Project Title: Crop Prediction

1. Introduction

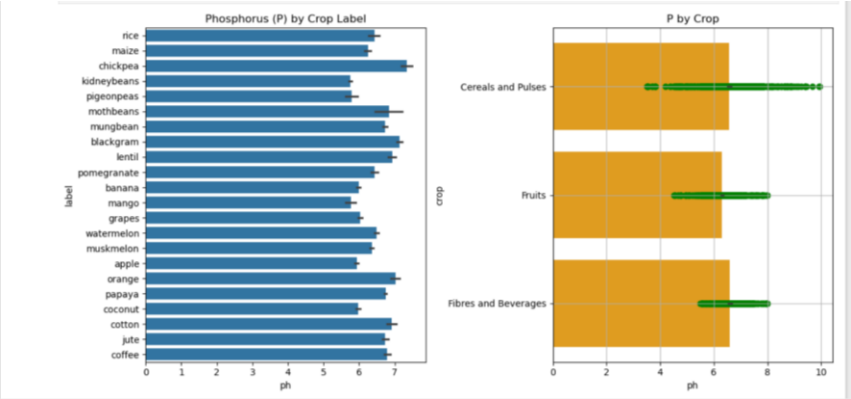
- Agriculture is a primary economic driver in India.
- More than 50% of the population depends on agriculture.
- The agriculture sector's contribution to India's GDP is less than that of other sectors.
- The Indian government has introduced initiatives to support farmers through various schemes and loans.
- Agriculture depends heavily on natural factors.
- Many farmers still grow traditional crops, unaware of the impact of global warming and climate change.
- This project aims to develop a model to predict crop types based on natural factors.

2. Data Exploration

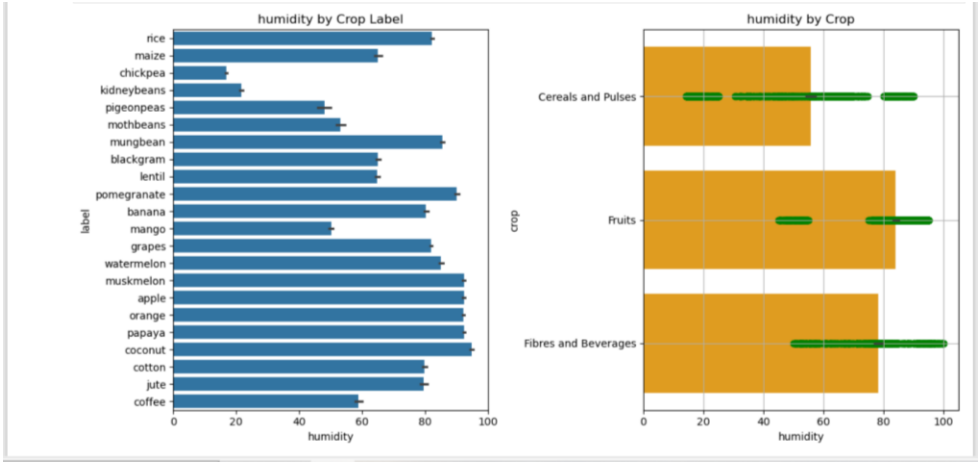
- The dataset used is "Crop_recommendation.csv".
- The dataset contains 2200 rows and 8 columns.
- The columns are:
    - N (Nitrogen)
    - P (Phosphorus)
    - K (Potassium)
    - Temperature
    - Humidity
    - pH
    - Rainfall
    - Label (Crop type)
- Data types of the columns:
    - 4 float columns
    - 3 int columns
    - 1 object column
- There are no missing values in the dataset.
- There are no duplicate rows.
- The "label" column has 22 unique values, representing different crop types.

3. Data Visualization and Analysis

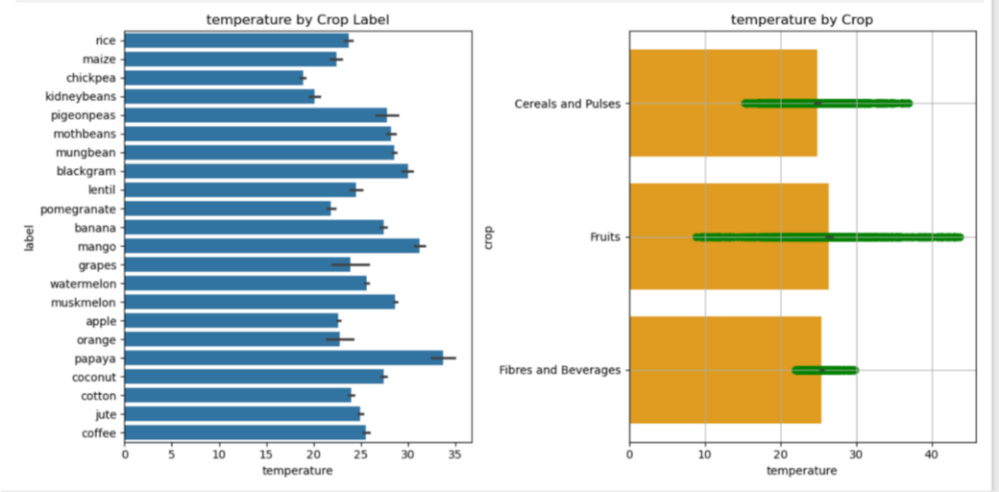
- The project includes visualizations to analyze the relationship between natural factors and crop types.
- Visualizations include:
    - Bar plots and scatter plots to show the relationship between:
        - Phosphorus (P) vs. Crop Label and Crop Category



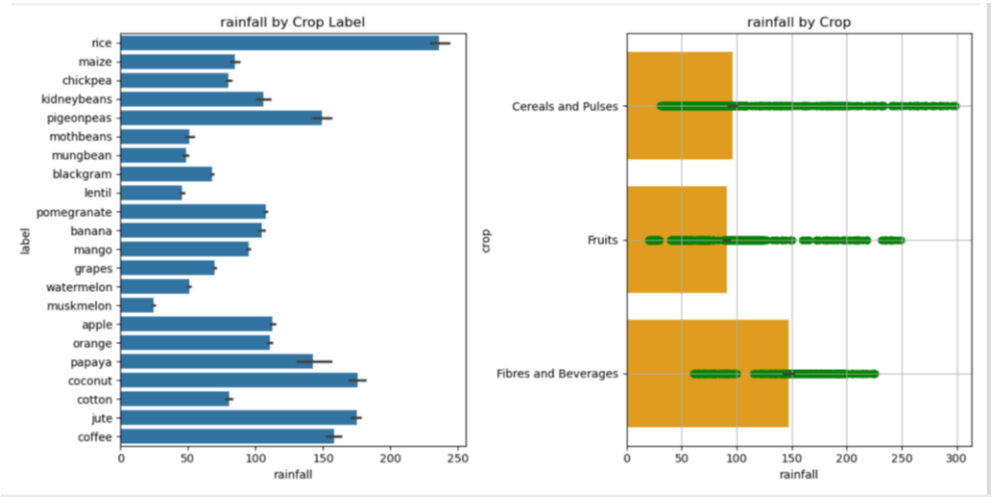
- Humidity vs. Crop Label and Crop Category



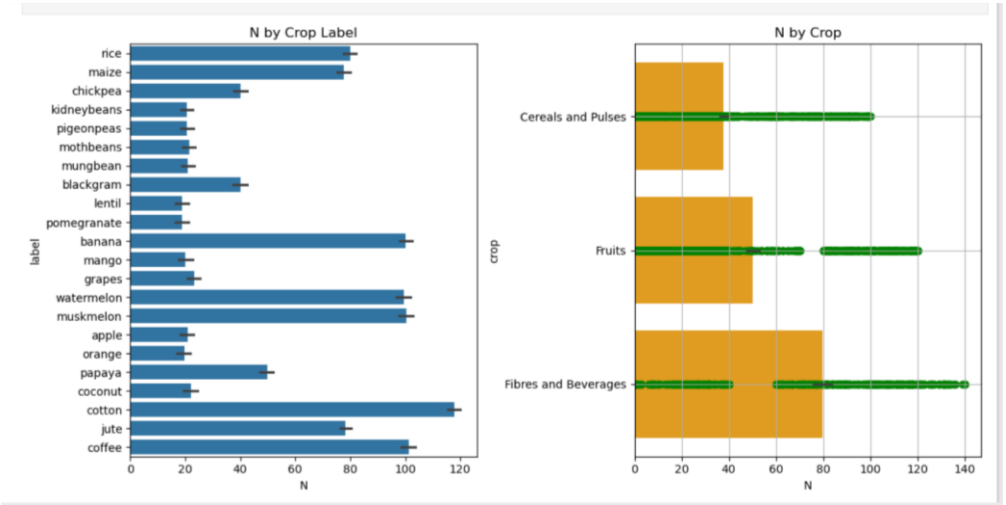
- Temperature vs. Crop Label and Crop Category



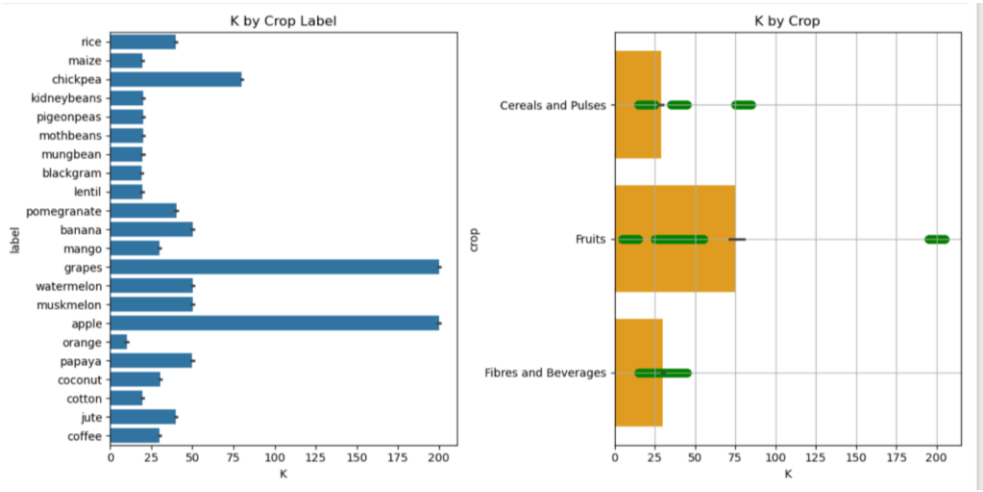
- Rainfall vs. Crop Label and Crop Category



- Nitrogen (N) vs. Crop Label and Crop Category



- Potassium (K) vs. Crop Label and Crop Category



  - Box plots to visualize the distribution of numerical features (N, P, K, temperature, humidity, pH, rainfall) and detect outliers.

4. Feature Engineering

- A new column, "crop", is created based on the "label" column.
- The "crop" column categorizes crops into three groups:
  - Cereals and Pulses
  - Fruits
  - Fibres and Beverages
- Crop column Label Encoded to 0 ,1 ,2

5. Handling Class Imbalance

- The distribution of values in the "crop" column is:
  - Cereals and Pulses: 900
  - Fruits: 900
  - Fibres and Beverages: 400

The original dataset has an imbalance in the distribution of crop types, with "Cereals and Pulses" and "Fruits" having significantly more samples than "Fibres and Beverages."

Oversampling was applied to address this imbalance.

The "Fibres and Beverages" category was oversampled to have an equal number of samples as the other two categories.

- Cereals and Pulses: 900
- Fruits: 900
- Fibres and Beverages: 900

6.Model Selection

- Machine Learning model used Logistic Regression , Decision Tree which provided accuracy_score 0.9893939393939394.
- Hyper Parameter Tuning with criterion ,max_depth, min_samples_leaf, min_samples_split,random_state .
- Performance Metrices evaluated with Confusion Matrix , Accuracy Score , Classification Report.

7. Conclusion and Future Work

- **Summary of Findings:**
- **Nutrient Levels (N, P, K):**
  - The bar plots for Nitrogen (N), Phosphorus (P), and Potassium (K) levels, shows the specific requirements of these nutrients for different crops. Crops have distinct needs for each nutrient, and these help in understanding the optimal fertilization strategies.
- **Environmental Factors (Temperature, Humidity, Rainfall):**
  - The temperature, humidity, and rainfall clearly illustrate the varying climatic needs of different crops. These visuals emphasize that crops like rice thrive in high rainfall areas, while others like cotton prefer drier conditions.

- **Limitations:**
  - As Natural Resources keep on changing over time and unusual factors like climate change , global warming , heat waves which are volatile makes it less predictable .

8. Code Repository:

https://github.com/Nikhil-G08/Crop-Prediction-Model

7. Summary

The project is designed to assist farmers in determining the most suitable crops to cultivate by analyzing natural factors. This is particularly important because while a significant portion of India's population relies on agriculture, the sector's contribution to the GDP is disproportionately low. Additionally, many farmers continue to grow the same crops their ancestors did, despite changes in climate and environmental conditions. The model aims to address these challenges by providing data-driven recommendations.