Report: Crop Recommendation System Using Random Forest Model

Abstract

This report presents a comprehensive analysis of multiple machine learning models developed for crop classification based on environmental and soil conditions. Building upon previous research, I evaluated seven distinct models: Random Forest, Logistic Regression, XGBoost, Gaussian Naive Bayes, K-Nearest Neighbors, Decision Tree, and Neural Network. The models were trained on a dataset comprising environmental features (rainfall, humidity, temperature, pH) and soil nutrients (Nitrogen, Phosphorus, Potassium), and were trained to classify 22 different crop types. Statistical analysis through ANOVA validated the significance of environmental factors in crop differentiation. After removing outliers, the models achieved accuracies ranging from 96.89% to 98.87%, with Random Forest emerging as the top performer. The insights gained from this model can inform decision-making for farmers, enabling them to optimize crop selection based on localized conditions, ultimately improving agricultural productivity and sustainability.

1. Introduction

In agricultural practices, optimizing crop selection is a critical factor for ensuring food security, efficient use of resources, and high yields. Crop recommendation systems, powered by machine learning algorithms, provide insights into which crops are best suited to specific environmental conditions. The data used in this analysis includes several features that are known to influence crop productivity, including **rainfall, humidity, temperature, Nitrogen (N), Phosphorus (P), Potassium (K), and pH** levels of the soil. The goal of this report is to analyze a Random Forest model that predicts the best crop to plant under given environmental conditions, and to interpret the results from feature importance and confusion matrix outputs. This dataset was built by augmenting datasets of rainfall, climate and fertilizer data that are available for India.

The model focuses on classifying 22 different crops based on the mentioned environmental factors. These crops include rice, maize, chickpea, kidney beans, pigeon peas, moth beans, mungbean, blackgram, lentil, pomegranate, banana, mango, grapes, watermelon, muskmelon, apple, orange, papaya, coconut, cotton, jute, and coffee. The Random Forest model was selected for its high accuracy and ability to handle large datasets, making it particularly suitable for agricultural prediction systems.

2. Evaluation Metrics

In this study, I employed several metrics to comprehensively evaluate model performance. The R-squared (R^2) score measures the proportion of variance in the dependent variable explained by the independent variables, with values closer to 1 indicating better fit. Mean Squared Error (MSE) quantifies prediction accuracy by measuring the average squared difference between predicted and actual values, where lower values indicate better performance. For classification performance, I utilized F1-scores, which represent the harmonic mean of precision and recall, providing a balanced measure of model accuracy. Recall metrics were used to assess each model's ability to correctly identify all instances of each crop class. These metrics collectively provide a robust framework for comparing model performance and reliability in crop prediction tasks.

3. Statistical Analysis (ANOVA)

Statistical analysis like ANOVA helps in understanding relationships between variables, the main predictive task focuses on classifying crop types based on various factors. The ultimate aim is to inform decision-making for optimal crop selection.

Prior to model development, I conducted an Analysis of Variance (ANOVA) to validate the significance of environmental factors in differentiating crop types. The analysis revealed compelling evidence for the importance of all environmental factors. Temperature exhibited a substantial F-statistic of 127.16 with a p-value of 2.87e-312, indicating strong statistical significance in differentiating between crop types. Humidity demonstrated exceptional significance with an F-statistic of 2315.85 and a p-value of 0.0, establishing it as a crucial factor in crop differentiation. The soil pH analysis yielded an F-statistic of 74.08 and a p-value of 5.70e-208, confirming its significant influence on crop selection. Rainfall analysis produced an F-statistic of 592.37 with a p-value of 0.0, demonstrating its critical role in determining crop suitability.

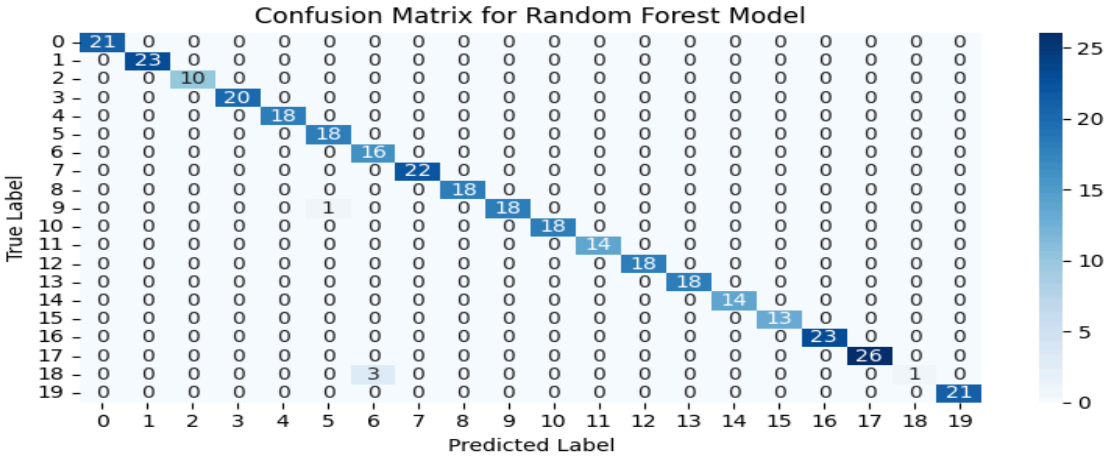
4. Model Performance

4.1 Random Forest Model

The Random Forest model demonstrated exceptional performance, achieving an accuracy of 98.87%. With an R-squared score of 0.9623, the model explained 96.23% of the variance in crop type predictions. The mean squared error of 1.27 indicated high prediction accuracy. The model showed remarkable consistency across all crop classes, with most achieving perfect classification scores. The confusion matrix revealed minimal misclassifications, primarily concentrated in classes with smaller sample sizes.

Figure 1: Confusion Matrix for Random Forest Model

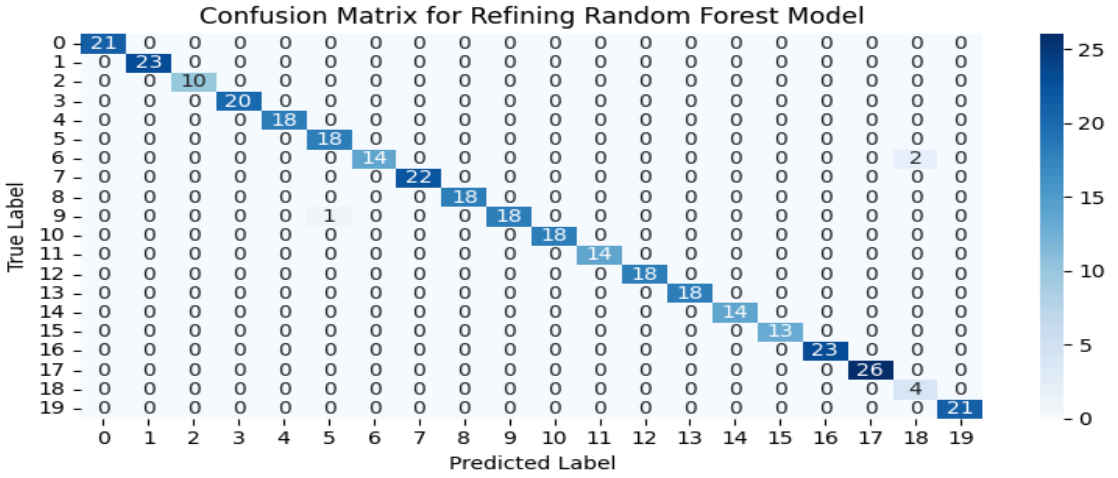
The confusion matrix shows a high level of agreement between predicted and actual labels, with most diagonal elements showing perfect classification. Misclassifications occur in two notable instances: class 6 shows lower precision (0.84) due to other classes being incorrectly classified as class 6, and class 18 shows poor recall (0.25) with 3 out of 4 instances being misclassified as class 6. Despite these specific cases, the model maintains strong overall performance with 98.87% accuracy.



Refining Random Forest

Through feature selection using SelectFromModel and hyperparameter tuning via GridSearchCV, the Random Forest model's performance was further improved. The refined model achieved an accuracy of 99.15%, with an R-squared score of 0.9744 and a reduced mean squared error of 0.8588. This optimization particularly enhanced the model's performance for challenging classes, with only minor misclassifications occurring in class 6 (2 instances) and class 9 (1 instance). The improved confusion matrix demonstrates near-perfect classification across most crop categories, suggesting that the feature selection and hyperparameter tuning effectively reduced model complexity while maintaining high predictive accuracy.

Figure 2: Confusion Matrix for Refining Random Forest Model



4.2 Logistic Regression

The Logistic Regression model achieved an accuracy of 96.89%, with an R-squared value of 0.908 indicating strong predictive power. The mean squared error of 3.07 suggested reasonable prediction accuracy. The model demonstrated particularly strong performance in classifying crops in categories 0, 3, 10, 13, and 17, achieving perfect precision and recall scores. Some variations were observed in classes 6 and 18, though they maintained acceptable F1-scores of 0.85 and 0.33 respectively.

4.3 XGBoost

The XGBoost model achieved an impressive accuracy of 98.59%, with an R-squared score of 0.95798 indicating excellent predictive capability. The mean squared error of 1.41243 demonstrated strong prediction accuracy. The model exhibited consistent performance across most crop classes, with perfect precision, recall, and F1-scores for the majority of crops. Minor performance variations were observed in classes 6 and 18, with precision and recall scores slightly lower than other classes.

4.4 Gaussian Naive Bayes

The Gaussian Naive Bayes classifier achieved an accuracy of 98.59% with an R-squared score of 0.9616. The mean squared error of 1.2909 indicated strong predictive accuracy. The model demonstrated robust performance across different crop classes, with macro-average precision, recall, and F1-scores of 0.97, 0.98, and 0.97 respectively. The confusion matrix showed strong diagonal dominance, indicating accurate predictions across most crop categories.

4.5 K-Nearest Neighbors

The K-Nearest Neighbors model achieved an accuracy of 97.46%, with an R-squared value of 0.9359 demonstrating strong predictive capability. The mean squared error of 2.1554 indicated good prediction accuracy. The model showed consistent performance across crop classes, with macro-average precision, recall, and F1-scores of 0.96, 0.95, and 0.96 respectively. The confusion matrix revealed strong classification performance with minimal misclassifications.

4.6 Decision Tree

The Decision Tree classifier achieved an accuracy of 97.18% with an R-squared score of 0.890. The mean squared error of 3.686 indicated reasonable prediction accuracy. The model demonstrated strong performance across most classes, with macro-average precision, recall, and F1-scores of 0.94. Some performance variations were observed in class 18, which showed lower precision, recall, and F1-scores of 0.25, indicating potential areas for improvement.

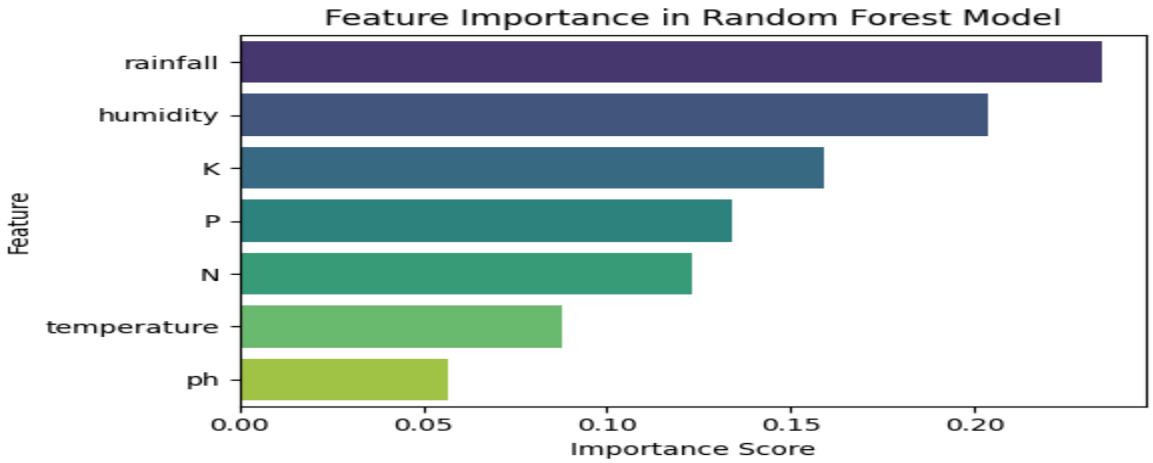
4.7 Neural Network

The Neural Network model achieved an accuracy of 96.89% with an R-squared score of 0.9040. The mean squared error of 3.2372 suggested good prediction accuracy. The model showed strong performance across most classes, with macro-average precision, recall, and F1-scores of 0.92, 0.93, and 0.93 respectively. Notable challenges were observed in predicting class 18, where the model struggled to achieve accurate classifications.

5. Feature Importance

One of the most valuable aspects of the Random Forest model is its ability to assess the importance of each feature in making predictions. The feature importance plot (Figure 3) provides insights into which environmental factors have the greatest influence on crop selection.

Figure 3: Feature Importance in Random Forest Model



From the plot, it is clear that **rainfall** and **humidity** are the most influential features, with importance scores of 0.235 and 0.204, respectively. This is followed by **Potassium (K)** with a score of 0.159, **Phosphorus (P)** (0.134), **Nitrogen (N)**(0.123), **temperature** (0.088), and **pH** (0.057). These rankings

provide a basis for understanding the factors that most affect crop growth and how these factors should be managed.

Let's break down each feature and its influence on crop selection.

Feature Impact on Crop Selection

4.1 Rainfall (0.235) Rainfall is the most critical factor influencing crop yield. Adequate water supply ensures that crops can complete their growth cycles, especially for water-intensive crops like rice, maize, and bananas. Rainfall not only affects soil moisture but also regulates temperature and humidity, which are crucial for seed germination, root development, and nutrient absorption. For example, crops like **rice**, which thrives in flooded conditions, require high rainfall, while **coconut** can tolerate less rainfall but still needs consistent water availability.

4.2 Humidity (0.204) Humidity plays a significant role in the photosynthesis process and affects how crops transpire. High humidity levels benefit tropical fruits like **mango**, **banana**, and **papaya**, which require a moist environment for proper development. Conversely, crops like **cotton** and **jute** thrive in regions with moderate humidity levels, where excessive moisture could damage fiber quality. Thus, humidity serves as a secondary factor to rainfall in determining the suitability of a crop for a particular region.

4.3 Potassium (K) (0.159) Potassium is essential for various plant physiological processes, including enzyme activation, photosynthesis, and water regulation. Its importance in agriculture cannot be understated, particularly for crops like **grapes** and **watermelon**, where potassium influences fruit size and sugar content. Adequate potassium levels promote stronger root systems, increase drought resistance, and enhance disease resistance, making it a crucial factor for crop health.

4.4 Phosphorus (P) (0.134) Phosphorus is a key element in energy transfer and root development. Crops such as **pigeon peas**, **lentil**, and **mungbean**, which rely on strong root systems to extract nutrients from the soil, benefit significantly from higher phosphorus levels. Phosphorus is also crucial during the early growth stages of crops, contributing to seed formation and overall plant vigor. Efficient phosphorus management is essential to prevent nutrient deficiency, which can drastically reduce crop yields.

4.5 Nitrogen (N) (0.123) Nitrogen is a primary component of chlorophyll, the compound that plants use for photosynthesis. It is particularly important for leafy crops like **maize** and **rice**, where high nitrogen levels result in better leaf area development and higher yields. Nitrogen also promotes protein formation, which is essential for **legumes** like **chickpeas** and **kidney beans**. While nitrogen is crucial, excessive amounts can lead to environmental damage, so managing nitrogen inputs is vital for sustainable agriculture.

4.6 Temperature (0.088) Temperature affects the rate of growth and development for all crops. Warm-season crops like **maize**, **cotton**, and **coffee** thrive at higher temperatures, whereas temperate crops such as **apple** and **orange** require cooler climates. Temperature extremes—either too high or too low—can cause stress to crops, reducing their productivity. The moderate importance of temperature in this model highlights its role in conjunction with rainfall and humidity in defining the overall growing environment for crops.

4.7 pH (0.057) Soil pH affects the availability of nutrients to plants. Most crops prefer neutral to slightly acidic soils (pH 6 to 7), but some crops, like **coffee** and **orange**, can tolerate slightly more acidic soils. pH influences the microbial activity in the soil, which is critical for nutrient cycling and availability. While

pH has a lower importance score in this model, it remains a crucial factor, particularly for crops sensitive to nutrient deficiencies caused by inappropriate pH levels.

5. Implications for Farmers and Stakeholders

The insights provided by this Random Forest model are invaluable for farmers and stakeholders involved in agricultural planning. By understanding the factors that most influence crop success, farmers can make informed decisions on crop selection based on their local environmental conditions. For instance, in regions with high rainfall and humidity, crops like **banana**, **coconut**, and **pomegranate** may be more suitable. Conversely, areas with less rainfall but good soil fertility may benefit from crops like **grapes**, **cotton**, or **lentil**.

This model also emphasizes the importance of soil fertility management, particularly the availability of nutrients like nitrogen, phosphorus, and potassium. By ensuring optimal levels of these nutrients, farmers can enhance crop yields and improve resource use efficiency. Furthermore, the impact of pH on nutrient availability means that soil testing and amendments may be necessary to optimize growing conditions for specific crops.

6. Conclusion

This comprehensive analysis demonstrates the effectiveness of machine learning approaches in crop recommendation systems. The high accuracy across multiple models, validated by statistical analysis, provides confidence in the reliability of these predictions. The Random Forest model's superior performance, combined with the clear identification of key environmental factors, offers valuable guidance for agricultural planning and resource management. These insights contribute to more sustainable and efficient agricultural practices, potentially improving food security and resource utilization in farming communities.

Model	Accuracy	R2	MSE	F1	Recall
Logistic Regression	0.9689	0.9087	3.0706	0.9673	0.9689
Random Forest	0.9887	0.9624	1.2655	0.9865	0.9887
Refine RF	0.9915	0.9745	0.8588	0.9919	0.9915
XGBoost	0.9859	0.958	1.4124	0.9854	0.9859
KNN	0.9746	0.9359	2.1554	0.9738	0.9746
Decision Tree	0.9718	0.8903	3.6864	0.9715	0.9718
Gaussian NB	0.9859	0.9616	1.291	0.9862	0.9859
Neural Network	0.9689	0.9037	3.2373	0.9638	0.9689

Reference:
https://www.kaggle.com/datasets/atharvaingle/crop-recommendation-dataset/data