INTEL UNNATI INDUSTRIAL TRAINING

PROJECT REPORT

ON

PROBLEM STATEMENT :

"Knowledge Representation and Insights Generation from Structured Dataset"

Institution: Baba Farid College of Engineering and Technology.

Submitted By: Team ByteSmiths

Team Members:

- Vishawjeet Singh
- Manjot Kaur
- Parmeet Kaur
- Arshdeep Singh
- Ratanveer Singh

ACKNOWLEDGEMENT

The successful completion of this project marks the beginning of an ongoing learning experience, transforming ideas and concepts into practical applications. This project has been a significant learning journey, enhancing our confidence to work in a professional setting. The experience gained will undoubtedly lead to bright prospects in the future.

We would like to extend our gratitude to our mentor Er. Gursewak Singh for the opportunity to work under their guidance, which has not only expanded our knowledge of the latest fields but also taught us the importance of team building.

With a deep sense of gratitude, We express our sincere thanks to the Intel Unnati Industrial Training team for their active support and continuous guidance, without which it would have been challenging to complete this project. I also appreciate my team members for their collaboration and dedication, as well as our peers for their valuable feedback and support throughout this endeavor

.

Team:
ByteSmiths

INTRODUCTION

In today's data-centric world, organizations are flooded with vast amounts of structured data. The challenge lies not just in storing this data but in extracting meaningful insights that can drive decision-making processes. This project, titled "**Knowledge Representation and Insight Generation from Structured Datasets**," addresses this challenge by developing an AI-based solution capable of effectively analyzing and interpreting structured data.

Problem Statement

Our project aims to develop an AI-based solution that addresses this challenge by automating the analysis of structured datasets. This solution will represent knowledge effectively and generate insights, providing organizations with a powerful tool for data-driven decision-making and enhancing their ability to respond to evolving business needs.

Objective

The primary objective of this project is to create a solution that harnesses the power of artificial intelligence to process and analyze structured data. This solution aims to:

1. **Represent Knowledge:** Use advanced techniques to structure and represent data in a way that highlights critical information and relationships.
2. **Generate Insights:** Analyze the data to identify patterns, trends, and anomalies, providing valuable insights that are not easily recognized through manual analysis.
3. **Aid Decision-Making:** Present the generated insights in a user-friendly manner, enabling stakeholders to make informed decisions based on accurate and comprehensive data analysis.

DATASET DESCRIPTION

Source: https://www.kaggle.com/datasets/varshitanalluri/crop-recommendation-dataset

About Dataset :

CROP RECOMMENDATION

INTRODUCTION

In modern agriculture, the precise recommendation of crops plays a crucialrole in achieving optimal yield and sustainability. As data-driven approaches become more prevalent, the importance of utilizing comprehensive datasets, especially those related to soil composition, becomes increasingly apparent. The dataset in focus includes extensive information on essential nutrients such as Nitrogen, Phosphorus, and Potassium, as well as environmental factors like Temperature, Humidity, pH levels, and Rainfall. Analyzing this dataset is vital for making informed decisions that can significantly enhance agricultural productivity, resource management, and crop health.

Features

The dataset contains several attributes that are crucial for determining the optimal crop for given conditions:

- **Nitrogen (N)**: The amount of nitrogen in the soil, measured in kg per hectare.
- **Phosphorus (P)**: The amount of phosphorus in the soil.
- **Potassium (K)**: The amount of potassium in the soil.
- **Temperature**: The temperature in degrees Celsius.
- **Humidity**: The percentage of humidity in the air.
- **pH_Value**: The pH level of the soil, indicating its acidity or alkalinity.
- **Rainfall**: The amount of rainfall in mm.

Preprocessing Steps:

Before applying machine learning algorithms, the dataset underwent the following preprocessing steps:

- **Data Cleaning**: Ensured there were no missing values or duplicate records in the dataset.
- **Exploratory Data Analysis (EDA)**: Performed EDA to understand the distribution and relationships between the features.
- **Label Encoding**: Converted the categorical crop names into numerical labels for model training.
- **Feature Scaling**: Applied RobustScaler to normalize the features, ensuring consistent ranges for model input.

Example Data Entry:

Here is an example of a single data entry from the dataset:

- **Nitrogen (kg/ha)**: 80
- **Phosphorus (kg/ha)**: 40
- **Potassium (kg/ha)**: 40
- **Temperature (Celsius)**: 30
- **Humidity (percentage)**: 82
- **pH_Value**: 6
- **Rainfall (mm)**: 200

Recommended Crop: Rice

This dataset provides a comprehensive foundation for developing our AI-based crop recommendation system, enabling precise and data-driven agricultural practices.

METHODOLOGY

Methods Used:

- **Data Collection and Cleaning:**
  - **Dataset Source:** The dataset used is the Crop Recommendation Dataset from Kaggle.
  - **Data Cleaning:** The dataset was checked for missing values and duplicates, but none were found, so no imputation or duplicate removal was needed.
- **Exploratory Data Analysis (EDA):**
  - **Unique Values and Dataset Information:** The dataset was examined for unique values, general information, and statistical summary.
  - **Visualization:** Various visualizations were used, including histograms, correlation heatmaps, crop distribution plots, and boxplots for each feature to understand the data better.
- **Knowledge Representation:**
  - **Data Visualization:** The data was visualized using histograms, correlation heatmaps, and bar plots to represent the distribution and relationships among different features.
- **Preprocessing:**
  - **Label Encoding:** The crop names were encoded into numerical values using LabelEncoder.
  - **Data Normalization:** The features were normalized using RobustScaler to remove the median and scale according to the Inter-Quartile Range.
- **Pattern Identification:**
  - **Model Training:** A Random Forest Classifier was used to train the model on the dataset.
  - **Model Evaluation:** The model's performance was evaluated using accuracy score, classification report, confusion matrix, and cross-validation scores.

- **Hyperparameter Tuning:**
  - **Randomized Search CV:** Hyperparameters were tuned using RandomizedSearchCV to optimize the model's performance.
- **Insights Generation:**
  - **Feature Importance:** The importance of each feature was determined using the feature_importances_ attribute of the Random Forest model and visualized using bar plots.
- **Predictive System:**
  - **Crop Prediction:** A predictive system was implemented to recommend the best crop based on input features.
  - **Requirements Suggestion:** The system can also suggest the optimal range of requirements for any given crop.

Tools:

- **Python:** Used for data manipulation, analysis, and implementation of machine learning models.
- **Pandas:** Employed for loading and preprocessing data, making it easy to handle and analyze.
- **NumPy:** Utilized for performing numerical operations and managing arrays, essential for data manipulation.
- **Matplotlib and Seaborn:** Libraries used for creating visual representations of data to understand patterns and insights.
- **Scikit-learn:** Provides tools for preprocessing data, training machine Learning models, evaluating their performance, and tuning their parameters.
- **Google Colab:** An online platform used for writing and executing code, offering collaboration and access to powerful computing resources.
- **Pickle:** A module used to save and load trained machine learning models, enabling reuse without retraining.
- **Next.js:** Utilized for frontend development, providing a seamless user experience.
- **Flask:** Used for backend development, handling data processing and server-side logic.
- **Vercel:** A platform used for deploying the frontend of the website, ensuring it is accessible and performs well.
- **Gemini AI:** Used for generating insights from data, transforming raw data into meaningful information.

RESULTS

The findings from our data analysis reveal significant patterns and relationships within the dataset. By applying various algorithms for knowledge representation and pattern identification, we have summarized our insights through visual representations. Key results include the optimal conditions for various crops, as well as the accuracy and effectiveness of our predictive model.

Visualizations:

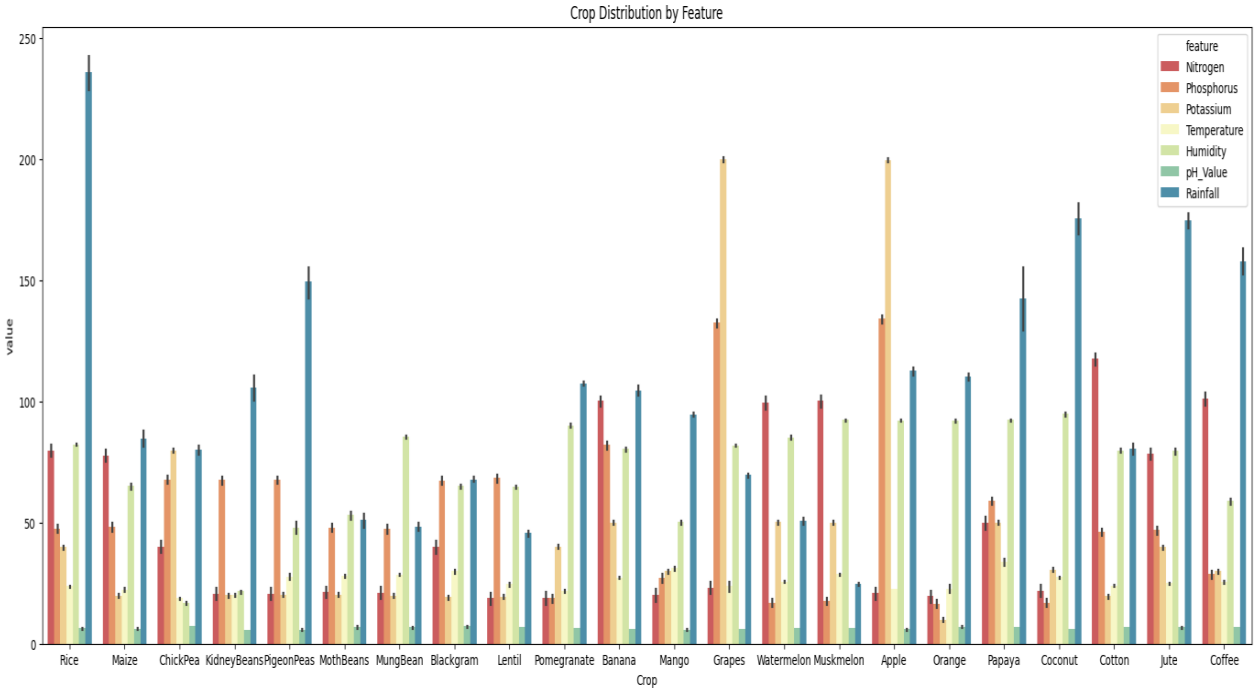
To interpret the results effectively, several charts and graphs were used:

- **Histograms:** Illustrated the distribution of each feature across the dataset.
- **Boxplots:** Showed the spread and variability of features for differentcrops.
- **Correlation Heatmap:** Highlighted relationships between features, helping in feature selection.
- **Crop Distribution Barplots:** Visualized the optimal conditions for each crop.
- **Confusion Matrix:** Evaluated the classification model's performance by comparing true vs. predicted values.
- **Feature Importance Chart:** Displayed the significance of each feature in the prediction model.

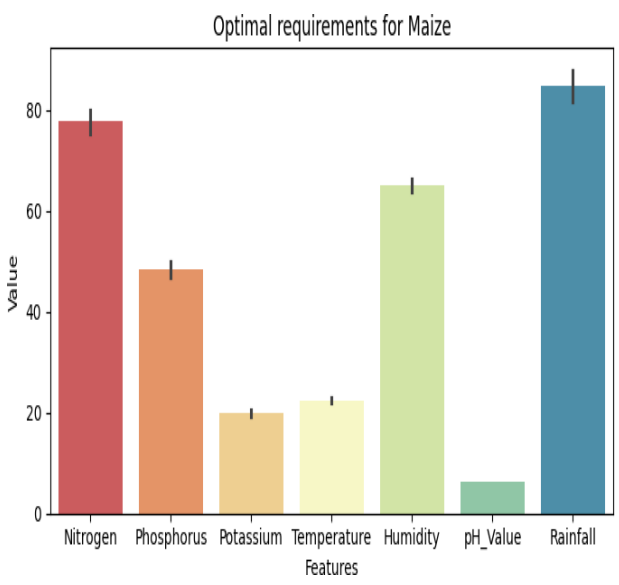
Insights Generated

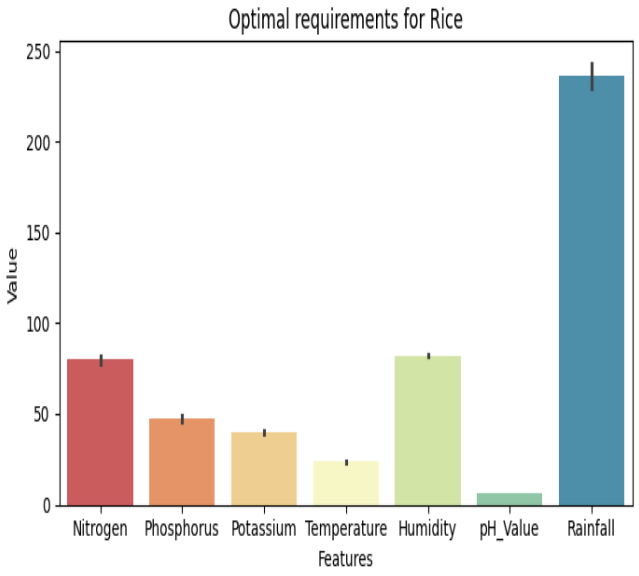
Based on the analysis and the model predictions, several valuable insights weregenerated:

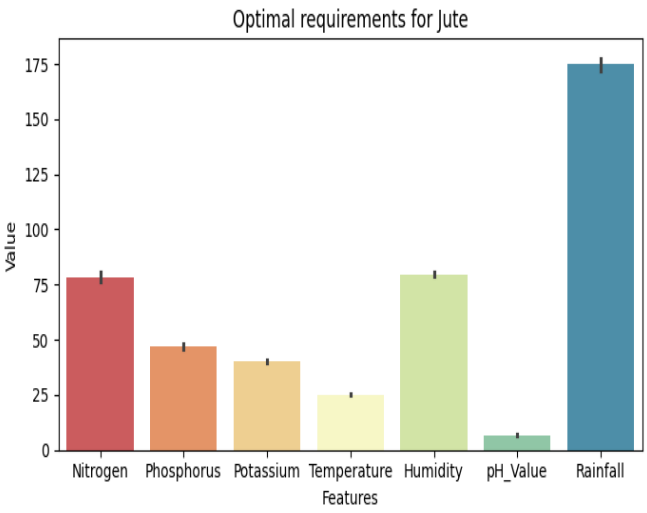
- **Key Features for Crop Prediction:**
    - **Nitrogen, Phosphorus, and Potassium** levels are critical for determining the suitability of a crop.
    - **Temperature, Humidity, pH Value, and Rainfall** also play significant roles in crop growth and yield.

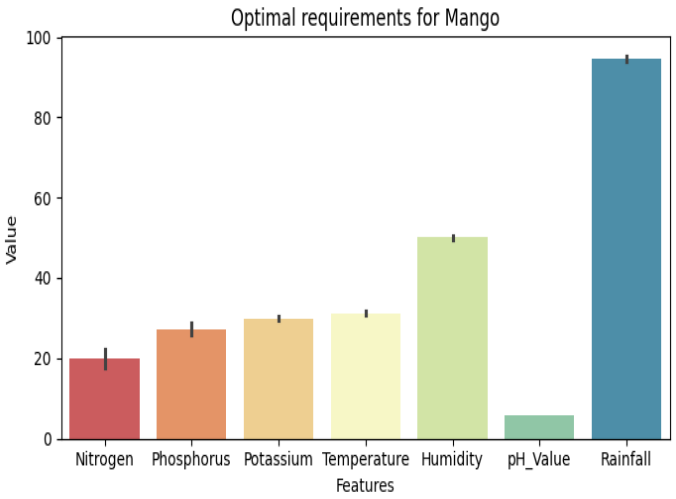


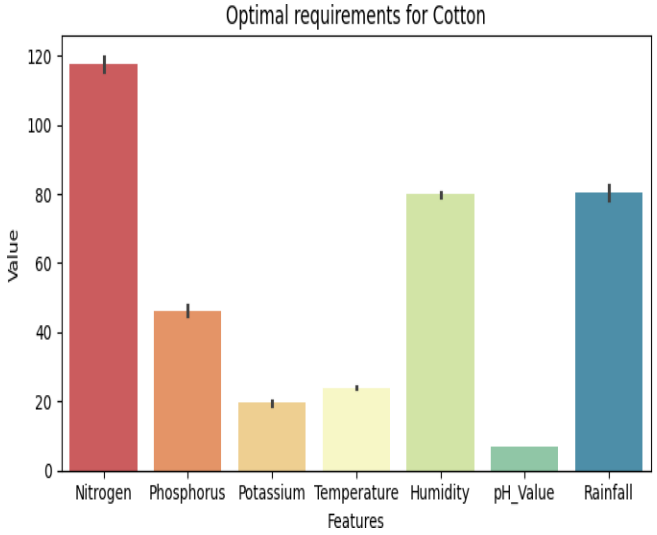
- **Optimal Conditions for Each Crop:**
  - The model identifies and recommends the best crops to cultivate based onthe current environmental conditions.
  - Each crop has specific optimal ranges for the features, such as nitrogen levels, temperature, and pH, which are crucial for maximizing yield.

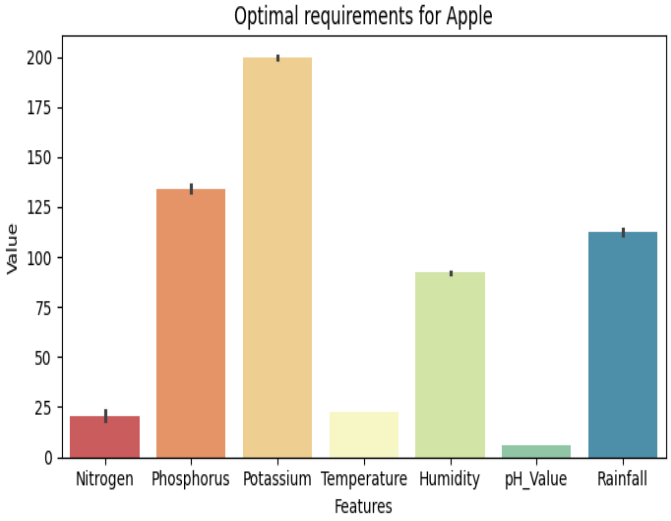


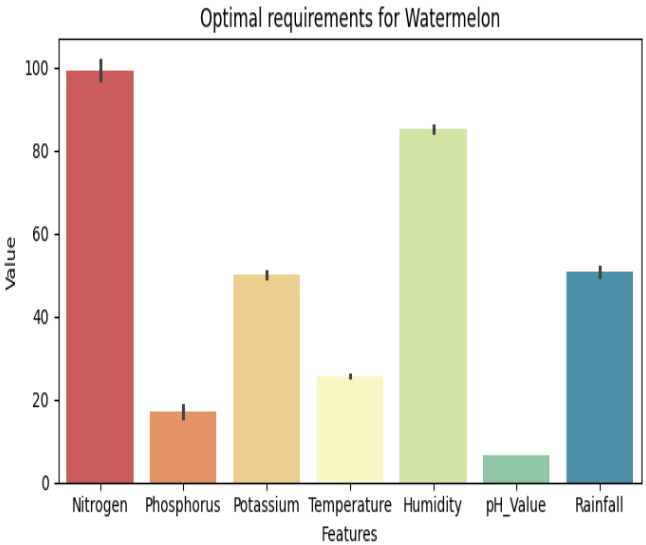


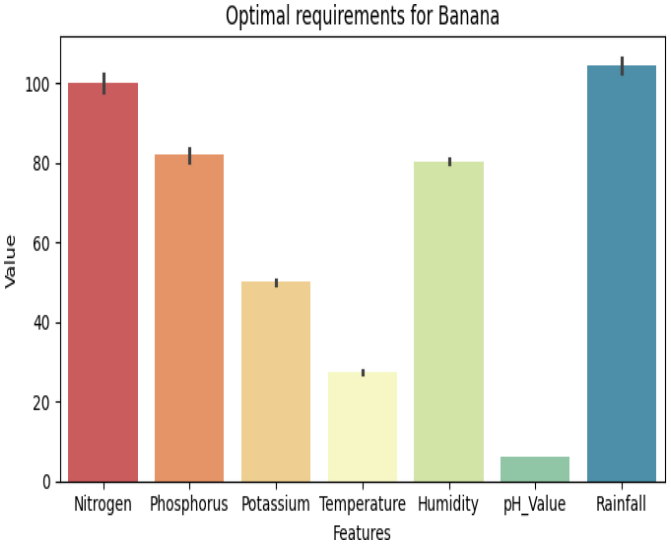

Optimal requirements for Jute
Value
Features
Nitrogen
Phosphorus
Potassium
Temperature
Humidity
pH_Value
Rainfall


Optimal requirements for Mango
Value
Features
Nitrogen
Phosphorus
Potassium
Temperature
Humidity
pH_Value
Rainfall


Optimal requirements for Cotton
Value
Features
Nitrogen
Phosphorus
Potassium
Temperature
Humidity
pH_Value
Rainfall

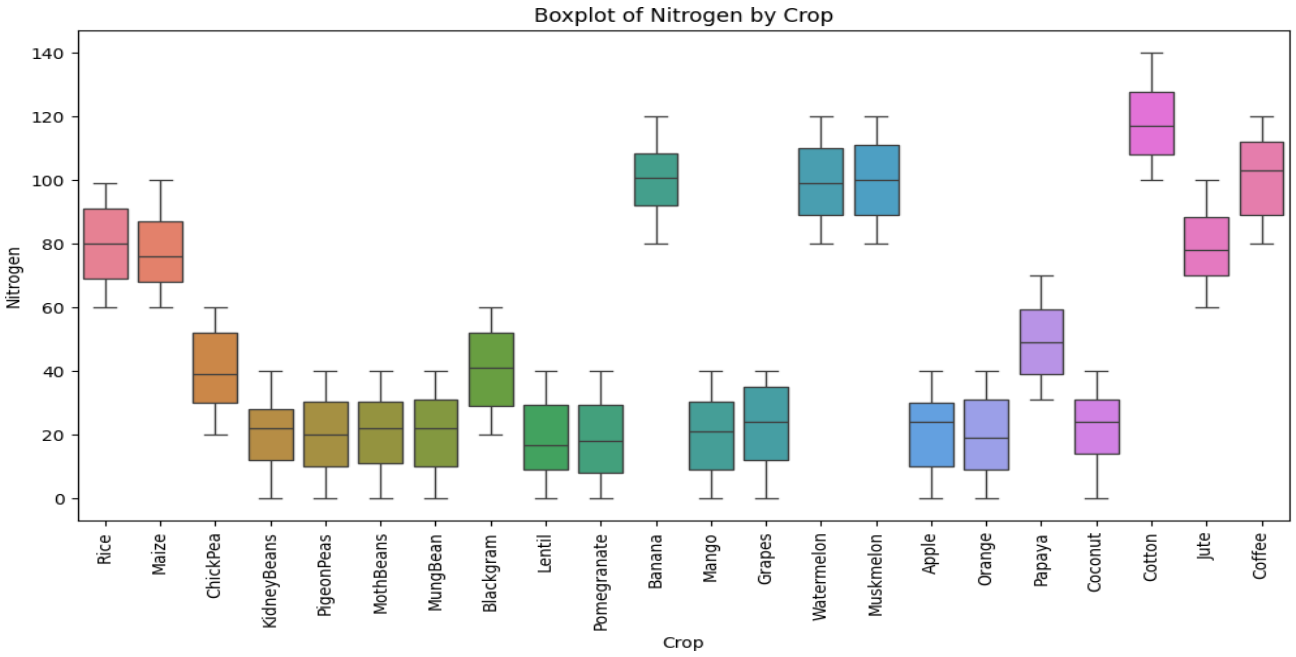

Optimal requirements for Apple
Value
Features
Nitrogen
Phosphorus
Potassium
Temperature
Humidity
pH_Value
Rainfall

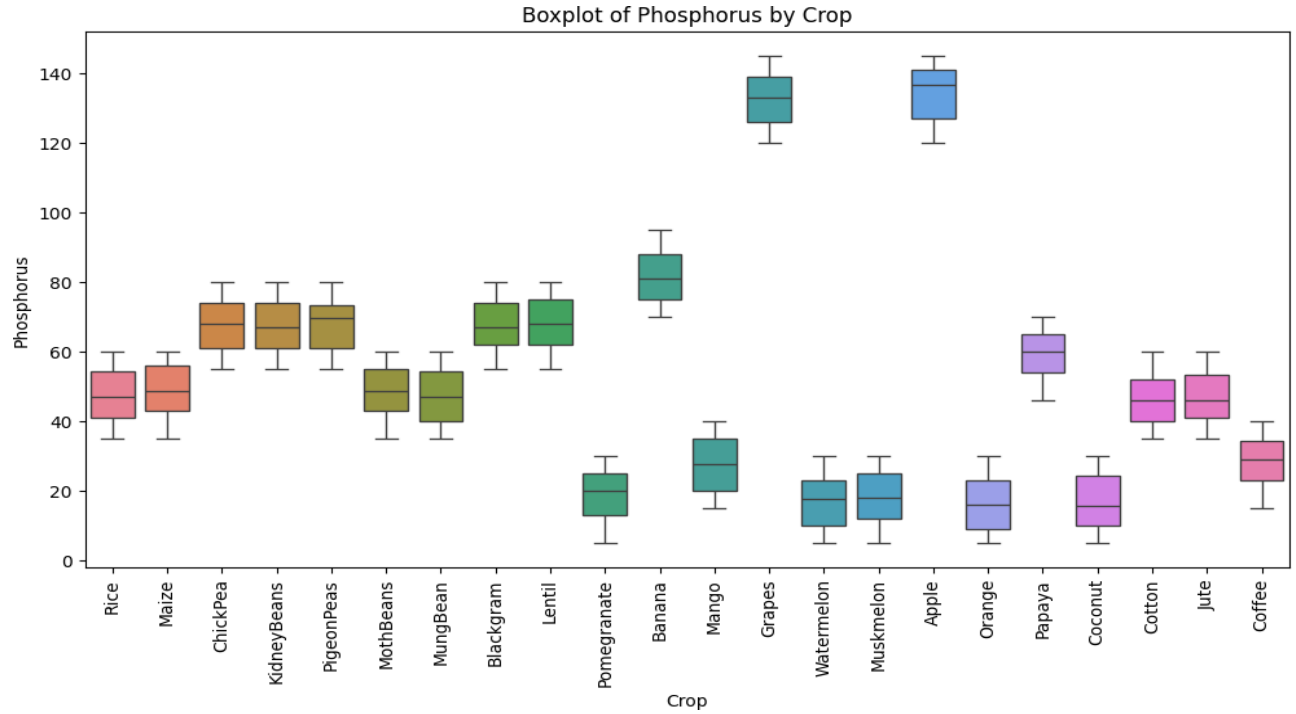

Optimal requirements for Watermelon
Value
Features
Nitrogen
Phosphorus
Potassium
Temperature
Humidity
pH_Value
Rainfall

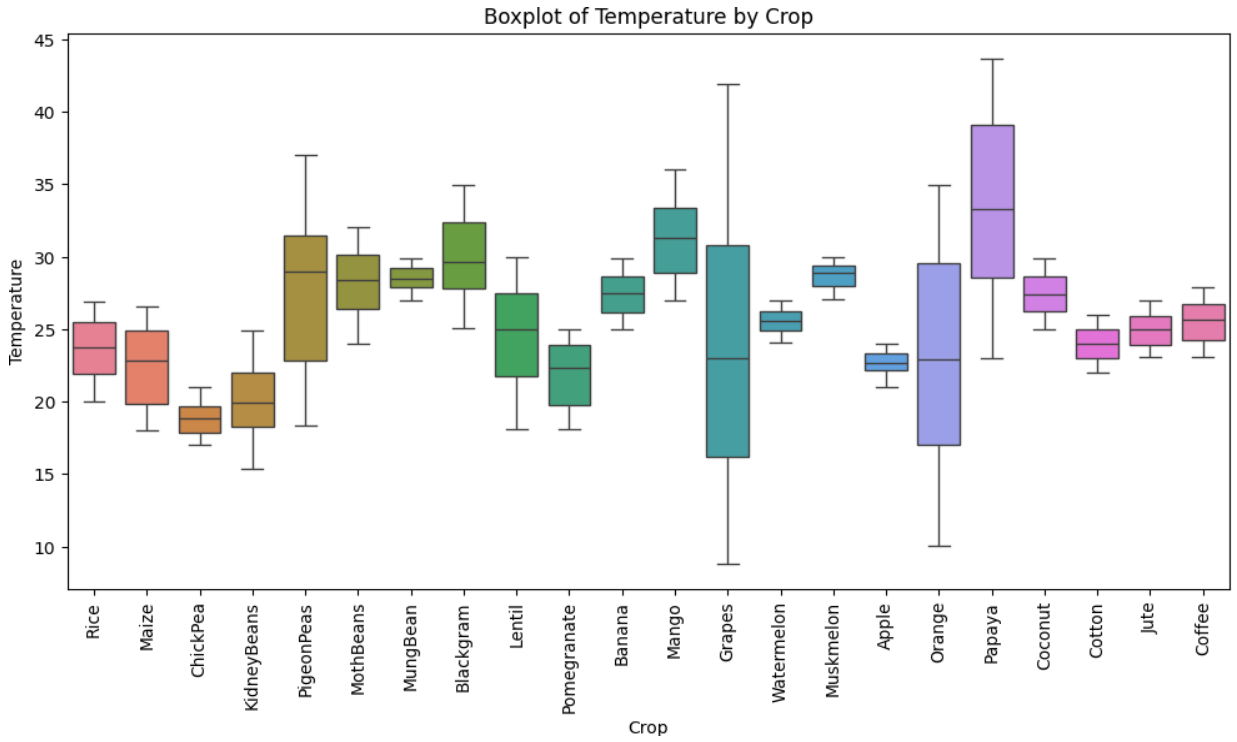

Optimal requirements for Banana
Value
Features
Nitrogen
Phosphorus
Potassium
Temperature
Humidity
pH_Value
Rainfall

- **Boxplots:**

  Showed the spread and variability of features for different crops.







- **Recommendations for Farmers:**
    - Optimal Crop Selection: The system recommended crops that are best suited to the given soil and environmental conditions, ensuring higher yield and sustainability.
    - Nutrient Optimization: It provided specific recommendations on the optimal levels of Nitrogen, Phosphorus, and Potassium for different crops, helping to enhance soil fertility and crop growth.
    - Performance Metrics**:** The system achieved a high accuracy rate in predicting the best crops for specific conditions, demonstrating the effectiveness of using AI in agricultural decision-making.

.

CONCLUSION

- The project successfully developed an AI-based solution for knowledge representation and insights generation using the Crop Recommendationdataset from Kaggle.
- Random Forest Classifier was employed, achieving high accuracy in predicting suitable crops.
- Key features such as Nitrogen, Phosphorus, and Temperature wereidentified as critical for crop growth.
- Thorough data cleaning, exploratory data analysis, and visualization wereconducted to understand patterns within the dataset.
- Preprocessing steps including label encoding and robust scaling ensureddata normalization.
- Hyperparameter tuning optimized the model's performance.
- Actionable insights were generated to aid farmers in selecting suitablecrops based on soil and climatic conditions.
- The project also included the development of a user interface (UI) where users can input basic details about soil conditions , temperature, and other relevant factors. This interface provides useful insights and recommendations on which crops are recommended for their specific conditions.
- Future work includes refining the model, incorporating real-time data,expanding the dataset, and exploring additional machine learning algorithms for improved accuracy.