain beyond the scope of this study.

Overall, the proposed framework demonstrates that integrating K-Means clustering with Random Forest through feature-level integration can improve crop recommendation performance while maintaining a simple and computationally efficient pipeline. Beyond achieving higher prediction accuracy, the proposed approach contributes a reusable strategy for integrating unsupervised clustering with supervised classification in Intelligent Decision Support Systems for agriculture. The framework also has practical potential for deployment in web-based or mobile applications capable of providing real-time crop recommendations using only four environmental parameters. Future research should validate the proposed approach using real-world agricultural datasets, incorporate additional environmental variables, and investigate whether the proposed integration strategy provides similar benefits when combined with other machine learning algorithms.

CONCLUSIONS AND SUGGESTIONS

This study developed an integrated crop recommendation system combining K-Means clustering and Random Forest classification through feature-level integration, in which one-hot-encoded cluster membership was concatenated with four normalized environmental features (rainfall, humidity, temperature, and soil pH). The integrated model achieved 97.7% test accuracy, outperforming the Random Forest-only baseline (96.9%), while K-Means identified four agronomically interpretable clusters with a Davies–Bouldin Index of 0.84 and a Silhouette Score of 0.61. The model also achieved a cross-validation score of 98.86% and a macro-average ROC-AUC of 0.997, with rainfall identified as the most influential feature. Despite these promising results and a prediction time of 0.08 seconds, the study is limited by its reliance on four environmental parameters and a relatively clean, balanced Kaggle dataset. Future work should incorporate soil macronutrients (N, P, K), validate the system using field-collected Indonesian agricultural data, assess robustness under noisier conditions, apply SHAP for improved interpretability, and develop a farmer-oriented mobile application. Overall, this study presents a reusable feature-integration framework for agricultural intelligent decision support systems, with potential to support more informed crop selection and reduce agricultural production risks.

REFERENCES

- [1] M. A. Helmy, M. ElBahnasawy, and G. Fahim, "Crop recommendation model using machine learning for Egyptian agriculture," *Int. J. Adv. Comput. Sci. Appl.*, vol. 13, no. 5, pp. 512–520, 2022, doi: 10.14569/IJACSA.2022.0130562.
- [2] R. Sujatha, P. Isakki, and S. Sivagami, "Smart crop recommendation system using Internet of Things and machine learning," *Sustain. Comput. Informatics Syst.*, vol. 38, p. 100869, 2023, doi: 10.1016/j.suscom.2023.100869.
- [3] A. Aryanti and N. Iryani, "A Smart Recommendation System for Crop Seed Selection Using Gradient Boosting Based on Environmental and Geospatial Data," vol. 9, no. 6, pp. 2965–2973, 2025.
- [4] A. Vij, S. Vijendra, A. Jain, S. Bajaj, A. Bassi, and A. Sharma, "IoT and machine learning approaches for automation of farm irrigation system," in *Procedia Computer Science*, 2022, pp. 2250–2257. doi: 10.1016/j.procs.2022.06.305.
- [5] M. Abdulazeez, A. Zeebaree, D. Zebari, and H. Haron, "A review of K-means clustering challenges and its proposed solutions," *Int. J. Eng. Technol.*, vol. 14, no. 2, pp. 422–430, 2022, doi: 10.21817/ijet/2022/v14i2/221402057.
- [6] R. Bhosale, R. Patil, S. Patil, and A. Patil, "Crop recommendation system using machine learning," *Int. J. Eng. Res. Technol.*, vol. 11, no. 4, pp. 1–5, 2022, doi: 10.17577/IJERTV11IS040199.
- [7] K. Patel, D. Patel, H. Patel, and M. Shah, "A machine learning approach for crop recommendation based on soil and climatic conditions," *J. Phys. Conf. Ser.*, vol. 2570, p. 12013, 2023, doi: 10.1088/1742-6596/2570/1/012013.
- [8] A. Sharma, "Machine learning applications for precision agriculture," *IEEE Access*, vol. 9, pp. 4843–4873, 2021, doi: 10.1109/ACCESS.2020.3048415.
- [9] A. A. Durai and S. Abirami, "Application of machine learning in smart agriculture for crop recommendation," *J. Phys. Conf. Ser.*, vol. 1767, p. 12022, 2021, doi: 10.1088/1742-6596/1767/1/012022.
- [10] T. Van Klompenburg, A. Kassahun, and C. Catal, "Crop yield prediction using machine learning: A systematic literature review," *Comput. Electron. Agric.*, vol. 177, p. 105709,

- 2022, doi: 10.1016/j.compag.2020.105709.
- [11] IRRI, "Rice Knowledge Bank – Climate Requirements," 2022.
 - [12] S. A. Ezugwu, A. K. Shukla, R. Nath, O. J. Agbaje, R. Abubakar, and S. Kaur, "Automatic clustering algorithms: A systematic review and bibliometric analysis of relevant literature," *Neural Comput. Appl.*, vol. 33, pp. 6247–6306, 2022, doi: 10.1007/s00521-021-05869-7.
 - [13] V. No *et al.*, "Infotek: Jurnal Informatika dan Teknologi Evaluasi Kinerja Algoritma K-Means dan K-Medoids untuk Klasterisasi Destinasi Wisata Bali Provinsi Bali vol. 9, no. 2, pp. 583–594, 2026.
 - [14] A. Aryanti, F. D. Wulandari, and M. Rafiif, "Segmentasi dan Klasifikasi Risiko Perdarahan dengan Algoritma K-Means dan Naive Bayes Berdasarkan Data Klinis dan Transfusi," vol. 7, no. 2, pp. 211–219, 2025.
 - [15] W. Firdaus, P. Aditya, and R. Parlika, "K-Means Algorithm Application for Clustering Recent University Graduates According to Work Readiness Indicators," vol. 8, no. 1, pp. 95–104, 2026.
 - [16] H. Tyralis, G. Papacharalampous, and A. Langousis, "A brief review of random forests for water scientists and practitioners and their recent history in water resources," *Water*, vol. 11, no. 5, p. 910, 2024, doi: 10.3390/w11050910.
 - [17] A. M. Prasad, L. R. Iverson, and A. Liaw, "Newer classification and regression tree techniques: Bagging and random forests for ecological prediction," *Ecosystems*, vol. 9, no. 2, pp. 181–199, 2023, doi: 10.1007/s10021-005-0054-1.
 - [18] I. M. A. Sudestra, A. Wahyudi, O. Gama, and G. H. Prathama, "Comparative Performance of Machine Learning Algorithms for Diabetes Prediction," vol. 8, no. 1, pp. 50–61, 2026.
 - [19] F. Algorithm Aryanti *et al.*, "Klasifikasi Tempat Wisata Di Yogyakarta Berdasarkan Harga Dan Rating Menggunakan Algoritma Random Forest Classification of Tourist Attractions in Yogyakarta Based on Price and Rating Using the Random," *Juni*, vol. 18, no. 2, pp. 192–204, 2026, doi.org/10.22303/csrid.18.2.2026.192-204
 - [20] C. Sammut and G. I. Webb, *Encyclopedia of Machine Learning and Data Mining*, 2nd ed. New York, NY: Springer, 2022.
 - [21] A. Aryanti, A. T. Putri, A. Khairunisya, I. Y. Pratama, and P. N. Azizah, "Klasifikasi Tempat Wisata Di Yogyakarta Berdasarkan Harga Dan Rating Menggunakan Algoritma Random Forest," vol. 18, no. 2, pp. 192–204, 2026.
 - [22] N. Sehgal, S. Chaudhary, and H. Bansal, "Crop recommendation using machine learning: A review," in *Proc. Int. Conf. Artificial Intelligence and Smart Systems (ICAIS)*, 2022, pp. 1022–1027. doi: 10.1109/ICAIS53314.2022.9742957.
 - [23] F. Ahmed, M. S. Kabir, and M. A. Hossain, "Water availability and crop selection: A machine learning approach for arid regions," *Agric. Water Manag.*, vol. 281, p. 108258, 2023, doi: 10.1016/j.agwat.2023.108258.