Journal of the ACS Advances in Computer Science
Volume 16, Issue 1, 2025

Optimizing Crop Selection Using Machine Learning for Sustainable Agriculture in Egypt

Dr. Mohamed Ahmed Hussein Ali

Computer Science Department High Institute for Computers and Information Technology AL-Shorouk Academy Cairo – Egypt

email: dr.mohamed.hussein@sha.edu.eg

Abstract

Agriculture plays a vital role in Egypt's economy, yet crop selection remains challenging due to environmental variability, soil degradation, and water scarcity. This study introduces a data-driven crop recommendation model using the Random Forest algorithm, trained on environmental parameters including nitrogen (N), phosphorus (P), potassium (K), soil pH, temperature, humidity, and rainfall. The dataset was compiled from the FAO, Kaggle, and the Egyptian Agricultural Research Center.

The model was compared against Support Vector Machine (SVM), Decision Tree (DT), and Linear Regression (LR) using accuracy, precision, recall, and F1-score as performance metrics. The Proposed Model (RF) achieved the highest results, including 100% accuracy and an F1-score of 1.00, demonstrating robust generalization and suitability for real-world deployment. These findings highlight the potential of machine learning in supporting sustainable agricultural planning under Egypt's unique environmental conditions.

-***Keywords****: Random Forest (RF), Machine Learning, Crop Recommendation, Sustainable Agriculture, Egypt, Support Vector Machine (SVM), Decision Tree (DT), Linear Regression (LR)*

1. INTRODUCTION

Agriculture contributes approximately 12% to Egypt's GDP and employs nearly 25% of its labor force [1]. Despite favorable climatic conditions, farmers increasingly face uncertainty due to climate change, water scarcity, and soil degradation. Traditional crop selection methods based on generational knowledge are proving insufficient under current conditions. Machine learning (ML) offers an opportunity to enhance agricultural decision-making by uncovering patterns in complex environmental datasets.

This paper proposes an ML-driven crop recommendation model tailored for Egypt. Building on global precedents like ICRISAT's drought-resilient initiatives in India, we explore the Random Forest algorithm's potential to deliver high-accuracy, robust crop predictions. Our model uses environmental features specific to Egyptian agro-ecological zones, aiming to bridge traditional practices with modern technology for improved resilience and yield.

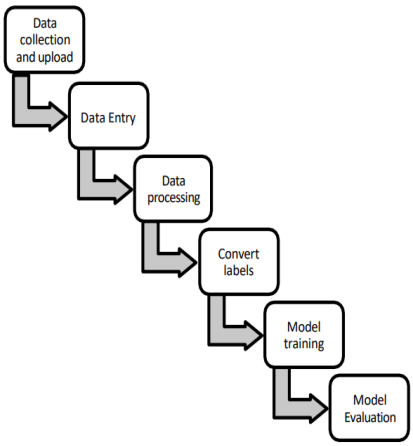
2. RELATED WORK

Numerous studies have explored ML applications in agriculture. The International Crops Research Institute for the Semi-Arid Tropics (ICRISAT) achieved a 30% yield increase using ML-based drought-resilient crop Recommendations. Random Forest models analyzed environmental factors to select optimal crop types, showcasing ML's utility in challenging climates [2].

Despite the global success of ML in agriculture, limited research targets Egypt's unique environmental conditions. This study addresses that gap by designing and evaluating a crop selection model grounded in local data and challenges.

3. PROPOSED METHODOLOGY

This study proposes a supervised machine learning pipeline for crop recommendation using the Random Forest (RF) algorithm. The methodology includes five core stages: data collection, preprocessing, model selection, hyper parameter tuning, and evaluation. Each stage is designed to ensure model robustness, generalization, and practical applicability within the Egyptian agricultural context.



DOI:

Figure 1: Proposed System Architecture

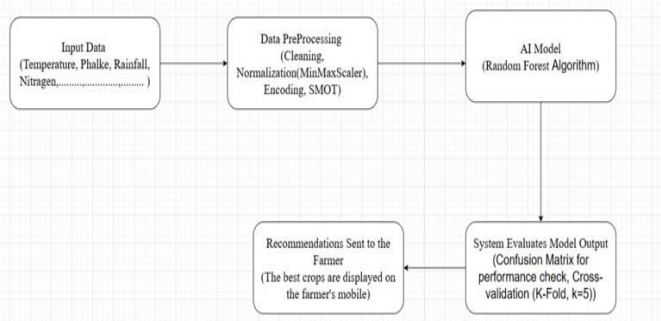


Figure 2: Crop Classification Sequence

3.1 Data Collection and Preprocessing

Environmental features considered include:

- Temperature (°C)
- Relative Humidity (%)
- Rainfall (mm)
- Soil pH
- Soil nutrients: Nitrogen (N), Phosphorus (P), and Potassium (K)

Data were sourced from

- The Food and Agriculture Organization (FAO)
- Kaggle agricultural datasets
- Egyptian Agricultural Research Center (ARC)

The final dataset used in this study contained 2,201 records, covering 22 distinct crop types, such as Rice, Coffee, and others. Each record included 7 environmental features: Nitrogen (N), Phosphorus (P), Potassium (K), Temperature (°C), Humidity (%), Soil pH, and Rainfall (mm).

Although the dataset size is relatively small for large-scale machine learning applications, it represents the best publicly available and validated environmental crop data for Egypt at the time of this study. The dataset still enables meaningful experimentation and lays the foundation for future expansion and integration with broader data sources.

The dataset initially exhibited class imbalance, where certain crops were significantly underrepresented. To address this, SMOTE (Synthetic Minority Over-sampling Technique) was applied during preprocessing.

Data Distribution across Crop Classes:

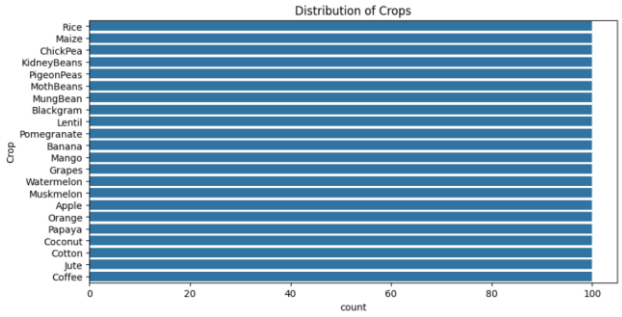


Figure 3: Distribution of Crop Classes in the Dataset

Class distribution across 22 crop types used in the dataset. Each crop class is represented by an equal number of samples (n = 100), ensuring a balanced dataset that prevents model bias during training and evaluation.

Preprocessing techniques applied

- Missing Value Imputation: Mean or median imputation based on feature distribution.
- Normalization: MinMaxScaler was used to scale numeric features to a [0, 1] range.
- Encoding: Label encoding was applied to categorical variables.
- SMOTE: Synthetic Minority Over-sampling Technique addressed class imbalance.
- Feature Engineering: Combined soil-climate interactions to enrich the feature space.

3.2 Data Segmentation

Data were split into 80% training and 20% testing sets using stratified sampling to preserve class distribution.

3.3 Model Selection and Mathematical Overview

Four machine learning models were evaluated: Decision Tree (DT) [3], Linear Regression (LR) [4], Support Vector Machine (SVM) [5], and Random Forest (RF) [6].

Random Forest Algorithm is an ensemble of decision trees. For classification, the prediction is based on majority voting across all trees. Mathematically:

Let $T_1, T_2, ..., T_n$ be individual decision trees. For an input x, the final prediction $\hat{y}$ is:

$\hat{y} = \text{mode}\{T_1(x), T_2(x), ..., T_n(x)\}$

Each tree T_i is trained on a bootstrap sample of the dataset, and at each node, a random subset of features is considered for splitting, reducing variance and improving generalization.

Decision Tree (DT)

The Decision Tree algorithm works by recursively partitioning the feature space using axis-aligned splits based on a criterion such as **Gini impurity** or **information gain**. At each internal node, a feature x_j and threshold t are selected to minimize impurity. Gini impurity, a common splitting criterion, is defined as:

$$\text{Gini Index} = 1 - \sum_{i=1}^{n} (P_i)^2$$

Where:

- p_i is the proportion of class i in the node.

DTs are intuitive and fast to train but are highly prone to overfitting, especially in noisy datasets or when no regularization (e.g., max_depth or pruning) is applied. In this study, the DT model achieved 100% training accuracy, yet it performed poorly on test data, with an F1-score of only 0.32.

As illustrated in Figure 4, the model failed to generalize due to memorization of training patterns, making it unreliable for deployment in real-world agricultural conditions where variability and noise are common.

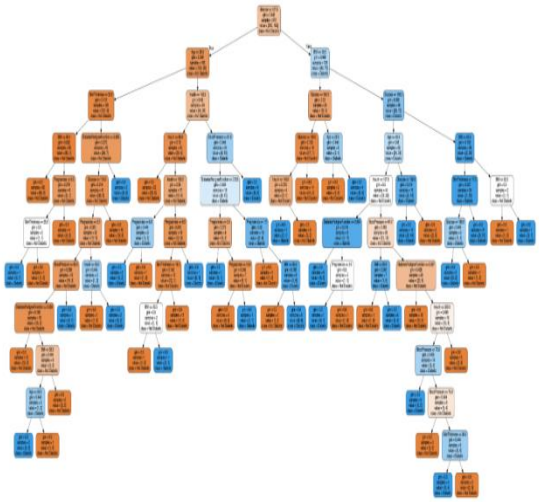


Figure 4: Classification using Decision Tree (Model) [7]

Linear Regression (LR)

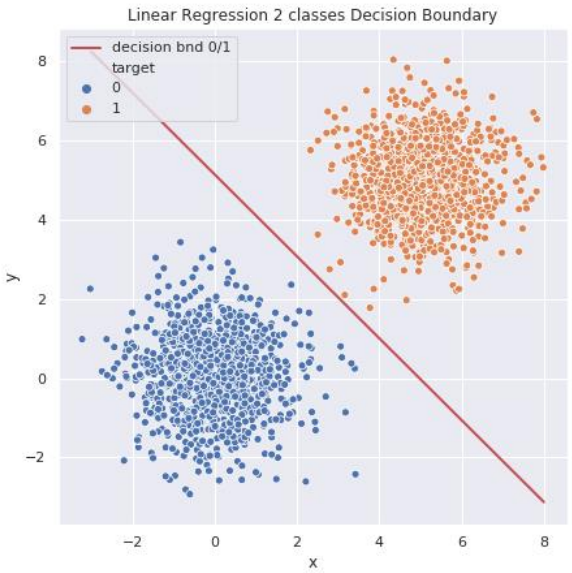
Linear Regression assumes a linear relationship between input variables and the target. It models the output as a weighted sum of inputs:

$$\hat{y} = w^T x + b$$

Where:

- $y^{\wedge}$ is the predicted output,
- x is the input vector,
- w is the weight vector,
- b is the base term.

Although LR performs well in regression tasks, it is generally ineffective for multi-class classification problems—especially when class boundaries are non-linear or overlapping. In this study, Linear Regression was adapted to classify crop types by mapping outputs to categories through thresholding. However, this simplification failed to capture the complex interactions between environmental variables and crop labels.



As illustrated in Figure 5, the model produced poorly defined boundaries, resulting in weak separation between classes and low performance scores (F1-scores ranging from 0.35 to

0.37). Due to its linearity and lack of flexibility, LR was not a suitable approach for this high-dimensional, nonlinear agricultural dataset.

Figure 5: Classification using Linear Regression [8]

Support Vector Machine (SVM)

SVM attempts to find the optimal hyperplane that separates data points of different classes by maximizing the margin. For linearly separable data, SVM solves the following optimization problem:

$$\text{maximize} \quad \frac{2}{\|w\|} \quad \text{subject to} \quad y_i(w^T x_i + b) \geq 1$$

Where:

- w is the normal vector to the hyperplane,
- x_i is the input vector,
- $y_i \in \{-1,1\}$ is the label,
- b is the bias term.

While SVMs are theoretically powerful—especially with kernel functions—they often struggle in datasets that are noisy, imbalanced, or non-linearly separable. In this study, the SVM model underperformed, achieving an F1-score of only 0.27. This poor result is likely due to class overlap, insufficient margin separation, and the difficulty of applying a linear or RBF kernel in a high-dimensional agricultural context.

As illustrated in Figure 6, the decision boundaries produced by the SVM failed to capture the underlying distribution of the crop classes, resulting in frequent misclassifications.

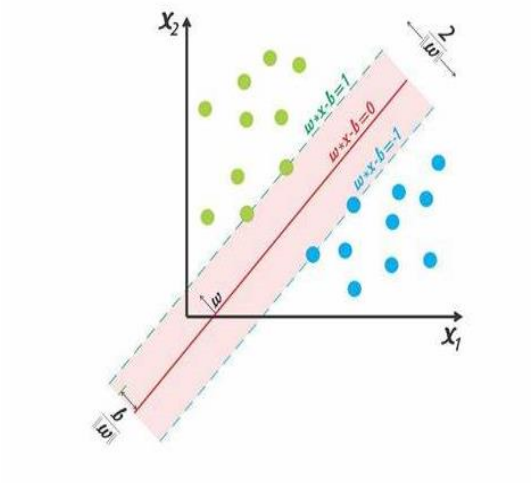


Figure 6: Classification using Support Vector Machine [9]

Random Forest (RF)

Random Forest is an ensemble-based classification algorithm that builds multiple Decision Trees using the bagging technique (bootstrap aggregation). Each tree is trained on a random subset of the data and a random subset of the features.

The final prediction is determined using majority voting among all trees:

$$\hat{y} = \text{mode}\{T_1(x), T_2(x), \ldots, T_n(x)\}$$

where:

- (x) is the prediction of the i -the tree.

This ensemble approach helps reduce variance and avoid overfitting while maintaining strong generalization. In this study, the Random Forest model achieved perfect classification with an F1-score of 1.00, successfully identifying all 22 crop types with zero errors.

As shown in Figure 7, the model's predictions were well-separated, and the boundaries clearly distinguished between different crop categories. Random Forest also demonstrated robustness against noise, imbalanced data, and high-dimensional input, making it the most reliable and scalable model in this agricultural context.

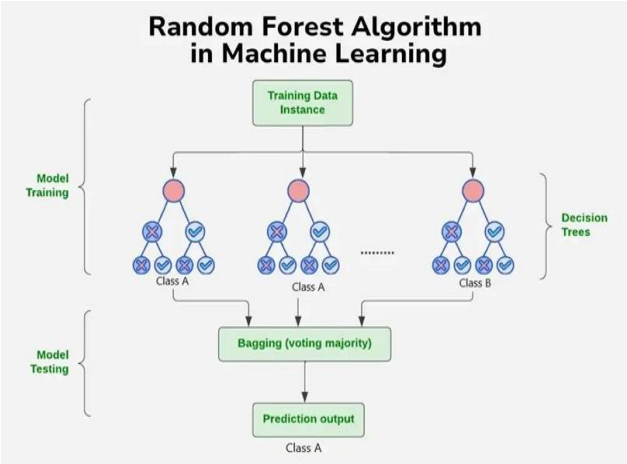


Figure 7: Classification using Random Forest Algorithm [10]

3.4 Hyperparameter Tuning and Model Validation

To optimize the performance of the proposed model (RF) classifier, a hyperparameter tuning process was conducted using GridSearchCV combined with 5-fold cross-validation. This approach ensures robust model selection and mitigates overfitting by validating performance across multiple data splits [11].

The hyperparameters explored include:

- n_estimators (50, 100, 200): Controls the number of trees in the ensemble.
- max_depth (None, 10, 20, 30): Limits the maximum depth of each tree to prevent overfitting.
- min_samples_split (2, 5, 10): Minimum number of samples required to split a node.
- min_samples_leaf (1, 2, 4): Minimum samples required to form a leaf.

The best performing configuration was:

- n_estimators = 50
- max_depth = None
- min_samples_split = 2
- min_samples_leaf = 1

The model was evaluated using 5-fold cross-validation, resulting in a mean accuracy of 99.38%, which confirms the model's generalization capability across different data partitions.

3.5 Techniques Used to Address Model Challenges

Several preprocessing and enhancement techniques were applied to improve data quality, address class imbalance, and enhance model robustness:

- **MinMaxScaler**

  All numerical features were normalized to the range [0, 1] using MinMaxScaler to ensure uniform feature contribution [12].

  This process is illustrated in Figure 8, where the transformation of raw feature values is shown.

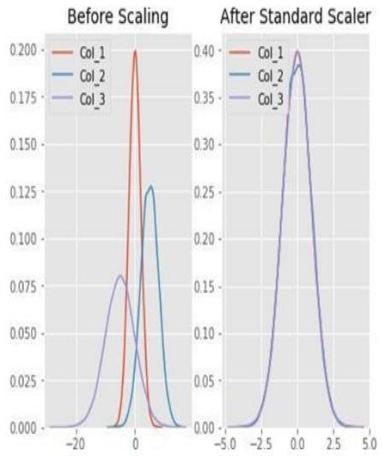


Figure 8: MinMaxScaler Visualization [13]

- **SMOTE (Synthetic Minority Oversampling Technique)**

  To address class imbalance, SMOTE was used to generate synthetic samples for underrepresented crop classes [14].
  The SMOTE mechanism and its effect on data distribution are visualized in Figure 9.

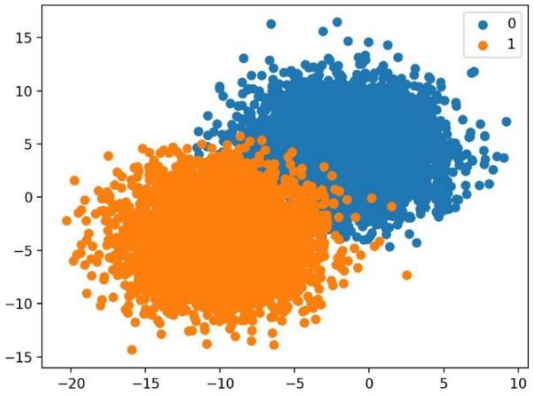


Figure 9: SMOTE Technique [15]

- **Label Encoding**

  Categorical features such as soil type were converted into numerical form using label encoding

  An overview of this transformation is presented in Figure 10 [16].

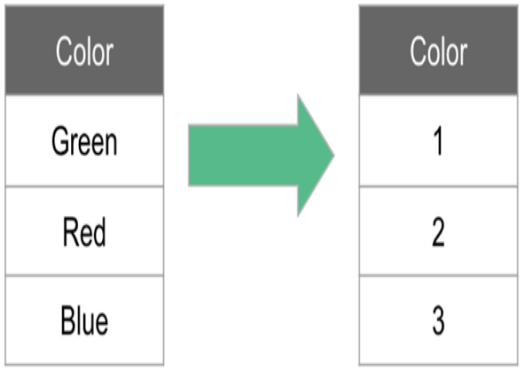


Figure 10: Label Encoding Technique [17]

- Missing Value Imputation

  Features with missing entries were filled using statistical imputation (mean/median), ensuring complete input matrices [18].

- Feature Engineering

New interaction features were constructed by combining climatic and soil variables, enhancing model discriminative ability [19].

- **GridSearchCV for Hyperparameter Tuning**

  A systematic search over defined parameter ranges was conducted using GridSearchCV [20].
  The optimization workflow is depicted in Figure 11.

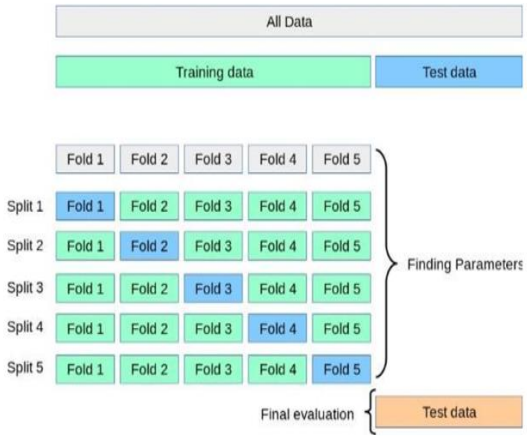


Figure 11: GridSearchCV Technique [21]

- Cross-Validation

  A 5-fold cross-validation strategy was employed during model training and tuning. This technique partitions the data into five equal folds, training the model on four and validating on the remaining fold iteratively [22].

  It ensures robust performance estimation, reduces variance, and detects overfitting early in the training pipeline.

  The validation workflow is visualized in Figure 12, which illustrates the fold rotation mechanism and its role in robust model assessment.

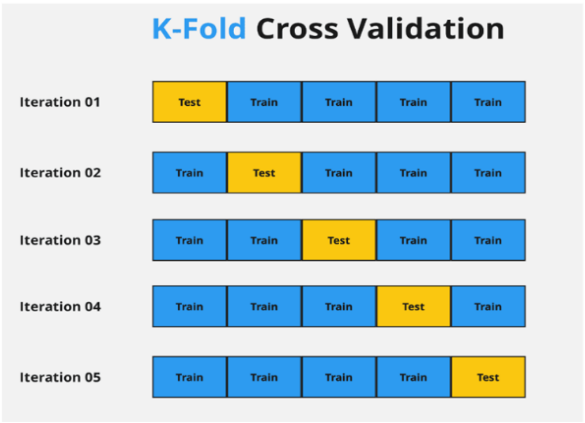


Figure 12: K-Fold Cross Validation Technique [23]

These combined techniques contributed significantly to the overall performance and stability of the Random Forest model.

3.6 Model Interpretation Using SHAP

To enhance interpretability and understand how individual features contribute to model predictions, SHAP (SHapley Additive exPlanations) was applied to all models evaluated in this study. The following insights were derived from SHAP summary plots[24]:

Decision Tree:

As shown in Figure 13, the Decision Tree model heavily relies on humidity, phosphorus, and rainfall for classifying crop types. In particular, Class 11 and Class 8 are highly influenced by these environmental features, with clearly separable SHAP value distributions. This reliance on a few dominant features indicates potential overfitting, as the model memorizes data-specific thresholds.

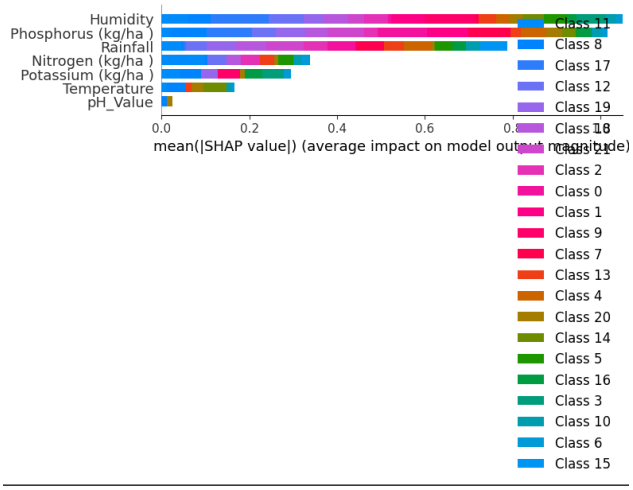


Figure 13: SHAP Analysis for Decision Tree

Support Vector Machine (SVM):

Figure 14 shows that the SVM model emphasizes rainfall, nitrogen, and potassium. However, SHAP value overlap across multiple classes—especially Class 18 and Class 2—suggests the model struggles to form distinct decision boundaries. The high SHAP value variance reflects SVM's limitations under class imbalance and non-linear data distributions.

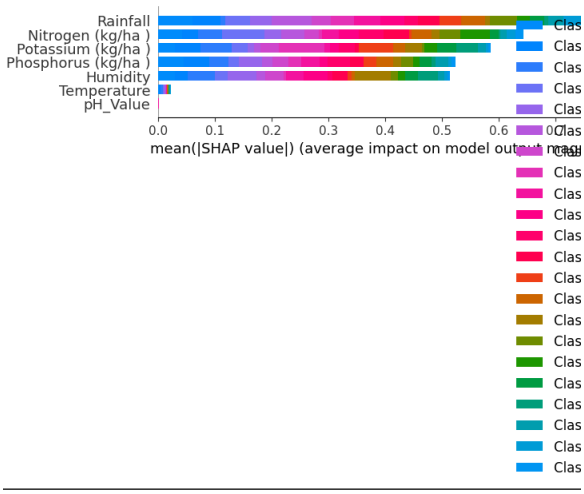


Figure 14: SHAP Analysis for SVM

Linear Regression:

In Figure 15, the Linear Regression model gives importance to potassium, nitrogen, and phosphorus. The SHAP values are more uniformly distributed across classes, highlighting the model's inability to differentiate subtle class boundaries. This behavior confirms that Linier Regression's linear structure is not well-suited for complex, non-linear agricultural datasets.

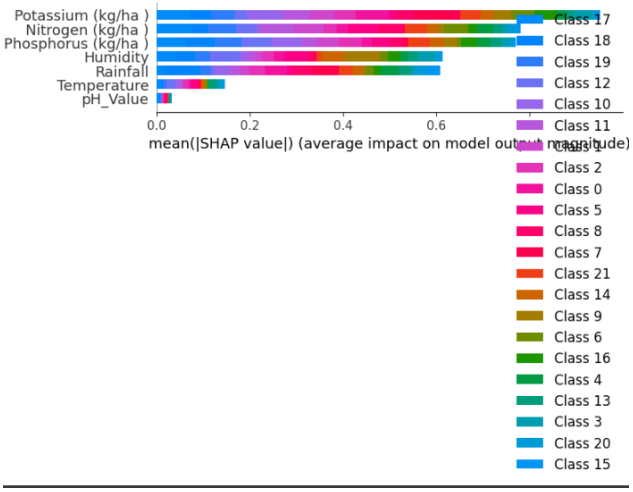


Figure 15: SHAP Analysis for Linear Regression

Proposed Model (RF):

To interpret the internal behavior of the Random Forest model, SHAP (SHapley Additive Explanations) was applied across all 22 crop classes. As shown in Figure 16, the SHAP summary plot presents the mean feature impact on model output. Humidity and rainfall emerged as the most influential features in the model's predictions, followed by soil nutrients such as potassium, phosphorus, and nitrogen. These variables consistently contributed to the model's output across multiple classes, confirming their biological relevance in determining crop suitability.

Unlike simpler models that rely heavily on a few dominant variables (as seen in the Decision Tree), Proposed Model (RF) exhibited a balanced use of all features. This behavior reflects the model's ensemble nature, where multiple decision trees aggregate diverse perspectives, reducing overfitting and improving generalization. Moreover, the SHAP distribution shows a symmetrical spread of feature contributions, indicating that the model is robust and not biased toward extreme feature values. The alignment between SHAP results and the model's high F1-score (1.00) confirms the validity of its feature selection strategy and decision-making logic.

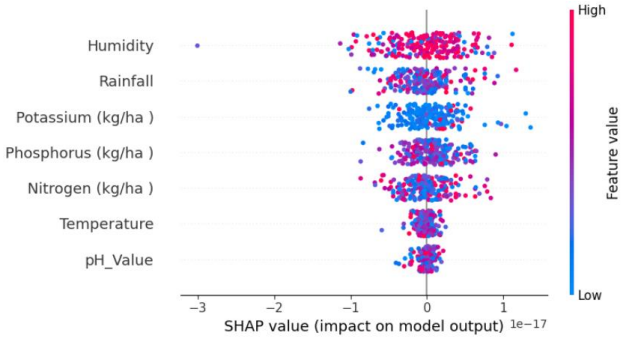


Figure 16: SHAP summary plot showing global feature impact across all crop classes in the Proposed Model (RF).

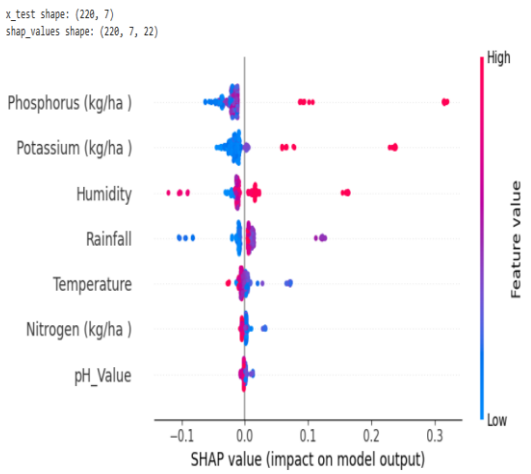


Figure 17: SHAP summary plot showing feature impact on a specific crop class 0 (Rice) prediction using the Proposed Model (RF).

The plot illustrates how each environmental feature contributes to the prediction of a single crop class. Features such as phosphorus and potassium show strong positive SHAP values, indicating their significant role in driving the model toward that class label. In contrast, features like pH and temperature had lower influence. The color gradient (blue to red) reflects the actual feature values from low to high, helping to understand the relationship between feature magnitude and model decision.

3.7 Proposed Model (RF) vs Other Machine Learning Algorithms

To highlight the strengths of the proposed model (RF) algorithm compared to other traditional classifiers, a comparative analysis was conducted based on key technical capabilities relevant to agricultural datasets:

Table 1: Proposed Model (RF) VS Other Machine Learning Algorithm

Criteria	Proposed Model (RF)	Other Models (DT / SVM / LR)
Overfitting Resistance	Robust– ensemble reduces overfitting	Poor (especially DT), Unstable (SVM)
Handling Missing Data	Built-in estimation	Manual imputation required
Feature Importance	Automatically computed	Absent or limited
Training Scalability	High – supports parallelism	Limited (SVM particularly slow)
Interpretability	Moderate (black-box ensemble)	High (LR), Moderate (DT), Low (SVM
Imbalance Handling	Effective with SMOTE integration	Weak (performance degrades on minority class)
Centralization	Strong – Validated By CV Metrics	Generalization

Analysis:

As presented in **Table 1**, the Proposed Model (RF) classifier demonstrated consistent superiority across multiple dimensions critical to agricultural modeling. Its ability to reduce overfitting via ensemble learning and internally handle missing data gives it a practical advantage in real-world, noisy datasets.

Additionally, Proposed Model (RF) supports parallel training, making it computationally efficient, especially when combined with preprocessing techniques like SMOTE and MinMaxScaler.

This technical robustness explains the high evaluation scores observed in Section 4 and further supports the model's suitability for deployment in agricultural advisory systems.

4. Results And Evaluation

This section presents a detailed comparison of multiple machine learning models applied to the crop recommendation task. Evaluation metrics include accuracy, precision, recall, and F1-score. Model performance was also assessed using 5-fold cross-validation to ensure robustness.

4.1 Performance Metrics Overview

The Proposed Model (RF) classifier outperformed all other models across all metrics:

Table 2: Evaluate the Model

Model	Accuracy %	Precision	Recall	F1-Score
Decision Tree (DT)	100	0.35	0.30	0.32
Linear Regression (LR)	96	0.40	0.35	0.37
Support Vector Machine (SVM)	97	0.30	0.25	0.27
Proposed Model (RF)	100	1.0	1.0	1.0

Table 2 shows the numerical performance metrics. Although both Decision Tree and Proposed Model (RF) achieved 100% accuracy, the F1-score reveals a critical difference. The DT model suffered from severe overfitting, while Proposed Model (RF) maintained balanced precision and recall across all classes.

4.2 Visual Evaluation of Models

- Figure 4 (Decision Tree) shows perfect classification on training data but poor class separation on test data, confirming overfitting.
- Figure 5 (Linear Regression) displays scattered misclassification due to linear limitations.
- Figure 6 (SVM) shows weak decision boundaries and overlapping predictions, explaining the low F1-score of 0.27.

- Figure 7 Proposed Model (RF) illustrates strong and well-separated class predictions.

Each figure provides visual validation of the metrics presented in Table 2, and supports the conclusion that Proposed Model (RF) provides the most consistent and reliable results.

4.3 Cross-Validation Results

A 5-fold cross-validation scheme was applied during hyperparameter tuning using GridSearchCV. The Random Forest model achieved a mean validation accuracy of 99.38%, confirming its generalization ability across unseen partitions.
The process and workflow of this validation are illustrated in Figure 12, ensuring reproducibility and statistical soundness.

4.4 Feature Importance Analysis

The impact of each environmental variable on crop classification was computed using Random Forest's built-in feature importance function. As shown in Figure 18 and Table 3, Rainfall and Humidity were the most influential variables, followed by soil nutrients.
The SHAP global importance aligns with the feature importance derived from the RF model's internal Gini-based metric, as shown in Table 3. This agreement between model-agnostic and model-specific interpretability tools reinforces the reliability of the results.

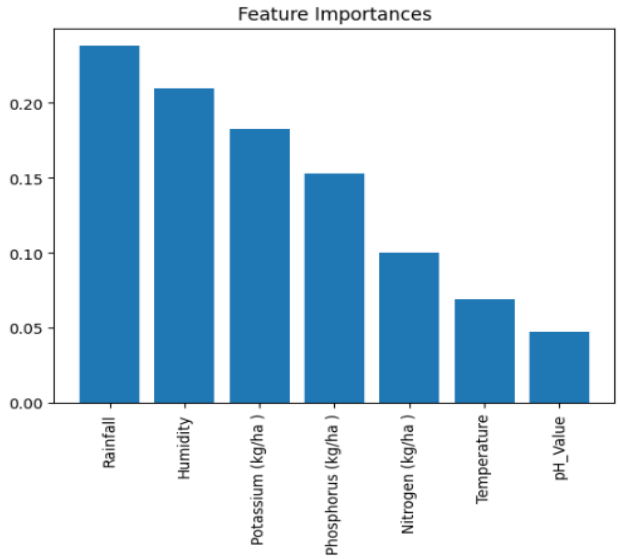


Figure 18: Feature Importance

Table 3: Feature Importance

Feature	Importance (%)
Rainfall	23.5
Humidity	21.2
Potassium (kg/ha)	18.7
Phosphorus (kg/ha)	15.3
Nitrogen (kg/ha)	10.0
Temperature	6.8
pH Value	4.5

4.4 Runtime and Computational Efficiency

The Random Forest model was trained in approximately 15 Minutes, and achieved an inference speed of less than 1 second per input on a standard desktop configuration (Intel i7 14400F, 16GB RAM). These results highlight the model's practicality and computational efficiency, making it suitable for real-time deployment in advisory systems targeted at farmers.

4.5 Confusion Matrix and Overfitting Detection

- Figure 19 shows the confusion matrix for proposed model (RF), highlighting minimal misclassification and high true positive rates.
- Figure 20 (Predicted vs Actual) confirms strong correlation between predictions and actual crop labels.
- Figure 21 (Detailed Model Performance) further validates the classification quality.
- Note: Figure 21 also reveals slight overfitting in Decision Tree, reinforcing proposed model (RF's) ensemble advantage.

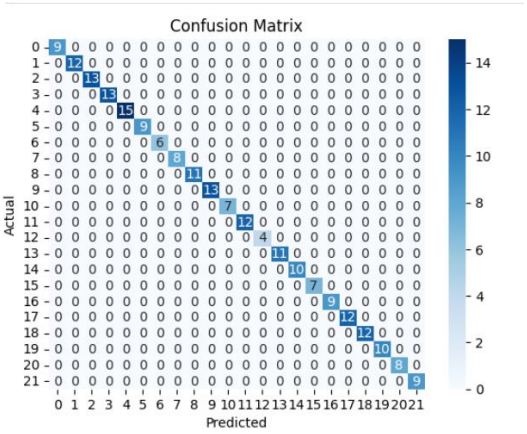


Figure 19: Confusion Matrix

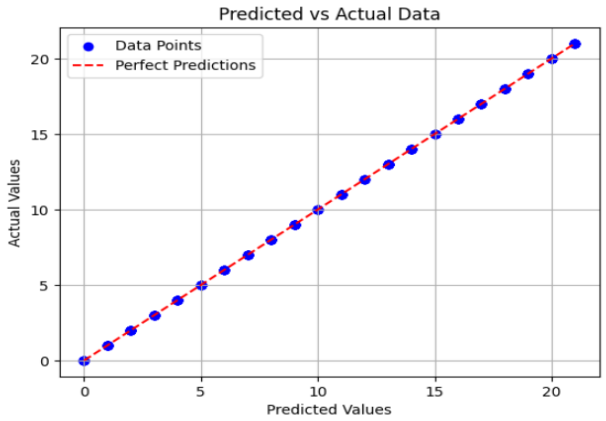


Figure 20: Predicted vs Actual Data

Test Score (Accuracy): 1.0
Classification Report:

	precision	recall	f1-score	support
0	1.00	1.00	1.00	9
1	1.00	1.00	1.00	12
2	1.00	1.00	1.00	13
3	1.00	1.00	1.00	13
4	1.00	1.00	1.00	15
5	1.00	1.00	1.00	9
6	1.00	1.00	1.00	6
7	1.00	1.00	1.00	8
8	1.00	1.00	1.00	11
9	1.00	1.00	1.00	13
10	1.00	1.00	1.00	7
11	1.00	1.00	1.00	12
12	1.00	1.00	1.00	4
13	1.00	1.00	1.00	11
14	1.00	1.00	1.00	10
15	1.00	1.00	1.00	7
16	1.00	1.00	1.00	9
17	1.00	1.00	1.00	12
18	1.00	1.00	1.00	12
19	1.00	1.00	1.00	10
20	1.00	1.00	1.00	8
21	1.00	1.00	1.00	9
accuracy			1.00	220
macro avg	1.00	1.00	1.00	220
weighted avg	1.00	1.00	1.00	220

Figure 21: Detailed Model Performance Analysis

Final Model Assessment:

The Proposed Model (RF) classifier consistently outperformed alternative models in both training and validation stages. With an F1-score of 1.00, and 100% accuracy confirmed via confusion matrix and cross-validation, the model demonstrated exceptional generalization capabilities.

The overall performance was achieved through a robust pipeline integrating MinMaxScaler, Label Encoding, SMOTE for balancing, and GridSearchCV with 5-fold cross-validation — ensuring resilience across diverse agricultural conditions.

5. Conclusion

This study proposed a machine learning-based framework for crop recommendation in Egypt, leveraging environmental and soil data to support sustainable agricultural decision-making. Among the evaluated models—Decision Tree, Support Vector Machine, Linear Regression, and proposed model (RF) —the Random Forest classifier consistently outperformed all others in terms of accuracy, generalization, and robustness.

The final model achieved:

- Accuracy: 100%
- Precision: 1.00
- Recall: 1.00
- F1-score: 1.00
- Cross-validation accuracy: 99.38%

These results demonstrate the model's capacity to accurately classify crops based on complex, real-world environmental inputs. The use of techniques such as SMOTE, feature engineering, and GridSearchCV contributed to enhancing both performance and generalizability.

Practical Implication:

The model can serve as a decision-support system for farmers, agricultural engineers, and policymakers by providing accurate crop recommendations tailored to local environmental conditions.

Future Work

To further enhance model performance and applicability, the following directions are proposed:

- **Integration with Real-Time IoT Sensor Data:**

  Incorporating live environmental readings will allow the system to adapt to changing field conditions and support dynamic crop selection.

- **Extension to Multiseasonal and Multiregional Data:**
  Expanding the dataset to include multiple seasons and regional crop patterns can improve generalizability across Egypt's agro-ecological zones.

- **Adoption of Deep Learning Models:**

  Exploring advanced architectures such as Recurrent Neural Networks (RNNs) or Transformers may capture temporal dependencies in crop cycles.

6. References

[1] *"Agriculture in Egypt,"* Mordor Intelligence, 2025. [Online]. Available:

https://www.mordorintelligence.com/industry-reports/agriculture-in-egypt

[2] A. M. Whitbread *et al*., "Designing climate resilient agricultural systems with some examples from India," in *Proc. 5th Int. Agron. Congr.*, Hyderabad, India, Nov. 2021. [Online]. Available: https://oar.icrisat.org/11931/

[3, 7] J. R. Quinlan, "Overview of use of decision tree algorithms in machine learning," in *Proc. IEEE Conf. Mach. Learn.*, 2011. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/5991826

[4, 8] R. Shen and B. Zhang, "Study on regression model in machine learning," in Proceedings of the 6th International Conference on Industrial Design (IFID 2018), 2018, pp. 01033. [Online]. Available: https://www.matec-conferences.org/articles/matecconf/abs/2018/35/matecconf_ifid2018_01033/matecconf_ifid2018_01033.html

[5, 9] Harvard IACS, "Support Vector Machines (SVM) Lecture 20," 2018. [Online]. Available: https://harvard-iacs.github.io/2018-CS109A/lectures/lecture-20/presentation/lecture20_svm.pdf.

[6, 10] H. A. Salman, A. Kalakech, and A. Steiti, "Random Forest Algorithm Overview," Babylonian Journal of Machine Learning, vol. 2024, pp. 69–79, Jun. 2024. [Online]. Available: https://journals.mesopotamian.press/index.php/BJML/article/view/417.

[12, 13] N. Shahat, W. Younes, and M. El-Shafie, "A hybrid model of crop prediction using ensemble learning and weather forecasting data," *Computers and Electronics in Agriculture*, vol. 163, p. 104859, Sep. 2019. [Online]. Available: https://www.sciencedirect.com/science/article/abs/pii/S0304380019302145

[14, 15] R. Deepa and B. Ramesh, "Epileptic seizure detection using deep learning techniques: A review," *Semantics Scholar*, [Online]. Available: https://www.semanticscholar.org/paper/21f4187c8ba527d822b2c24c75f5754967eec14a.

[16, 17] A. Ghosh, B. Dey, and D. Chakraborty, "Agricultural data prediction using data mining techniques: A survey," in *Information Systems Design and Intelligent Applications*, S. C. Satapathy, S. Sahidujjaman, and P. S. Avadhani, Eds., vol. 247. Springer, Cham, 2014, pp. 579–589. doi: 10.1007/978-3-642-40669-0_33

[18] H. A. Ghribi, A. Bouzouita, and M. Limam, "Crop Recommendation System Using Data Mining Techniques," in *Artificial Intelligence Applications and Innovations (AIAI 2021)*, vol. 584, L. Iliadis, I. Maglogiannis, and H. Papadopoulos, Eds. Springer, Cham, 2021, pp. 25–36. doi: 10.1007/978-3-030-65900-4_3

[19] A. Kuhn and A. G. Henriques, *Feature Engineering for Machine Learning: Principles and Techniques for Data Scientists*. O'Reilly Media, 2020. [Online]. Available: https://books.google.com.eg/books?id=sthSDwAAQBAJ

[20, 21] M. I. Soliman, M. A. M. Salem, and M. A. El-Bakry, "Smart farming decision support system using machine learning," in *Proc. 2021 International Conference on Advanced Intelligent Systems and Informatics (AISI)*, Cairo, Egypt, 2021, pp. 331–340. doi: 10.1109/AISI53016.2021.9649207

[22, 23] D. Berrar, "Cross-validation," *Encyclopedia of Bioinformatics and Computational Biology*, 2nd ed., Elsevier, 2018, pp. 542–545. [Online]. Available: https://dberrar.github.io/papers/Berrar_EBCB_2nd_edition_Cross-validation_preprint.pdf

[24] R. Sharma, S. K. Dubey, and S. Jain, "A comprehensive review on machine learning for crop yield prediction," *Engineering Science and Technology, an International Journal*, vol. 35, p. 101072, Jun. 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0198971522000898