United States International University – Africa

School of Science and Technology

FS 2025

Capstone Project

Project Title: Smart Crop Recommendation System

Team Members

NAME	REG NO	ROLE
Muhia, Wilson Junior Wambugu	669024	Data & Preprocessing Lead
ZAKARIYA SHAFI	668596	Deployment & Documentation Lead
Ambachow Kahsay	670550	Modeling & Evaluation Lead

Course: DSA3020- Principles of Machine Learning

Date: 11/30/2025

Instructor: Dr. Dennis Njagi

Abstract

Agriculture is a critical sector for food security, economic stability, and sustainable development. However, farmers often face challenges in identifying the optimal crops to cultivate based on soil and environmental conditions. This project proposes a *Smart Crop Recommendation System* that leverages machine learning to recommend the most suitable crops, maximizing yield and minimizing resource wastage. Using real-world agricultural datasets, the project applies classification algorithms—including Logistic Regression, Decision Trees, Random Forest, and XGBoost—to predict the optimal crop based on features such as nitrogen, phosphorus, potassium, soil pH, temperature, humidity, and rainfall. The study ensures a rigorous machine learning workflow by considering **data balancing**, **data leakage prevention**, and appropriate **feature selection**. The best-performing model, a reduced and hyperparameter-tuned Random Forest Classifier, achieves a test accuracy of **99%**, precision and recall of **0.99**, and an F1-score of **0.99**, demonstrating its robustness and reliability. This system has the potential to empower farmers with data-driven insights, contributing to sustainable agriculture.

Problem Statement

Agriculture remains one of the most resource-intensive sectors, with crop yields heavily influenced by soil fertility, climatic factors, and environmental conditions. Farmers often make crop selection decisions based on intuition, tradition, or general recommendations, which can lead to suboptimal yields, resource wastage, and financial losses. Effective crop selection is therefore crucial for maximizing yield, reducing losses, and ensuring sustainable agricultural practices. To address this challenge, this project develops a **machine learning-based Smart Crop Recommendation System** that predicts the most suitable crop for a given set of soil and environmental conditions. The system leverages measurable soil properties—such as nitrogen, phosphorus, potassium, and pH—alongside environmental factors like temperature, humidity, and rainfall. The target variable is the crop type, which is categorical, making this a **multiclass classification problem**. Machine learning is particularly well-suited for this task because it can model complex, non-linear interactions between soil nutrients, environmental factors, and crop suitability, enabling precision agriculture and data-driven decision-making for farmers (Géron, 2023).

Table of Contents

Figures

Literature Review

Several studies have explored the application of machine learning techniques in agriculture, particularly for tasks such as crop yield prediction and crop recommendation. Logistic Regression (LR) is a widely used linear classifier suitable for multiclass problems when the relationship between features and target classes is approximately linear. Its primary strength lies in its simplicity, ease of implementation, and interpretability, as the coefficients directly indicate the influence of each feature on the prediction. However, LR struggles to capture complex non-linear relationships inherent in agricultural data (Hosmer et al., 2013).

Decision Trees (DTC) and Random Forests (RFC) are ensemble-based, non-linear models capable of capturing intricate patterns and interactions among soil and environmental variables. Decision Trees are highly interpretable, allowing visualization of decision rules, which is valuable for understanding feature importance and model logic. Random Forests, being an ensemble of multiple trees, improve predictive performance and reduce overfitting but at the cost of reduced interpretability compared to single decision trees (Breiman, 2001).

Support Vector Machines (SVM) are particularly effective in high-dimensional feature spaces and can model complex, non-linear relationships using kernel functions. SVMs provide robust classification boundaries but are less interpretable and computationally intensive for large datasets.

XGBoost (XGBC) combines gradient boosting with regularization, allowing the model to achieve high accuracy while controlling overfitting. Although XGBoost is highly performant, its complexity can make it difficult to interpret, requiring feature importance metrics or SHAP values for explanation (Chen & Guestrin, 2016).

Across all models, careful preprocessing, including normalization, handling of categorical variables, and ensuring balanced datasets, is essential to achieve reliable and interpretable predictions. While simpler models like LR offer clarity and ease of interpretation, ensemble and boosting methods provide superior accuracy in capturing the non-linear interactions typical of agricultural data, albeit with increased complexity. Selecting the appropriate model involves balancing predictive performance with interpretability, depending on the end-user needs for actionable insights in precision agriculture.

Methodology

The project followed the **CRISP-DM framework**. The first step involved understanding the business problem: providing farmers with reliable crop recommendations based on soil and environmental data. Data were collected from publicly available datasets and included features such as Nitrogen, Phosphorous, Potassium, pH, temperature, humidity, and rainfall.

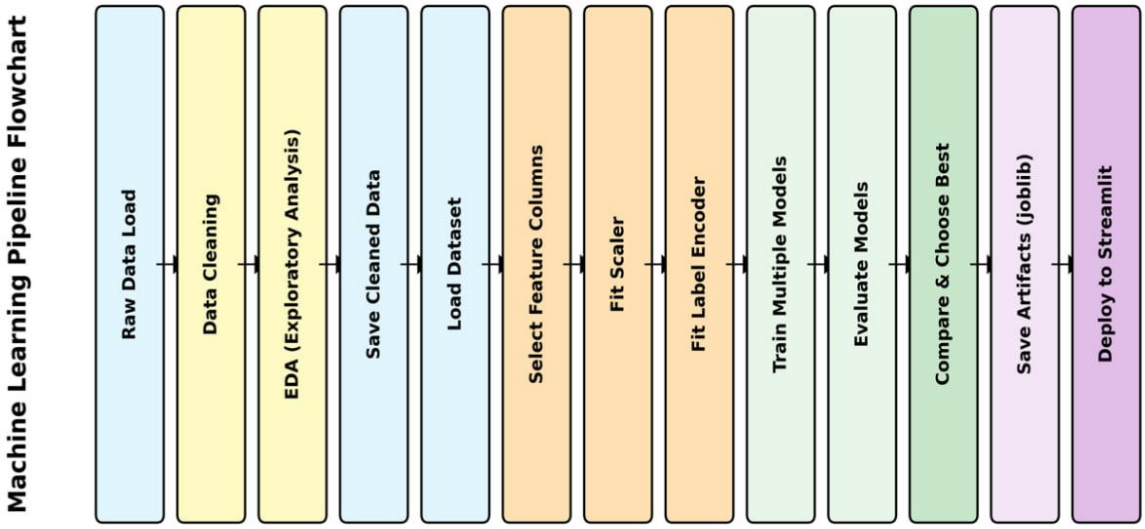

Figure 1:Machine Learning pipeline flowchart

Data Description & Preprocessing

The dataset utilized in this study encompasses key soil properties Nitrogen (N), Phosphorus (P), Potassium (K), and pH as well as environmental factors, including temperature, humidity, and rainfall, collected from multiple agricultural regions. The target variable is the crop type, which is categorical and includes a diverse set of crops such as apples, maize, rice, and chickpea. Descriptive statistics and visualizations were employed to understand the distribution of each feature. For instance, histograms and boxplots highlighted the spread and potential outliers in soil nutrients and climatic variables. Correlation analysis was performed to identify multicollinearity among features, guiding the selection of the most informative variables for model training. Furthermore, the dataset was assessed for class imbalance. Each crop category had nearly equal representation, mitigating the risk of biased predictions.

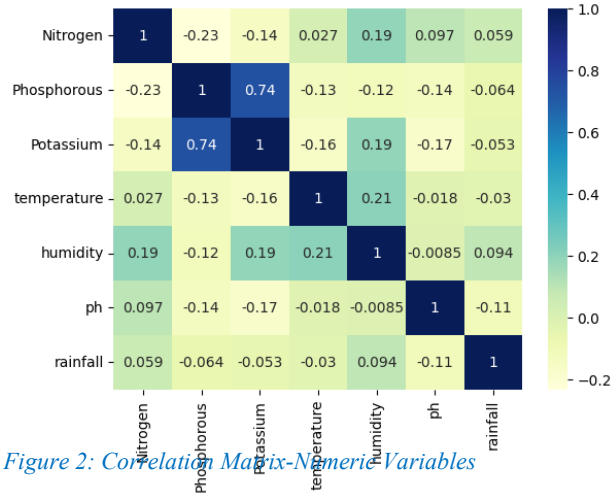

Figure 2: Correlation Matrix-Numeric Variables

Potassium and Phosphorous show moderate positive correlation (0.74). For **linear models** like Logistic Regression, this can cause **multicollinearity**, making coefficient interpretation unstable, though overall accuracy may remain high; mitigation includes **removing one feature, applying PCA, or using L2 regularization**. For **tree-based models** (Decision Trees, Random Forest, XGBoost), correlation is generally not an issue for predictions, but feature importance may be **split between correlated variables**, affecting interpretability.

Train Test Split

Prior to modeling, the dataset was split into training (80%) and test sets (20%) to prevent data leakage, ensuring that the models were evaluated on unseen data. The training features were standardized using StandardScaler to ensure all features contributed equally to the model, particularly important for distance-based algorithms like SVM and for gradient-boosted models like XGBoost. The target variable was encoded using LabelEncoder, transforming categorical crop labels into numeric codes suitable for classification algorithms.

Model Development & Evaluation

For this project, multiple machine learning algorithms were explored to identify the most suitable model for crop recommendation. The chosen models included Logistic Regression (LR), Decision Tree Classifier (DTC), Random Forest Classifier (RFC), Support Vector Machine (SVM), and XGBoost Classifier (XGBC). Each model was trained on the training set and evaluated on the test set using metrics suitable for multiclass classification, including precision, recall, F1-score, accuracy, and mean absolute error (MAE).

Logistic Regression was used as a baseline linear model. Both standard and L2-regularized versions were trained to evaluate the impact of regularization on performance. Prior to L2 regularization, the model achieved a training accuracy of 0.97 and test accuracy of 0.96. After applying L2 regularization, the training accuracy slightly decreased to 0.94 and test accuracy reduced to 0.9, indicating a minor underfitting effect introduced by regularization. Logistic Regression offers high interpretability, allowing users to understand the contribution of each feature (e.g., soil nutrients or rainfall) to crop prediction. However, it assumes linear relationships between features and the log-odds of the output, which may limit performance on complex patterns

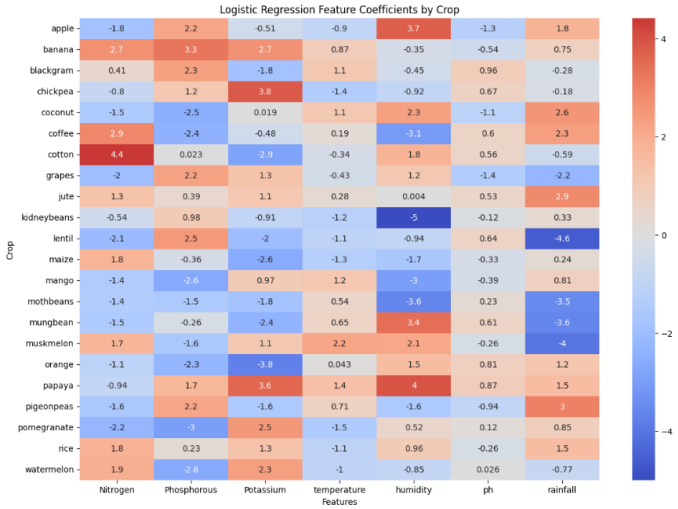

Figure 3:Logistic Regression Coefficients

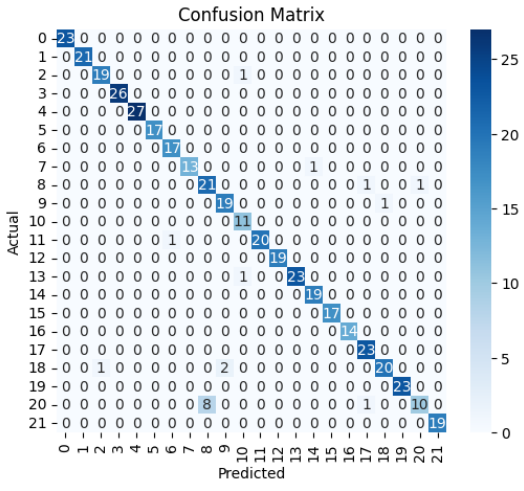

Figure 4: Confusion Matrix - Logistic Regression

Decision Tree Classifier (DTC) was employed to capture non-linear relationships between soil and environmental features. Initially, the DTC achieved training and test accuracies of 0.92 and 0.89, respectively. Hyperparameter tuning, including adjustments to max_depth, min_samples_split, min_samples_leaf, and pruning with ccp_alpha, improved the model substantially. The tuned Decision Tree reached a cross-validated accuracy of 0.981 and test

accuracy of 0.98. Decision Trees are intuitive and easy to visualize, making them useful for communicating model decisions to non-technical stakeholders. However, they are prone to overfitting, especially on small datasets or with many features, which necessitate careful tuning.

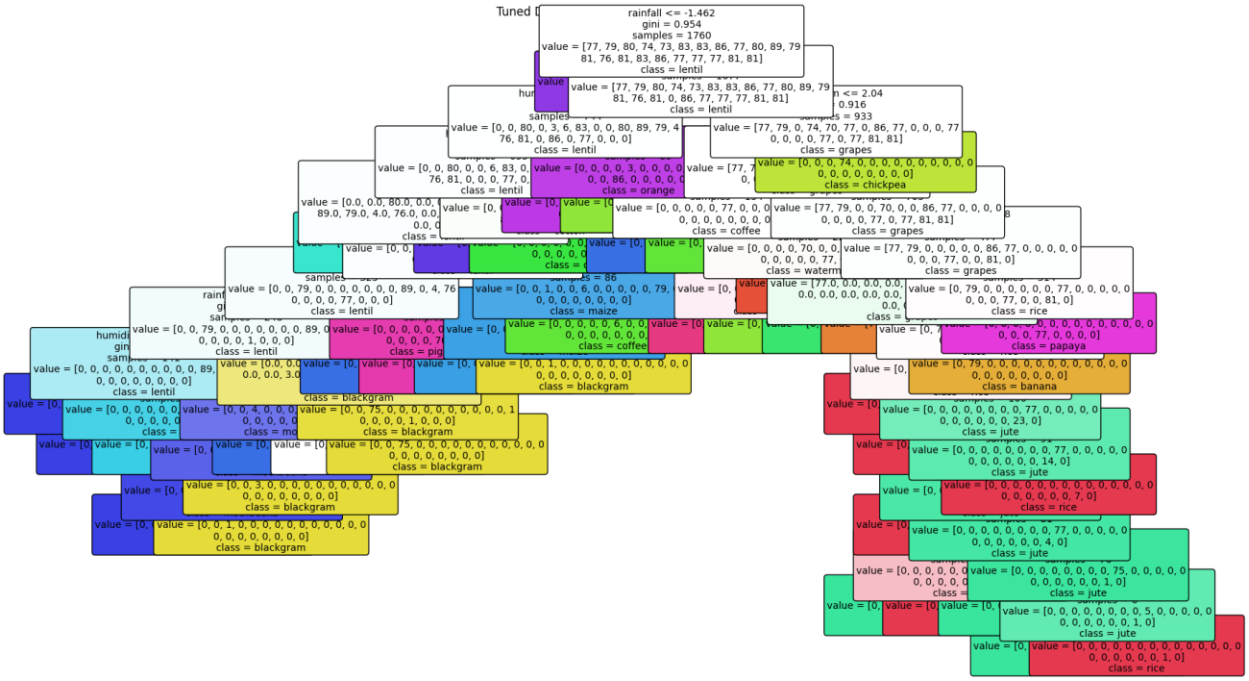

Figure 5: Tuned Decision Tree visualization

Random Forest Classifier (RFC), an ensemble of Decision Trees, was leveraged to further enhance performance and reduce overfitting. Initially trained with default parameters, RFC achieved test accuracy of 0.99. Hyperparameter tuning with RandomizedSearchCV optimized parameters including n_estimators=274, max_depth=16, min_samples_split=7, min_samples_leaf=3, and criterion='entropy', resulting in cross-validated accuracy of 0.99. To improve efficiency and interpretability, a Reduced & Tuned Random Forest Classifier (RTRFC) was trained using only five selected features: Potassium, Humidity, Rainfall, Nitrogen, and Phosphorous. This optimized model maintained high performance with a test and train accuracy of 1.00, macro-averaged precision of 0.99, recall of 1.00, and F1-score of 0.99 (Figure 6: Feature importance of RTRFC). Random Forest models are highly robust to noise and capable of capturing complex interactions but are less interpretable than linear models.

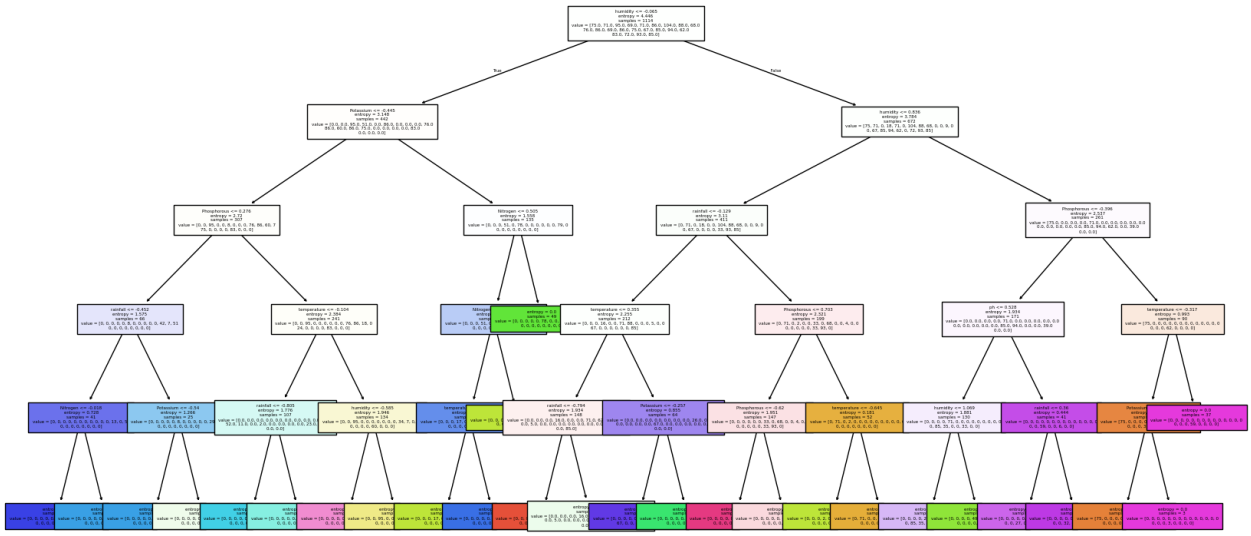

Figure 6: One tree Visualization-SVM

Support Vector Machines (SVM) and XGBoost Classifier (XGBC) were also evaluated to ensure a comprehensive comparison of machine learning approaches. The SVM model achieved a test accuracy of 0.97, demonstrating strong capability in handling high-dimensional and non-linear data patterns. XGBoost, leveraging gradient boosting with additional regularization, attained a test accuracy of 0.99, making it highly resistant to overfitting. Both models, however, are less interpretable than simpler approaches such as Logistic Regression or Decision Trees and require careful hyperparameter tuning to achieve optimal performance.

During model evaluation, careful attention was given to data quality considerations. The dataset was confirmed to be balanced across all crop categories, eliminating bias toward overrepresented classes. Data splitting into training and test sets was performed prior to any scaling or encoding, preventing leakage and ensuring unbiased evaluation. Multiclass considerations were incorporated through macro-averaged metrics, treating all crop classes equally, and model performance was further analyzed using confusion matrices, meaning absolute error (MAE), and class-wise F1-scores Confusion matrices SVC. Cross-validation was applied to assess model stability and generalization.

The XGBoost Classifier was trained on the full feature set, including soil properties (Nitrogen, Phosphorus, Potassium, pH) and environmental factors (temperature, humidity, rainfall). After hyperparameter tuning, XGBC achieved a training accuracy of 1.00 and a test accuracy of 0.99, with macro-averaged precision, recall, and F1-score all at 0.99, demonstrating excellent generalization across all crop classes. Feature importance analysis highlighted Potassium and Humidity as the most influential predictors for crop classification. Although XGBC provides superior accuracy, its complexity reduces interpretability, which may be a concern for stakeholders seeking insight into decision-making.

Results and Discussion

The results indicate that a reduced set of features is sufficient for accurate crop prediction, simplifying interpretability. Hyperparameter tuning improved Decision Tree and Random Forest models. Ensemble models (Random Forest, XGBoost) outperformed single estimators, confirming their suitability for complex, multiclass agricultural tasks. Feature importance analysis highlighted Potassium, humidity, and rainfall as key predictors.

Figure 8: Comparison bar chart of test accuracies for all models

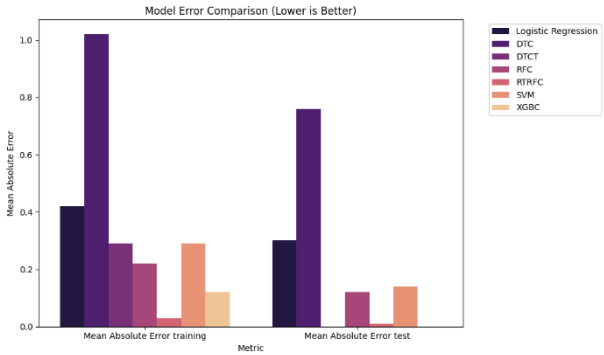

Figure 7: Comparison bar chart of test accuracies for all models

Overall, the Reduced & Tuned Random Forest Classifier (RTRFC) emerged as the best-performing model when balancing accuracy and interpretability. RTRFC achieved near-perfect predictive performance while allowing visualization of feature importance, making it suitable for practical deployment. While simpler models like Logistic Regression and Decision Trees are valuable for interpretability and visual insights, XGBoost offers a robust alternative when maximum predictive performance is required. The comparative evaluation demonstrates that ensemble and gradient boosting techniques are highly effective for multiclass crop recommendation tasks. To reduce dimensionality and improve computational efficiency, feature selection was performed based on domain knowledge and model-driven importance scores. The five key features **Potassium, Humidity, Rainfall, Nitrogen, and Phosphorous** were selected for training the optimized models. This choice balances predictive accuracy and model interpretability, enabling stakeholders to understand the most influential factors in crop recommendation.

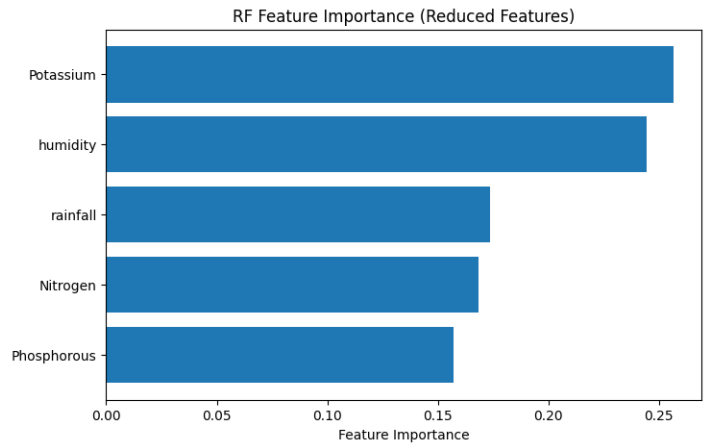

Figure 9: Feature Importance-Random Forest

Model Deployment

The final crop recommendation model was developed using a refined set of five highly informative features: **Potassium**, **Humidity**, **Rainfall**, **Nitrogen**, and **Phosphorous**. These features were selected based on domain relevance and their strong contribution to predictive performance during exploratory analysis. This pipeline forms the foundation for training the **Reduced & Tuned Random Forest Classifier (RTRFC)**, which was identified as the optimal model due to its balance of accuracy, robustness to non-linear interactions, and interpretability through feature importance analysis.

The deployed Streamlit application enables farmers to input soil and environmental data and instantly receive crop recommendations, translating model predictions into actionable decisions.

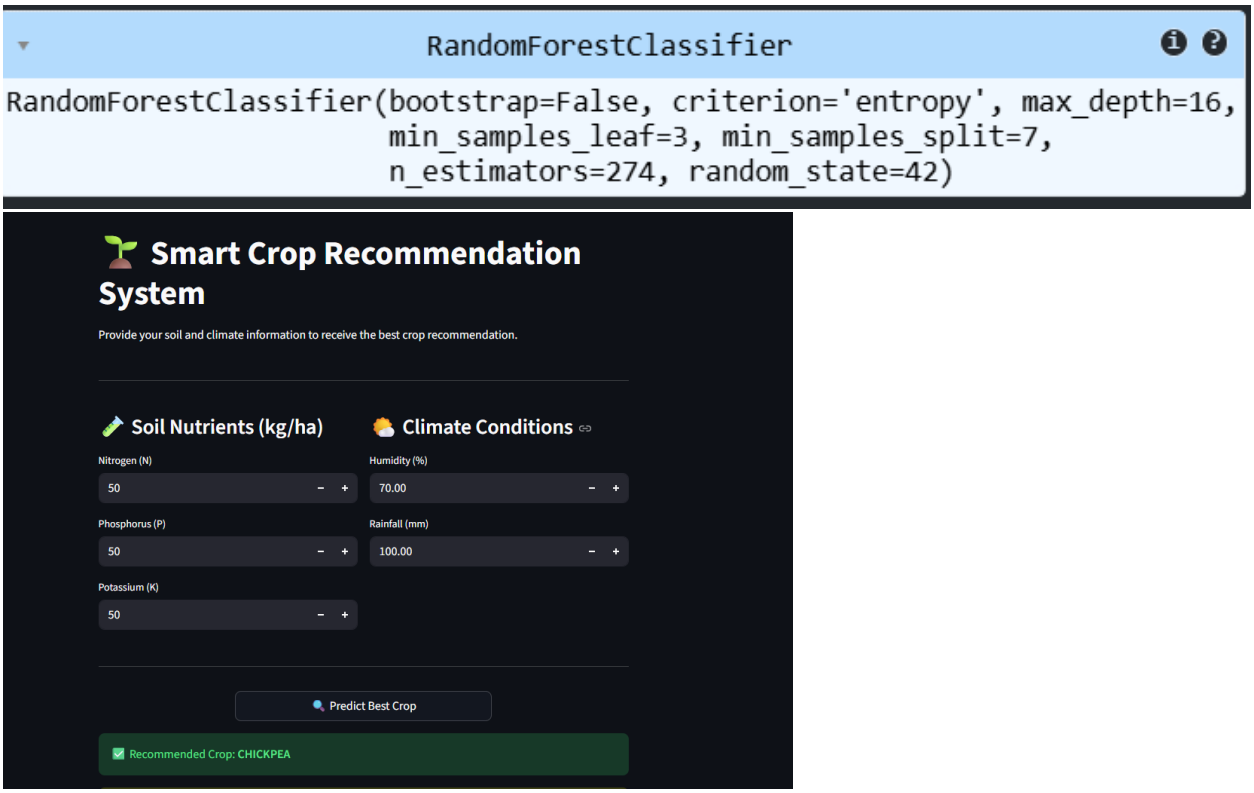

Figure 10: Demo

Conclusion & Future Work

The project successfully developed a high-performing crop recommendation system using machine learning. Feature reduction and hyperparameter tuning produced a reliable, interpretable system. Future work could integrate additional environmental parameters (e.g., sunlight, soil texture, microclimate) and explore deep learning models for more complex interactions. Real-time sensor integration could further enhance model accuracy and utility.

References

Géron, A. (2023). *Hands-On Machine Learning with Scikit-Learn, Keras & TensorFlow*. O'Reilly Media.

Kuhn, M., & Johnson, K. (2019). *Feature Engineering and Selection: A Practical Approach for Predictive Models*. CRC Press.

Kaggle Datasets: Crop Recommendation Dataset. https://www.kaggle.com/

Scikit-learn Documentation. https://scikit-learn.org/stable/