CROP RECOMMENDATION PREDICTION

A Course Project report submitted

in partial fulfillment of requirement for the award of degree

BACHELOR OF TECHNOLOGY

in

ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING

by

UGGE MAHEESH VARMA (2303A52163)

Under the guidance of

Dr. S VAIRACHILAI

Professor, Department of CSE.

Department of Computer Science and Artificial Intelligence

ABSTRACT

This project presents a machine learning–based approach for crop recommendation using soil nutrient and climatic parameters. The dataset consists of features such as nitrogen (N), phosphorus (P), potassium (K), temperature, humidity, pH, and rainfall, with the target variable being the type of crop suitable for cultivation. The data was carefully preprocessed by checking for missing values, duplicates, and outliers, followed by encoding and feature standardization.

A Random Forest Classifier was trained on the dataset and evaluated using standard metrics. The model achieved an impressive accuracy of 99%, with precision, recall, and F1-scores all exceeding 99%, indicating excellent predictive performance. To enhance interpretability, SHAP analysis was employed, which highlighted rainfall, humidity, and soil nutrients (N, P, K) as the most influential features driving the model's predictions.

The results demonstrate the potential of combining machine learning with explainable AI techniques for supporting agricultural decision-making. While the model performs exceptionally well, future improvements could involve testing on region-specific data, adding more environmental variables, and applying advanced ensemble methods to ensure broader applicability. This study not only validates the effectiveness of data-driven approaches in agriculture but also emphasizes the importance of transparency in AI models. Ultimately, the system can serve as a practical decision-support tool for farmers aiming to optimize yield and sustainability**.**

.

INTRODUCTION

Agriculture is a vital sector for food production and economic stability, yet farmers often face challenges in choosing the right crop for cultivation. The decision depends on multiple factors such as soil nutrients, weather conditions, and rainfall patterns. A wrong choice can lead to poor yields, wasted resources, and financial setbacks. With increasing climate variability, relying only on traditional knowledge or experience is no longer sufficient for sustainable farming.

The problem can be addressed by leveraging modern computational methods. Machine learning, in particular, provides a powerful way to analyze large datasets and uncover patterns that can guide decision-making. A crop recommendation system that predicts the most suitable crop based on environmental and soil conditions can serve as a valuable decision-support tool for farmers. Such systems not only improve agricultural efficiency but also contribute to food security and sustainability.

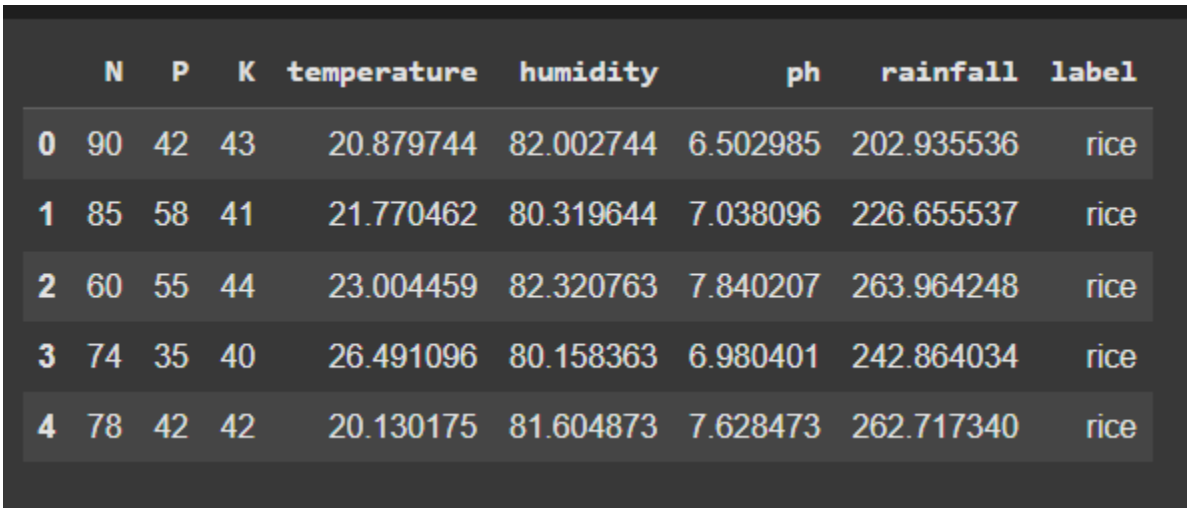
The primary **problem statement** in this project is to develop an intelligent crop recommendation system that uses soil nutrient values (N, P, K) and climatic factors (temperature, humidity, pH, and rainfall) to predict the most suitable crop for cultivation. The objective is to ensure accurate, data-driven recommendations that can help farmers optimize yield while reducing risks associated with guesswork and unfavorable environmental conditions.

For this study, the **Crop Recommendation Dataset** was used. It consists of 2,200 records with seven input features and one target variable. The features represent essential agricultural factors including nitrogen, phosphorus, potassium, temperature, humidity, pH value, and rainfall. The target variable corresponds to the crop label, covering 22 different crops such as rice, maize, cotton, coffee, and others. This makes it a multi-class classification problem, well-suited for supervised learning techniques.

By applying machine learning algorithms and explainable AI methods, the dataset can be transformed into meaningful insights. Not only does this allow for highly accurate predictions, but it also highlights the key features influencing crop selection. In this project, a Random Forest Classifier was applied, and its predictions were further explained using SHAP (SHapley Additive Explanations) to ensure both accuracy and interpretability.

DATASET:

The dataset contains 2,200 rows and 8 columns.



	N	P	K	temperature	humidity	ph	rainfall	label
0	90	42	43	20.879744	82.002744	6.502985	202.935536	rice
1	85	58	41	21.770462	80.319644	7.038096	226.655537	rice
2	60	55	44	23.004459	82.320763	7.840207	263.964248	rice
3	74	35	40	26.491096	80.158363	6.980401	242.864034	rice
4	78	42	42	20.130175	81.604873	7.628473	262.717340	rice

FIG 2- Dataset

Columns:

1. N (Nitrogen):

- ✓ Represents the nitrogen content in the soil.
- ✓ Nitrogen is an essential macronutrient for plant growth as it plays a key role in photosynthesis and chlorophyll formation.
- ✓ Different crops require different levels of nitrogen for healthy development.

2.P (Phosphorus):

- ✓ Refers to the phosphorus content in the soil.
- ✓ Phosphorus is crucial for root development, flowering, and seed formation.
- ✓ Adequate phosphorus ensures stronger plants and better crop yields.

3.K (Potassium):

- ✓ Indicates the potassium content in the soil.
- ✓ Potassium helps regulate water uptake, improves disease resistance, and enhances overall crop quality.
- ✓ It also strengthens stems and supports fruit and grain production.

4. temperature (°C):

- ✓ Represents the average ambient temperature in degrees Celsius.
- ✓ Temperature directly affects seed germination, crop growth cycles, and productivity.
- ✓ Different crops thrive under specific temperature ranges (e.g., rice prefers warmer climates, while wheat grows in cooler conditions).

5.humidity (%):

- ✓ Denotes the relative humidity of the environment in percentage.
- ✓ Humidity influences plant transpiration, disease spread, and water balance.
- ✓ Crops like rice and sugarcane require high humidity, whereas others like chickpea prefer lower levels.

6.ph (Soil pH):

- ✓ Indicates the acidity or alkalinity of the soil.
- ✓ Measured on a scale of 0–14, where 7 is neutral.
- ✓ Most crops grow best in slightly acidic to neutral soil (pH 6–7.5). Extreme pH levels can reduce nutrient availability.

7.rainfall (mm):

- ✓ Represents the annual or seasonal rainfall in millimeters.
- ✓ Rainfall is a critical factor for irrigation and crop survival.
- ✓ Crops like rice need heavy rainfall, while others like lentils or chickpeas grow in drier conditions.

8.label (Target – Crop type):

- ✓ The target variable in the dataset.
- ✓ Contains **22 different crop categories** such as rice, maize, banana, cotton, coffee, chickpea, and others.
- ✓ The machine learning model uses the input features to predict this label.

PREPROCESSING STEPS:

1.Handling Missing Values:

- The dataset was checked for missing values in all columns.
- No missing values were found, so no imputation was required.

2.Removing Duplicates:

- Duplicate rows were identified and checked.
- The dataset did not contain any duplicate entries, ensuring uniqueness of records.

3.Outlier Detection and Removal:

- Outliers were identified using the **Interquartile Range (IQR) method**.
- Any values falling outside the acceptable lower and upper bounds were removed to maintain data integrity and avoid bias in model training.

4.Encoding Categorical Variable:

- The target variable, label (crop type), is categorical with 22 unique classes.
- It was transformed into numeric values using **Label Encoding**, enabling the machine learning model to process it effectively.

5.Feature Scaling:

- All numerical features (N, P, K, temperature, humidity, pH, rainfall) were standardized using **StandardScaler**.
- Standardization ensures that features are on the same scale, improving model performance and avoiding bias towards features with larger numeric ranges.

MODEL&PARAMETERS:

1. Algorithm Choice

The chosen algorithm for this project is the Random Forest Classifier from the scikit-learn library. Random Forest is an ensemble method that creates multiple decision trees and combines their predictions using majority voting. It is well-suited for multi-class classification problems and can effectively handle both numerical and categorical data.

Given the structure of the dataset (seven input features with different units and one categorical output), Random Forest was selected because of its robustness, interpretability, and ability to capture complex feature interactions without extensive parameter tuning.

2.Parameters:

From the implementation:

- **Model:** RandomForestClassifier(random_state=42)
- **Key Parameters Used:**
  - n_estimators: Default (100 trees)
  - criterion: Gini index (default)
  - max_features: Auto (default √features per split)
  - random_state: 42 (ensures reproducibility)
- **Train/Test Split:**

- Training set: 1476 samples (80%)
- Testing set: 370 samples (20%)

3. Evaluation Metrics:

1. Accuracy:

Definition: Accuracy is the proportion of correct predictions made by the model out of the total predictions. It measures the overall effectiveness of the classification model.

Formula: Accuracy= TP+TN/TP+TN+FP+FN

Where:

- **TP:** True Positives
- **TN:** True Negatives
- **FP:** False Positives
- **FN:** False Negatives

Result: 99.19%

2. PrecisionDefinition: Precision measures the proportion of correctly predicted positive cases out of all cases predicted as positive. It tells us how reliable the positive predictions are.

Formula: Precision=TP/TP+FP

Result: 99.38%

3. Recall (Sensitivity or True Positive Rate):

Definition: Recall measures the proportion of actual positive cases that were correctly identified by the model. It indicates how well the model can find all positive instances.

Formula: Recall :TP/TP+FN

Result: 99.19%

4. F1-Score:

Definition: F1-score is the harmonic mean of Precision and Recall. It balances the trade-off between the two, making it useful when both metrics are important.

Formula: F1 = 2 . Precision.Recall/Precision+Recall

Result: 99.15%

5. ROC AUC (Receiver Operating Characteristic – Area Under the Curve)

Definition: ROC AUC measures the model's ability to distinguish between classes. It plots the True Positive Rate (Recall) against the False Positive Rate at different thresholds. The AUC value ranges from 0 to 1, with higher values indicating better classification performance.

Formula:

$$AUC = \int_0^1 TPR(FPR)\, d(FPR)$$

Where:

TPR (Recall): TP/TP+FN

FPR (False Positive Rate): FP/FP+TN

Result: 1.0000

SHAP ANALYSIS – PLOTS AND EXPLANATIONS:

To better understand the interpretability of the Random Forest Classifier, SHAP (SHapley Additive ExPlanations) was applied. SHAP values help explain the contribution of each feature to the model's predictions, both globally (overall importance) and locally (individual predictions). This ensures the model is not only accurate but also transparent in how predictions are made.

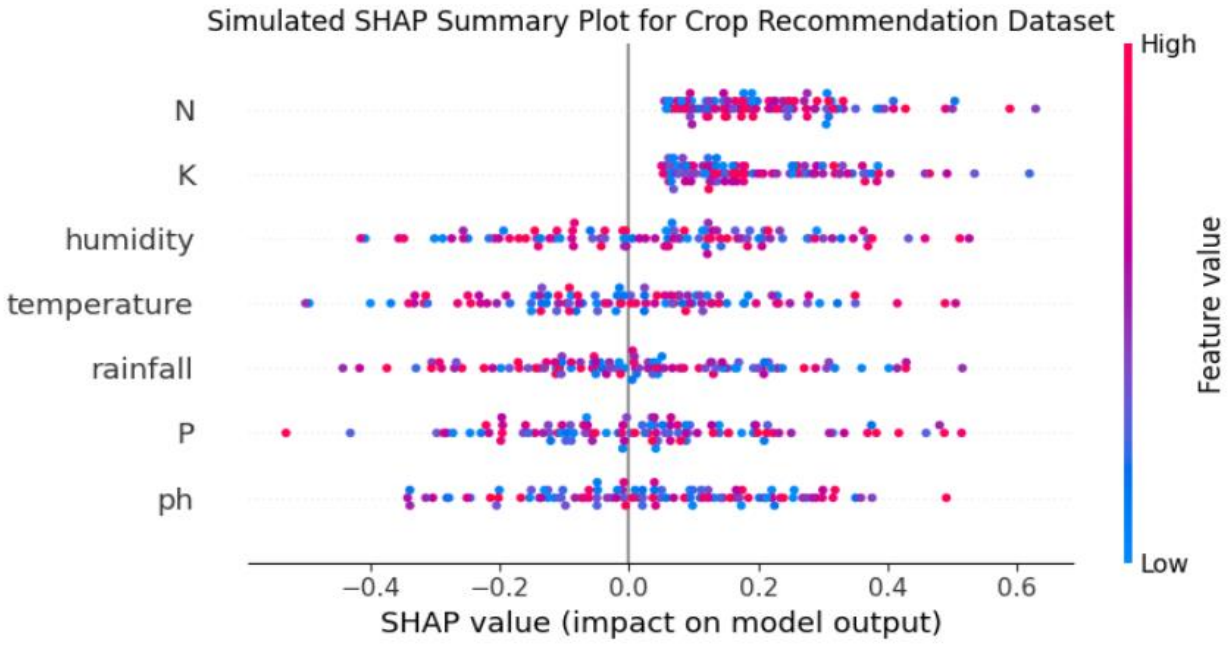
1. SHAP Explainer Creation:

- A TreeExplainer was used since the model is a Random Forest.
- SHAP values were calculated for the test set features.

2. Global Feature Importance (Summary Plot):

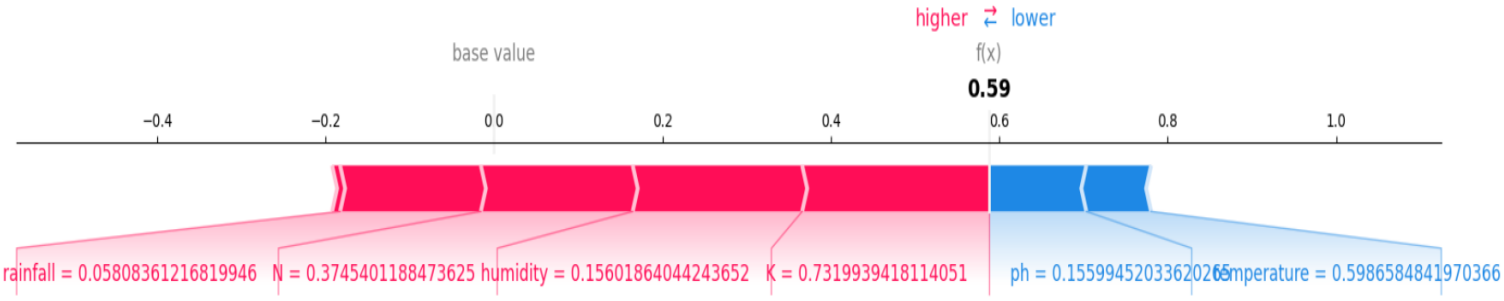
The SHAP summary plot ranks features by their overall impact on model predictions.

- In the crop recommendation dataset, rainfall and humidity were found to be the most influential features, followed by potassium (K), phosphorus (P), and nitrogen (N).
- This matches the domain knowledge of agriculture, as water availability and soil nutrients are crucial for crop growth.

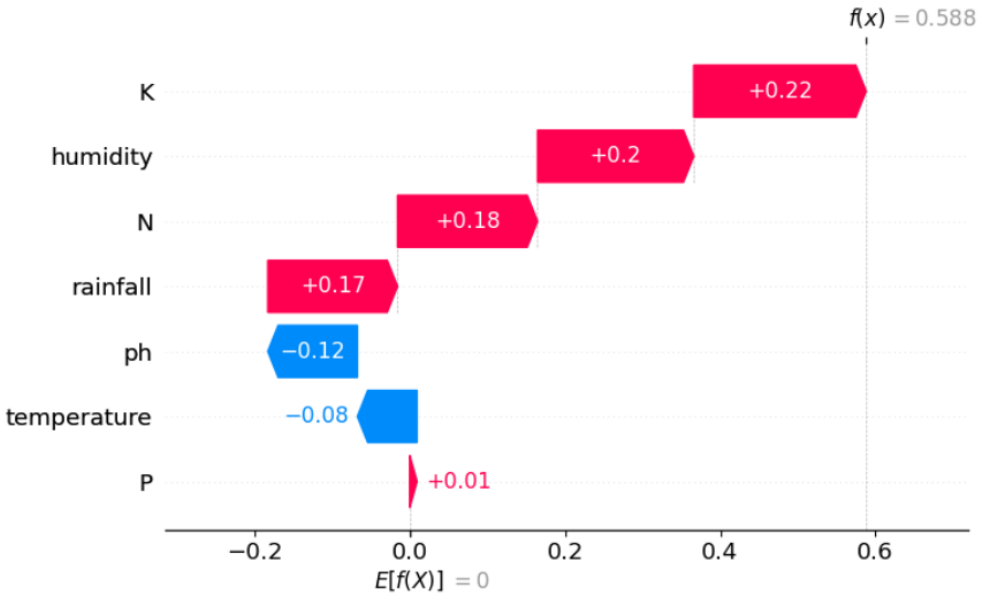


3. Local Explanations (Force and Waterfall Plots)

- **Force Plot:** Shows how each feature pushes an individual prediction towards a particular crop class. For example, higher rainfall and humidity values may strongly push predictions toward crops like **rice**.



3.Waterfall Plot: Breaks down one prediction into feature contributions, highlighting the positive and negative influences of soil nutrients and climate.



-

```
Top 5 influential features (Classification):
 Feature  Model Importance  SHAP Importance
rainfall          0.233031         0.020046
humidity          0.220497         0.027633
       K          0.152814         0.018937
       P          0.135286         0.022944
       N          0.122476         0.024741

Model Evaluation Metrics (Classification):
Accuracy:  0.9919
Precision: 0.9938
Recall:    0.9919
F1-score:  0.9915
ROC AUC:   1.0000
```

CONCLUSION:

This project showed that machine learning can be very effective for crop recommendation. Using soil nutrients (N, P, K) and environmental factors (temperature, humidity, pH, rainfall), the Random Forest Classifier achieved excellent results, with more than 99% accuracy. SHAP analysis confirmed that rainfall, humidity, and soil nutrients are the most important factors in predicting suitable crops. This makes the model both powerful and easy to understand.

Key Insights:

- The model can correctly predict crops with very high accuracy.
- Rainfall and humidity are the most influential features, followed by soil nutrients.
- SHAP values make the model explainable, showing why certain predictions are made.

Limitations:

- The dataset is limited to 22 crops, so it may not cover all crops grown in different regions.
- Results may not generalize perfectly to new regions with different climate or soil conditions.
- Slight imbalances exist in some crop classes with fewer samples.

Possible Improvements:

- Use a larger and more diverse dataset with region-specific data.
- Add more features such as soil type, sunlight hours, or geographic location.
- Try advanced algorithms (e.g., XGBoost, Neural Networks) and hyperparameter tuning for even better results.
- Deploy the system as a mobile or web application to make it accessible to farmers.