Optimizing Crop Recommendations Using Advanced Machine Learning Techniques

Marwa M. Eid
Faculty of Artificial Intelligence
Delta University for Science
& Technology Mansoura 35111, Egypt

Jadara University Research Center
Jadara University
Jordan
mmm@ieee.org

***Abstract*—This work aims to find the best crop recommendation system where we employ several machine learning algorithms using a dataset containing soil and environmental factors. Some of the variables in the dataset are Nitrogen; Phosphorus; Potassium; Temperature; Humidity; pH value; and Rainfall. We also compare the models, including Logistic Regression, Support Vector Machine (SVM) classifier, K-Nearest Neighbors (KNN), Decision Tree, and Extra Trees, to determine the most appropriate crops according to the growing conditions. Random Forest is used for feature selection to determine which features have the most influence on the models. According to the evaluation results of the various models, the best one, Extra Trees, has a test accuracy of They achieve a test F1 score of and the test precision is and a test recall of. On the other hand, the other models show slightly lower accuracy with test accuracy of for both Logistic Regression and Support Vector Machine models. KNN achieving The Random forest algorithm was found to be the most accurate model.**

***Index Terms*—Crop recommendation , machine learning , feature engineering , model evaluation , agricultural productivity**

I. Introduction

Crop recommendation is an important focal area in modern agriculture that is central to the achievement of high yields, proper utilization of resources, and promotion of sustainable practices. Considering the crops that are suitable to be grown on the given type of soil and the environment determines the overall yield in agriculture and how effectively the resources are utilized. This study, however, seeks to overcome this challenge by adopting a broad dataset that comprises factors like Nitrogen, Phosphorus, Potassium, Temperature, Humidity, pH Value, and Rainfall. These variables are very essential to determine the level of fertility of soils as well as determine the environmental conditions that are conducive to the growth of various crops.

To get accurate crop recommendations, we perform different types of Machine Learning Algorithms such as Logistic Regression, SVM, KNN, Decision Trees, and Extra Trees. This is done to establish the efficiency of the models in identifying the appropriate crops relevant to specific environments. It also involves proper data pre-processing methods to increase the accuracy of the models tested on the dataset. Further, the result of converting features with StandardScaler, and encoding with OrdinalEncoder and LabelEncoder guarantees proper preparation of the data for analysis.

In recent years, optimization and metaheuristic algorithms have become increasingly relevant in the broader field of agricultural analytics and computational intelligence [1] [2]. Although not directly implemented in this research, such methods—like Genetic Algorithms, Particle Swarm Optimization, Grey Wolf Optimizer, and Whale Optimization Algorithm—have demonstrated strong capabilities in addressing complex optimization challenges across agriculture [3]. These algorithms are inspired by natural phenomena such as evolution, animal behavior, and swarm intelligence, enabling them to find optimal or near-optimal solutions where traditional methods may fail. In the context of crop recommendation, optimization approaches can be used to fine-tune model parameters, improve feature selection, and enhance decision-making by efficiently exploring large solution spaces [4]. Their adaptability and robustness make them valuable for solving multidimensional agricultural problems, such as yield maximization, irrigation scheduling, and nutrient optimization, thereby complementing data-driven and machine learning-based systems [5].

This research work aims to take advantage of data-driven methods to achieve accurate crop advisement, thus promoting improved crop management in agriculture. In the context of this study, it is planned to enhance the efficiency of resource utilization, increase crop productivity, and advance sustainable agricultural practices by incorporating refined machine learning algorithms and algorithms for data preprocessing. This paper offers a comprehensive analysis of model performance and recommendations on the ability of data-driven approaches to improve agricultural yield and hence the overall welfare of agricultural businesses and farmers.

II. Literature Review

This literature review synthesizes several related studies that explain how machine learning models and data preprocessing

approaches have been used in different disciplines. Therefore, machine learning in crop recommendation systems is vast, especially in data preprocessing including feature engineering leading to high accurate Extra Trees model [6]. Further, while proposing ensemble learning models for sunshine duration prediction, trials have shown feature selection and optimization methods such as Grey Wolf and Stochastic Fractal Search for enhanced performance [7].

For disease diagnosis, various machine learning models have been used in the medical field especially the convolutional neural networks with an understanding that diagnosis requires proper feature balancing as well as classifier optimization as embraced by [8] in the diagnosis of COVID-19. Recent literature on class imbalance demonstrates that the traditional measures of model performance can fall short, especially if datasets are imbalanced, while the use of measures such as the Matthews Correlation Coefficient can offer a clearer picture of a model's performance and functioning [9].

Decision tree and random forest methods have been found to fit into the ensemble family of methods used in clinical research for the prediction of an outcome that exhibits good performance with small sample sizes. These models are useful as dependable decision-aiding systems within the healthcare domain [10]. Moreover, the use of machine learning frameworks presented in detecting the risk of Autism Spectrum Disorder using logistic regression analysis and feature analysis also displays how diagnostic tools can be of assistance in early detection to cut the costs of high-impact diseases [11]. Recent advancements in machine learning and optimization techniques have shown significant potential in agricultural applications, particularly in crop recommendation systems. Network intrusion detection methods utilizing feature selection and hybrid metaheuristic optimization [12] demonstrate the effectiveness of combining traditional machine learning approaches with modern optimization algorithms. Similarly, wind power prediction models based on machine learning and deep learning [13] showcase the versatility of these techniques across different domains, providing valuable insights into how predictive models can be adapted for agricultural forecasting. The integration of image processing and artificial intelligence for early diagnosis applications [14] further illustrates the potential for visual-based agricultural monitoring systems, while optimized ensemble algorithms for parameter prediction [15] offer robust frameworks that can be applied to crop yield estimation and recommendation systems.

III. Material and Methods

As Shown in figure 1 The following are the main steps that have been proposed for the development of a crop recommendation system: First, data collection is conducted on soil nutrient characteristics – Nitrogen, Phosphorus, and Potassium levels, environment characteristics – Temperature, Humidity, pH, Rainfall, and crop characteristics. Missing values are handled by imputation, and outliers are removed by using the IQR method; besides, skewed features are transformed logarithmically, using root functions, and power functions. Features are then normalized by standardization using the Standard-Scaler while the categorical data is encoded with the help of the Ordinal-Encoder and Label-Encoder. Feature engineering creates new variables like the ratio of Nitrogen, Potassium, and Phosphorus (NP_Ratio, NK_Ratio, PK_Ratio) and environmental indices which include temperature humidity index and rainfall humidity index (Temp_Humidity_Index, Rainfall_Humidity_Index), and mean NPK (NPK_Average), and categorizing the soil ph level. Feature selection is done using an algorithm called Random Forest in the context of crop prediction based on the most influential features. The selected features are then applied to create and develop machine learning models where amongst them are Logistic Regression, SVM, KNN, Decision Tree, and Extra Trees. For each model, the number of correct instances, accuracy, F1 measure, precision, and recall are reported. Extra Trees gave better results against other models with an accuracy of 98. 86% accuracy and an F1 score of 98. 86%, so it is considered to be the optimal option for the crop recommendation system.

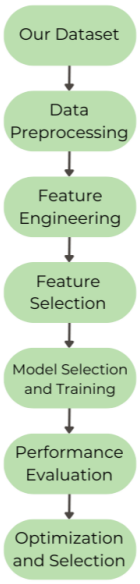

Fig. 1. Proposed Framework

A. Dataset Description

The dataset adopted for the current research comprises relevant information on agricultural and environmental factors that are essential for crop recommendations. Each feature provides valuable information about soil and climatic conditions affecting crop growth and yield:

- **Nitrogen (N):** Estimates the quantity of nitrogen in the soil, which is an essential nutrient crucial for plant growth and development.
- **Phosphorus (P):** Represents the proportion of phosphorus in the soil, important for energy transfer and root formation.

- **Potassium (K):** Indicates the proportion of potassium, an essential nutrient responsible for water regulation and overall plant growth.
- **Temperature:** The ambient temperature in degrees Celsius around the plant, significantly influencing metabolic and growth processes.
- **Humidity:** Represents the percentage of moisture in the air, affecting plant respiration rates, transpiration rates, and water availability.
- **pH_Value:** The soil pH, indicating soil acidity or alkalinity, which significantly affects nutrient solubility and availability to plants.
- **Rainfall:** The total rainfall measured in millimeters, determining water availability and significantly influencing plant hydration and growth.

B. Data Preprocessing

Data preprocessing is the first step in the data science process where input data is cleaned, transformed, and formatted to fit the downstream machine learning processes. This process makes sure that the data they obtain is valid, reliable, and ready for model building and assessment. The preprocessing phase encompasses several critical steps: missing value management, feature scaling, feature encoding, dealing with skewed data, and feature creation and feature reduction. Every stage includes particular procedures and mathematical operations that are aimed at optimizing the quality of data and boosting the efficacy of the model.

C. Descriptive Analysis of the Data

The Descriptive Analytics Section is the initial investigation of the given dataset, utilizing tabular analysis and a set of visual tools to identify the patterns and connections inherent in the agricultural data. This section focuses on the relationship between these core quantitative variables Nitrogen, Phosphorus, Potassium, Temperature, Humidity, pH Value, and Rainfall, and presents the correlation matrix with a heatmap to depict the values involved and their correlation or dependency. The pair plot further categorizes these relationships by presenting scatter plots and histograms that depict interactions and distributions within various categories. Conversely, the box plot provides an unobstructed view of the distributional nature of these variables with focal points on variation and outlier presence. Thus, using these data visualizations, the section provides invaluable information to explain the structure and variability of the data under analysis. They serve as useful instruments for further analysis of the dataset and potential decision-making based on it while giving an insight into what the set consists of. In other words, the Descriptive Analytics Section provides the big picture for the data and unravels the patterns that agricultural decision-makers ought to consider in their decisions.

This is well depicted in the heatmap illustrated in Figure 2 where the agricultural variables include Nitrogen, Phosphorus, Potassium, Temperature, Humidity, pH Value, and Rainfall. The heatmap incorporates the use of color where red represents a high positive correlation, green represents a low correlation, and blue represents a high negative correlation among other colors.

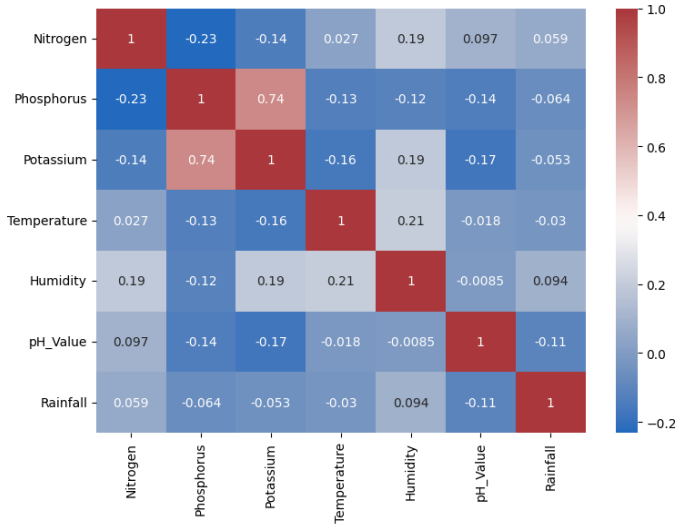

Fig. 2. Heatmap depicting correlation among agricultural variables, illustrating both correlation strength and direction.

In Figure 3, there is a box plot depicting Nitrogen, Phosphorus, Potassium, Temperature, Hardness, pH Value, and Rainfall. The plot also indicates the IQR, the median, and the outliers for each of the variables, which makes it easier to assess their distribution and variability.

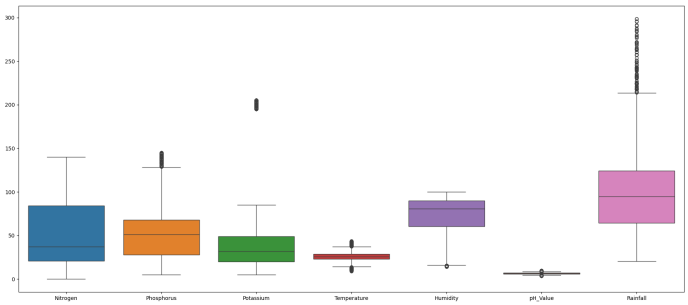

Fig. 3. features a box plot summarizing the distribution and variability of variables like Nitrogen, Phosphorus, Potassium, Temperature, Hardness, pH Value, and Rainfall.

These are essential in getting a general feel of the data set and identifying features and trends that are significant enough to warrant further analysis.

D. Machine Learning Models

In this research, we use the classification models to introduce the appropriate crop. Classification algorithms such as Logistic Regression, SV Machine, KNN, Decision Tree, and Extra Trees are the main tools used to reveal hidden patterns in the agricultural dataset. To assess the efficiency of each of the models presented, we use indicators that measure the effectiveness of a predictive model, including accuracy and

F1, precision, and recall. Such measures make it possible to evaluate the performance of each classification model in providing accurate crop recommendations and choosing the most suitable approach [10].

1) Logistic Regression: Logistic regression was originally developed for statistical data analysis, where it is used to determine the relationships between one or more independent variables and a dependent variable. In predictive problems, the input dataset generally consists of two possible values for the dependent variable (target class). The goal is to model the relationship between the independent variables and the target class using a logic function based on probabilities. Further details on classification using multinomial logistic regression are available in [11-12].

2) Support Vector Machine (SVM): SVM is one of the most flexible models since it can be used for classification, regression, and outliers [13-14]. They are beneficial for classification tasks, especially for large datasets. The SVM classification technique operates by mapping the features that are not linearly separable into a higher dimensional space using a kernel function to fit a hyperplane into the classes. This is made possible by the kernel trick where kernels such as the linear, polynomial, or Gaussian RBF are used and the added features' computational complexity is reduced. The distance between classes is defined by specific instances in the dataset called support vectors while the kernel parameters define the margin as well as the acceptable margin violations. Like other classifiers, even if SVM was originally designed for binary classification, it is possible to use it for multinomial classification tasks as well.

$$\sum_{i=1}^{n} \alpha_i y_i K(x_i, x) + b \tag{1}$$

where $K(x_i, x)$ is the kernel function, α_i and y_i are the Lagrange multipliers, and b is a bias term.

3) K-Nearest Neighbors (KNN): The approach described in the current work employs a wrapper approach for feature selection, where the K-nearest neighbor or the KNN classifier is a monitored learning formula [15]. In KNN, every example is classified into a particular class label depending on the classification done by most of its K-neighbors. In the case of KNN, to decide on the class of the unknown circumstances, it is necessary to use training circumstances rather than the structure versions. KNN is employed for categorization problems to evaluate the quality of the chosen part of functions. The Euclidean range (EucD) between functions of the training information and features of the screening information is computed to determine the nearest K next-door neighbors.

4) Decision Tree: Speaking of its nature, a Decision Tree (DT) is a tree-like model of decision making, where the leaves are associated with the outcome labels and the branches are associated with sets of input features leading to these or those outcomes [16].

As in the binary decision tree, the data (the parent node) is split into two subgroups, each called a child node, by selecting the feature with the highest split criterion. The two obtained subsets become new parent nodes again, which are divided into two child nodes each in turn. This process of splitting cases into two categories continues until all the observations have been grouped. The algorithm is nonparametric; it does not need to assume any distributional form of the predictor or the target variable.

5) Extra Trees: Extra-Trees is a classifier of agricultural data and is a random decision tree of one or more random trees designed for each subset of the data [17-18]. This technique aids in preventing overfitting and enhances predictability by averaging the findings. Shown below are some of the agricultural applications of the ExtraTree classifier; predicting crop yields, and determining the health of the soil.

E. Performance Metrics

Performance indicators are crucial for evaluating the effectiveness of machine learning models and algorithms. These metrics provide quantitative assessments of how well a model performs its intended task [19]. Key measures include accuracy, precision, recall, and the F1 score. Accuracy indicates the overall correctness of the model, while precision and recall assess the classifier's effectiveness in identifying relevant instances. The F1 score balances precision and recall to provide a single metric that reflects their combined performance. Understanding these metrics is essential for assessing model performance and guiding improvements in predictive accuracy.

TABLE I
CRITERIA FOR EVALUATING CLASSIFICATION RESULTS

Metric	**Formula**
Accuracy	$\frac{TP+TN}{TP+TN+FP+FN}$
Precision	$\frac{TP}{TP+FP}$
Recall	$\frac{TP}{TP+FN}$
F1 Score	$2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$

IV. EXPERIMENTAL RESULTS

This section discusses the performance gained by the proposed crop recommendation system using different machine learning models and the data preprocessing steps used to enhance the outcome.

A. Model Performance

Among the five machine learning models tested for crop recommendation—Logistic Regression, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Decision Tree, and Extra Trees models—the performance was evaluated based on accuracy, F1 score, precision, and recall. The Extra Trees model exhibited the highest performance, achieving an accuracy of 0.9886, an F1 score of 0.9886, precision of 0.9892, and recall of 0.9886. Other models, such as Logistic Regression, SVM, KNN, and Decision Tree, also demonstrated

good performance with slightly lower accuracy and F1 scores. Detailed model performance is provided in Table II:

TABLE II
CRITERIA FOR MODEL PERFORMANCE

Metric	Logistic Regression	SVM	KNN	Decision Tree	Extra Trees
Accuracy	0.979	0.979	0.984	0.981	0.988
F1 Score	0.979	0.979	0.984	0.981	0.988
Precision	0.981	0.981	0.984	0.982	0.989
Recall	0.979	0.980	0.984	0.981	0.988

V. CONCLUSION

The results of this study revealed the machine learning models' success in reconsidering crop recommendations using detailed data obtained from soil and environment. This proposed the application of data preprocessing methods such as improve the performance of the models. As a result, the model called Extra Trees has been chosen as the most accurate with the highest values of accuracy, F1, precision, and recall.

The results support the hypothesis that there is a need to apply several data preprocessing stages, as well as select the most relevant features for improving the performance of machine learning models in agriculture. The confusion matrices as the continuation of the visualizations also helped in understanding the strengths and weaknesses of the models for classification.

Finally, this research shows the possibilities of combining data analysis to advance agricultural practices and make practical crop suggestions to increase yield rates, efficiency, and sustainability. The enhancement of featured data by utilizing other data types, such as weather data or remote sensing imagery, could be addressed in the next steps of research to improve the proposed analytics and adapt them for practical use in the agricultural environment.

REFERENCES

[1] El-Sayed M El-Kenawy, Abdelhameed Ibrahim, Amel Ali Alhussan, Doaa Sami Khafaga, Ayman EM Ahmed, and Marwa M Eid. Smart city electricity load forecasting using greylag goose optimization-enhanced time series analysis. *Arabian Journal for Science and Engineering*, pages 1–19, 2025.

[2] El-Sayed M Elkenawy, Amel Ali Alhussan, Doaa Sami Khafaga, Zahraa Tarek, and Ahmed M Elshewey. Greylag goose optimization and multilayer perceptron for enhancing lung cancer classification. *Scientific Reports*, 14(1):23784, 2024.

[3] Ahmed El-Sayed Saqr, Mohamed S Saraya, and El-Sayed M El-Kenawy. Enhancing co2 emissions prediction for electric vehicles using greylag goose optimization and machine learning. *Scientific Reports*, 15(1):16612, 2025.

[4] El-Sayed M Elkenawy, Amel Ali Alhussan, Marwa M Eid, and Abdelhameed Ibrahim. Rainfall classification and forecasting based on a novel voting adaptive dynamic optimization algorithm. *Frontiers in Environmental Science*, 12:1417664, 2024.

[5] Amel Ali Alhussan, El-Sayed M. El-kenawy, Doaa Sami Khafaga, Amal H. Alharbi, and Marwa M. Eid. Groundwater resource prediction and management using comment feedback optimization algorithm for deep learning. *IEEE Access*, pages 1–1, 2025.

[6] T. Ayoub Shaikh, T. Rasool, and F. Rasheed Lone. Towards leveraging the role of machine learning and artificial intelligence in precision agriculture and smart farming. *Computers and Electronics in Agriculture*, 198:107119, 2022.

[7] Y. Akkem, S. K. Biswas, and A. Varanasi. Smart farming using artificial intelligence: A review. *Engineering Applications of Artificial Intelligence*, 120:105899, 2023.

[8] Modelling speech emotion recognition using logistic regression and decision trees. *International Journal of Speech Technology*. Retrieved October 15, 2024.

[9] A. De Caigny, K. Coussement, and K. W. De Bock. A new hybrid classification algorithm for customer churn prediction based on logistic regression and decision trees. *European Journal of Operational Research*, 269(2):760–772, 2018.

[10] A. Correa Bahnsen, D. Aouada, A. Stojanovic, and B. Ottersten. Feature engineering strategies for credit card fraud detection. *Expert Systems with Applications*, 51:134–142, 2016.

[11] M. Christ, N. Braun, J. Neuffer, and A. W. Kempa-Liehr. Time series feature extraction on basis of scalable hypothesis tests (tsfresh – a python package). *Neurocomputing*, 307:72–77, 2018.

[12] R. Alkanhel, E.-S. El-kenawy, A. Abdelhamid, A. Ibrahim, M. Alohali, M. Abotaleb, and D. Khafaga. Network intrusion detection based on feature selection and hybrid metaheuristic optimization. *Computers, Materials & Continua*, 74(2):2677–2693, 2022.

[13] Z. Tarek, M. Shams, A. Elshewey, E.-S. El-kenawy, A. Ibrahim, A. Abdelhamid, and M. El-dosuky. Wind power prediction based on machine learning and deep learning models. *Computers, Materials & Continua*, 74(1):715–732, 2022.

[14] E. S. Mira, A. M. S. Sapri, R. F. Aljehanı, B. S. Jambı, T. Bashir, E.-S. M. El-Kenawy, and M. Saber. Early diagnosis of oral cancer using image processing and artificial intelligence. *Fusion: Practice and Applications*, (1):293–308, 2024.

[15] E.-S. El-kenawy, A. Ibrahim, S. Mirjalili, Y.-D. Zhang, S. Elnazer, and R. Zaki. Optimized ensemble algorithm for predicting metamaterial antenna parameters. *Computers, Materials & Continua*, 71(3):4989–5003, 2022.