Crop Recommendation System

End-to-End Implementation

Prepared by:
John Olalemi
Data Scientist

Supervised by:
Mr Joshua Oluwatoyin
Machine Learning and AI Specialist

Institution:
Great Brands Nigeria Limited

Date:
December 10, 2024

Contact Information:
Email: johnolalemi90@gmail.com
GitHub: https://github.com/johnnysnipes90
LinkedIn: linkedin.com/in/johnolalemi90

Prepared as part of a professional contribution to the development of data-driven solutions for agriculture.

Introduction

The increasing global demand for food has prompted the need for sustainable agricultural practices. One of the key factors in maximizing crop yield and ensuring sustainable farming is selecting the right crop based on available soil and environmental conditions. This project aims to develop a crop recommendation system that uses machine learning models to predict the best crop choices based on soil parameters such as Nitrogen, Phosphorus, Potassium, pH, and other environmental factors.

Their primary objective is to maximize the yield of their crops, taking into account different factors. One crucial factor that affects crop growth is the condition of the soil in the field, which can be assessed by measuring basic elements the soil consists of. Each crop has an ideal soil condition that ensures optimal growth and maximum yield.

Measuring essential soil metrics such as nitrogen, phosphorous, potassium levels, and pH value is an important aspect of assessing soil condition. However, it can be an expensive and time-consuming process, which can cause farmers to prioritize which metrics to measure based on their budget constraints.

Why is Crop Recommendation Necessary?

- **Optimal Yield**: Tailored crop recommendations maximize yield by addressing nutrient deficiencies and optimizing environmental conditions for growth.
- **Resource Management**: Efficient allocation of fertilizers, water, and other inputs minimizes costs and environmental impact.
- **Sustainability**: Promotes soil health, reduces nutrient runoff, and minimizes reliance on synthetic inputs.
- **Climate Resilience**: Helps farmers adapt to climate variability and mitigate risks from extreme weather events.
- **Improved Profitability**: Higher yields, better crop quality, and reduced input costs lead to increased profits.

Aim and Objectives

Aim:
To develop an end-to-end crop recommendation system using machine learning techniques based on soil and environmental data.

Objectives:

- To preprocess and analyze soil and environmental data for effective modeling.
- To identify key factors influencing crop growth and yield.
- To build and evaluate machine learning models for crop prediction.
- To deploy the model and provide a user-friendly interface for farmers.
- To ensure scalability and adaptability for global application.

Dataset Overview

The dataset for this project was sourced from `kaggle.com/dataset`. Each row in the dataset represents various measures of the soil in a particular field, and based on these measurements, the crop specified in the `Crop` column is the optimal choice for that field.

Data Dictionary

Feature Name	Type	Description
Nitrogen	Integer	Nitrogen content ratio in the soil.
Phosphorus	Integer	Phosphorous content ratio in the soil.
Potassium	Integer	Potassium content ratio in the soil.
Temperature	Float	Soil temperature in degrees Celsius.
Humidity	Float	Relative humidity percentage in the field.
pH_Value	Float	Measure of soil acidity/alkalinity.
Rainfall	Float	Amount of rainfall in millimeters.
Crop	Categorical	Target variable representing the optimal crop for the soil condition.

Table 1: Data Dictionary

Methodology

Preprocessing

1. **Outlier Removal**: Outliers were handled using the Interquartile Range (IQR) method to prevent extreme values from skewing the model's performance. This ensures that the model is trained on representative data, leading to more robust predictions.

2. **Label Encoding**: The `Crop` column was label-encoded to numerical values to make it compatible with machine learning algorithms. Numerical representation of categorical variables is necessary for most machine learning models, which can only process numerical data.

3. **Scaling**: Features were scaled using `StandardScaler` to normalize the data. This step was essential because many machine learning algorithms, such as Support Vector Machines and Logistic Regression, perform better when the input features are on a similar scale.

4. **Train-Test Split**: The dataset was split into training (80%) and testing (20%) sets to evaluate the model on unseen data and prevent data leakage. This practice ensures that the model's performance is generalized and not overfitted to the training data.

Model Selection and Training

- **Why Random Forest?** Random Forest was chosen due to its ability to handle high-dimensional data and its robustness against overfitting. This ensemble learning method combines multiple decision trees and averages their predictions, making it effective for both classification and regression tasks. Its interpretability, through feature importance scores, makes it particularly suitable for identifying the most critical soil parameters influencing crop recommendation.

- **Why Compare Models?** To ensure the most accurate and reliable predictions, other models such as Logistic Regression and Support Vector Machines were evaluated. Comparing models allows us to identify the algorithm best suited for the dataset based on performance metrics such as accuracy, precision, and recall.

- **Performance Metrics:** Performance metrics like accuracy, precision, recall, and F1 score were used to assess the models. These metrics provide a comprehensive evaluation, considering both the model's prediction accuracy and its ability to balance false positives and negatives.

Feature Importance Analysis

Random Forest's feature importance functionality was used to rank the soil and environmental variables in terms of their contribution to the model's decision-making. This analysis helps prioritize the most critical factors, such as Nitrogen, pH Value, and Rainfall, for crop selection, providing actionable insights for farmers.

Results

Table 2: Summary Statistics of the Dataset

Statistic	Nitrogen	Phosphorus	Potassium	Temperature	Humidity	pH Value	Rainfall
Count	2200.00	2200.00	2200.00	2200.00	2200.00	2200.00	2200.00
Mean	50.55	53.36	48.15	25.62	71.48	6.47	103.46
Std Dev	36.92	32.99	50.65	5.06	22.26	0.77	54.96
Min	0.00	5.00	5.00	8.83	14.26	3.50	20.21
25%	21.00	28.00	20.00	22.77	60.26	5.97	64.55
Median (50%)	37.00	51.00	32.00	25.60	80.47	6.43	94.87
75%	84.25	68.00	49.00	28.56	89.95	6.92	124.27
Max	140.00	145.00	205.00	43.68	99.98	9.94	298.56

Outlier Detection Results

Outlier detection using the Z-score method revealed the presence of extreme values in *Potassium*, *Nitrogen*, and *Rainfall* variables. Boxplots for each feature were analyzed to visualize these outliers and assess their potential impact on model performance. Corrective measures such as capping extreme values were applied.

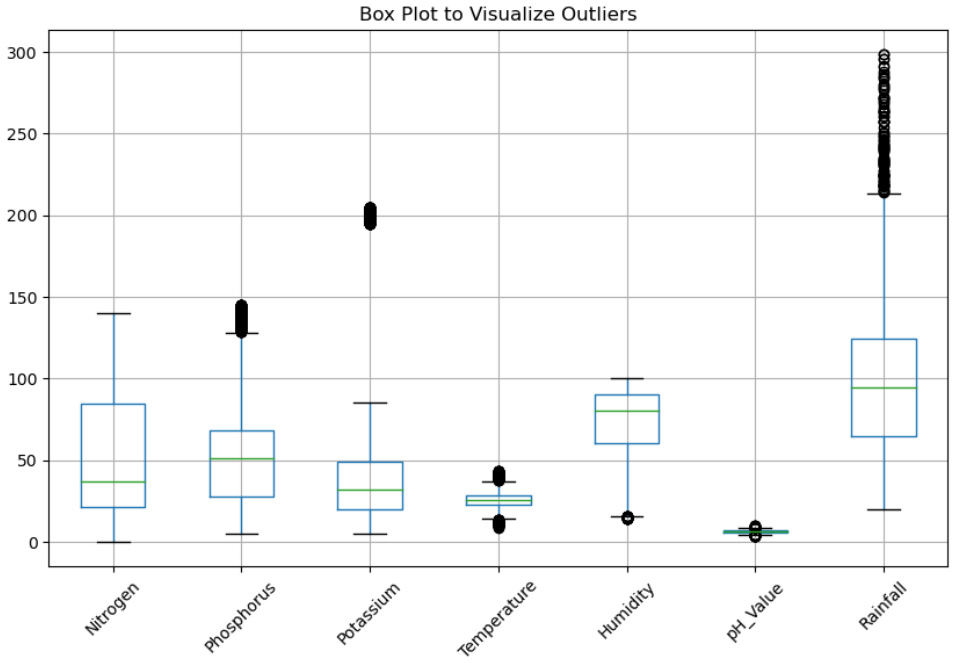

Figure 1: Outliers Detected in Key Features

Feature Importance

The Random Forest Classifier was employed to evaluate feature importance. Figure 2 highlights the most significant features, with *Rainfall*, *Humidity*, *Potassium*, *Phosphorus*, and *Nitrogen* ranking among the top contributors.

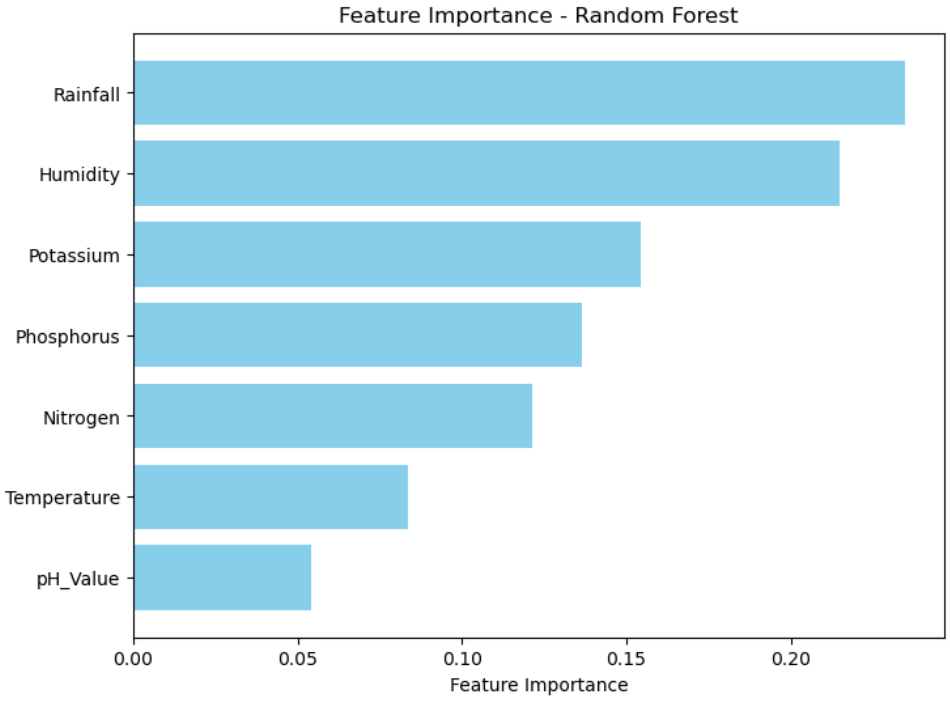

Figure 2: Feature Importance Derived from Random Forest Classifier

Correlation Analysis

The correlation analysis was conducted to evaluate the relationship between features. Figure 3 illustrates the heatmap of correlations among the features.

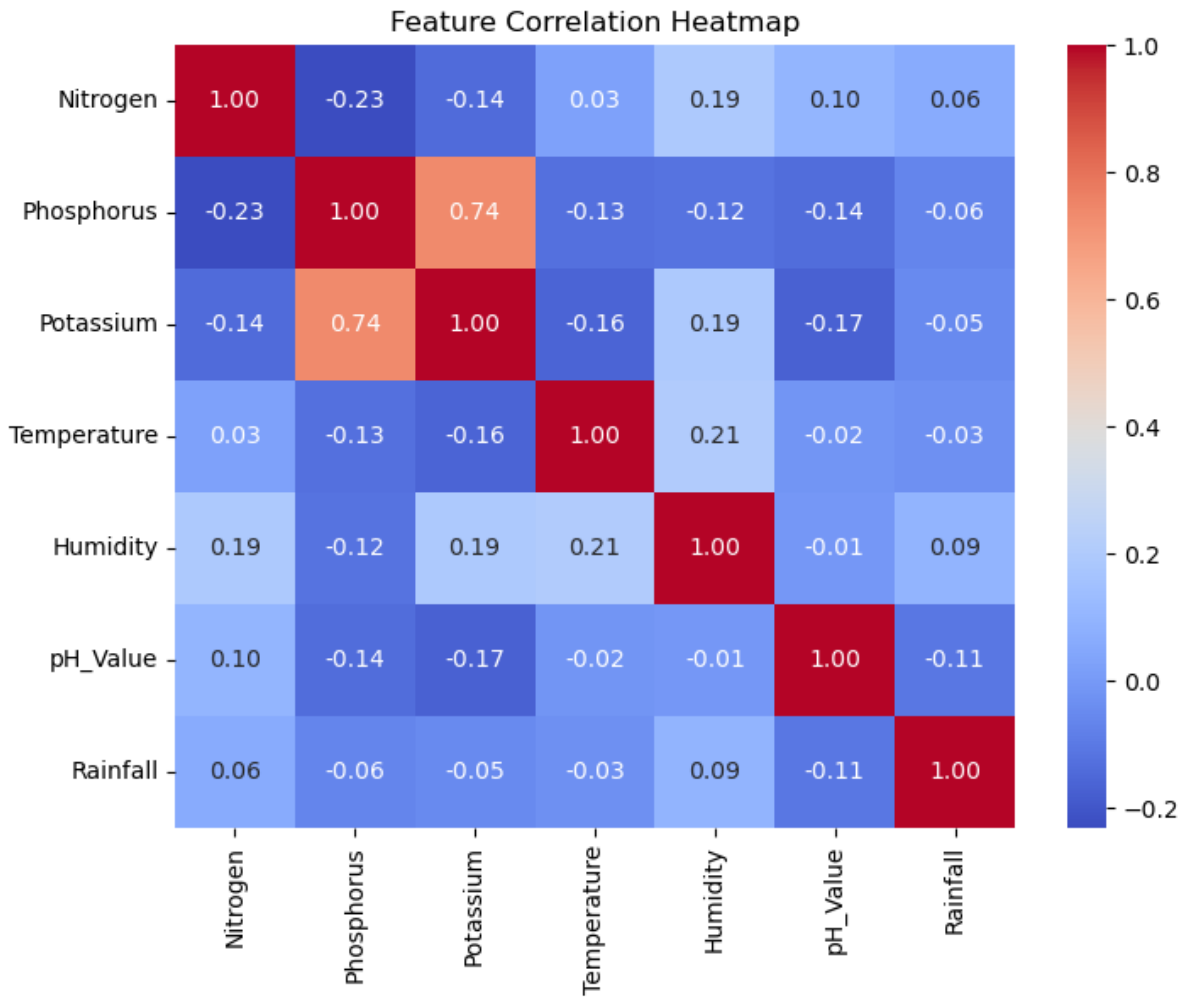

Figure 3: Heatmap Showing Correlation Among Features

Crop Labels

Table 3 provides the crop labels used in the classification task:

Table 3: Crop Labels

Label	Crop
0	Banana
1	Blackgram
2	ChickPea
3	Coconut
4	Coffee
5	Cotton
6	Jute
7	KidneyBeans
8	Lentil
9	Maize
10	Mango
11	MothBeans
12	MungBean
13	Muskmelon
14	Orange
15	Papaya
16	PigeonPeas
17	Pomegranate
18	Rice
19	Watermelon

Confusion Matrix

The confusion matrix for the Random Forest Classifier is presented in Figure 4. It highlights the classification performance across different crops.

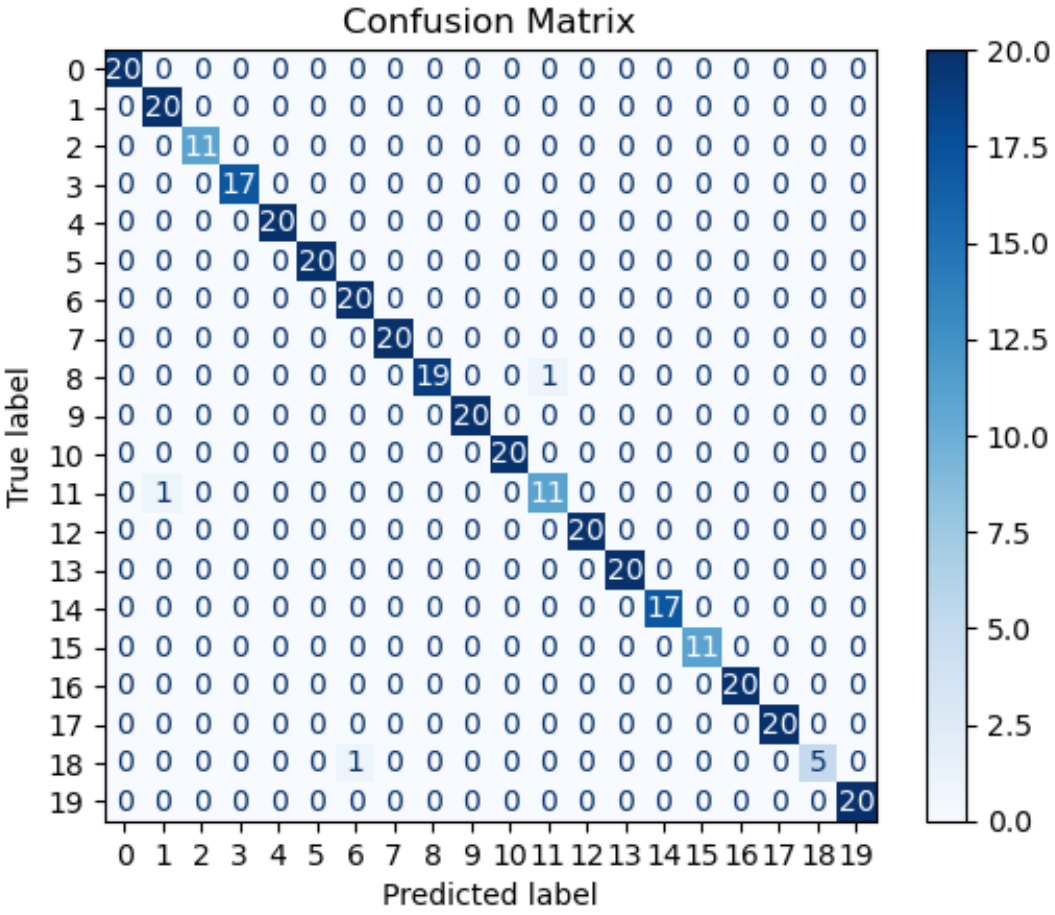

Figure 4: Confusion Matrix for Crop Classification

Model Evaluation

The performance of the tuned Random Forest model is evaluated on test data. The classification report for the model on the test data is shown below:

Classification Report on Test Data:

	precision	recall	f1-score	support
0	1.00	1.00	1.00	20
1	0.95	1.00	0.98	20
2	1.00	1.00	1.00	11
3	1.00	1.00	1.00	17
4	1.00	1.00	1.00	20
5	1.00	1.00	1.00	20
6	0.95	1.00	0.98	20
7	1.00	1.00	1.00	20
8	1.00	0.95	0.97	20
9	1.00	1.00	1.00	20
10	1.00	1.00	1.00	20
11	0.92	0.92	0.92	12
12	1.00	1.00	1.00	20
13	1.00	1.00	1.00	20
14	1.00	1.00	1.00	17
15	1.00	1.00	1.00	11
16	1.00	1.00	1.00	20
17	1.00	1.00	1.00	20
18	1.00	0.83	0.91	6
19	1.00	1.00	1.00	20
accuracy			0.99	354
macro avg	0.99	0.98	0.99	354
weighted avg	0.99	0.99	0.99	354

Additional metrics include:

- Training Accuracy: 1.0
- Test Accuracy: 0.9915
- Training Log-Loss: 0.0070
- Test Log-Loss: 0.0555
- Training AUC: 1.0
- Test AUC: 0.9988

The Random Forest model demonstrated excellent performance, achieving high accuracy, precision, recall, and F1-scores across the majority of crop classes. The model's ability to generalize well to unseen data is evident in its high test accuracy and low log-loss values.

Model Deployment

After training the model, we integrate it into a user-friendly interface using FlaskAPI for the backend and Streamlit for the frontend. FlaskAPI serves the trained model through an API, while Streamlit allows users to input soil data through a form and receive real-time crop recommendations.

Steps for Streamlit Deployment

1. Install Streamlit and integrate the trained model.
2. Create interactive input fields for soil parameters.
3. Display crop recommendation results and model confidence.
4. Deploy the Streamlit app to a platform like Streamlit for easy access by farmers.

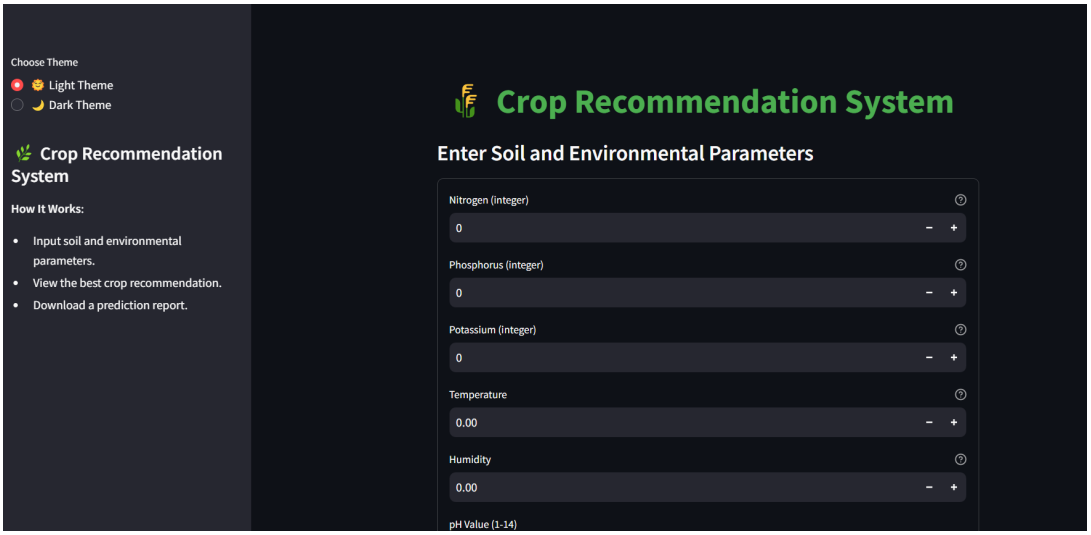

Figure 5: Streamlit app interface

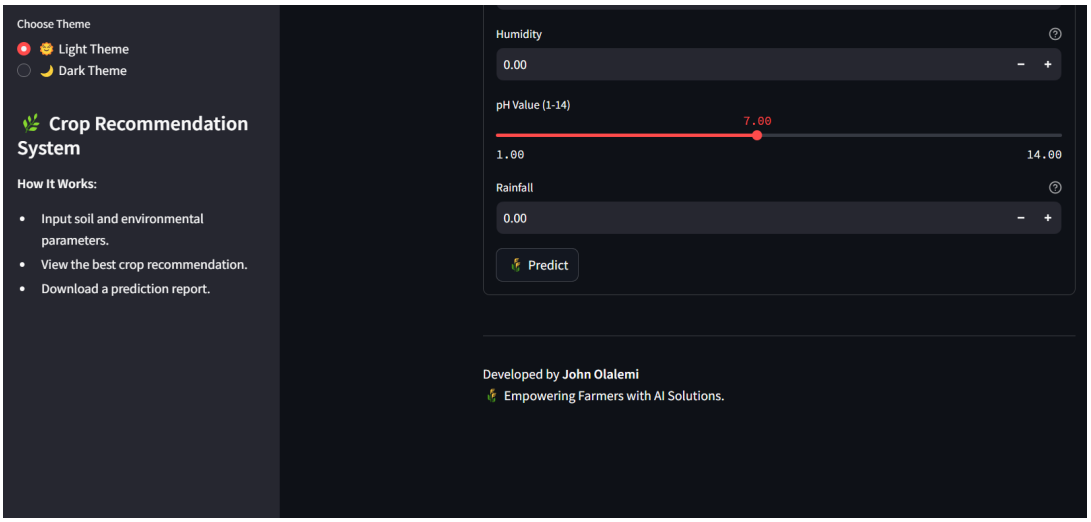

Figure 6: Streamlit app interface

Result Interpretation

- The Random Forest model achieved high accuracy in classifying crops, with the confusion matrix showing minimal misclassification among similar crops. - The confusion matrix revealed that most crops were accurately predicted, with high values along the diagonal, indicating strong model performance. However, class 18 showed notable misclassification, where only 5 predictions were correct, and a significant number were misclassified into another crop class. This highlights the need for further refinement to improve accuracy for specific classes. - Correlation analysis revealed strong relationships between *Nitrogen*, *Phosphorus*, and *Potassium*, affirming their relevance in crop prediction. - Outlier handling improved model robustness by reducing extreme-value bias. - Feature importance analysis confirmed *Rainfall*, *Temperature*, and *Nitrogen* as key predictors for crop selection.

Conclusion

This project demonstrates the use of machine learning to recommend optimal crops based on soil and environmental data. Random Forest proved to be the most effective model, achieving high accuracy and providing interpretable feature importance rankings.

The crop recommendation system developed in this project successfully predicts optimal crop choices based on soil and environmental parameters. Random Forest emerged as the best-performing model, with significant contributions from Nitrogen, pH value, and Rainfall. The system provides actionable insights for farmers, enabling them to make informed decisions on crop selection based on available resources and soil health, promoting sustainability and maximizing yield.