

**Summer-2024**

**Project Report**

**Course Code : CSE303**

**Course Title : Statistics for Data Science**

**Section : 03**

**Group Number : 04**

**Submitted by:**

|  |  |
| --- | --- |
| Student ID | Student Name |
| Al-Imam Uddin | **2021-3-60-260** |
| Sifat Noor Siam | **2022-1-60-368** |
| Tasnim Saima Raita | **2022-1-60-291** |

**TABLE OF CONTENTS**

|  |  |
| --- | --- |
| Content | Page |
| 1. Introduction | 3 |
| 2. Dataset description | 3 |
| 3. Data Preprocessing | 3 |
| 4. Exploratory Data Analysis (EDA) | 4-10 |
| 5. Machine learning model | 11-12 |
| 6. Conclusion | 12 |

# Introduction

This dataset is related to crop recommendation based on several environmental factors such as nitrogen (N), phosphorus (P), potassium (K) content in the soil, temperature, humidity, pH, and rainfall. This analysis aims to identify patterns and insights that can assist in recommending the most suitable crops based on these environmental parameters. The ultimate target class is ‘level’ column data of the dataset to improve decision-making in agriculture, helping farmers choose crops that match the conditions of their land and climate.

2. Dataset Description

The dataset consists of 2,200 rows and 8 columns. Each row represents a different environmental scenario with the following features:

* **N**: Nitrogen content in the soil (integer).
* **P**: Phosphorus content in the soil (integer).
* **K**: Potassium content in the soil (integer).
* **Temperature**: The average temperature in degrees Celsius (float).
* **Humidity**: The average relative humidity in percentage (float).
* **pH**: The pH value of the soil (float).
* **Rainfall**: The average rainfall in mm (float).
* **Label**: The recommended crop type (categorical, object).

The dataset contains numerical values for all features except the label, which is categorical and represents different types of crops.

# 3. Data Preprocessing

For the analysis, no missing values were found in the dataset as all columns contain complete data. Steps taken to clean and prepare the dataset:

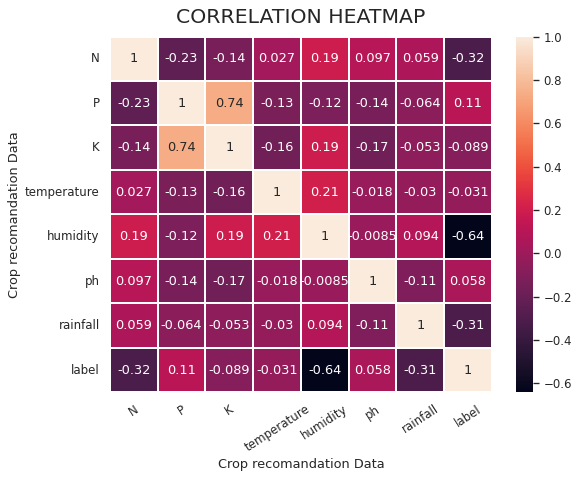
* No need for any step taken since all values were present.
* The categorical values in the **label** column represent different crops, and these need to be encoded into numerical values.

For exploration and visualizations, the dataset column **label** value was replaced with a numerical value in its current form.

# 4. Exploratory Data Analysis (EDA):

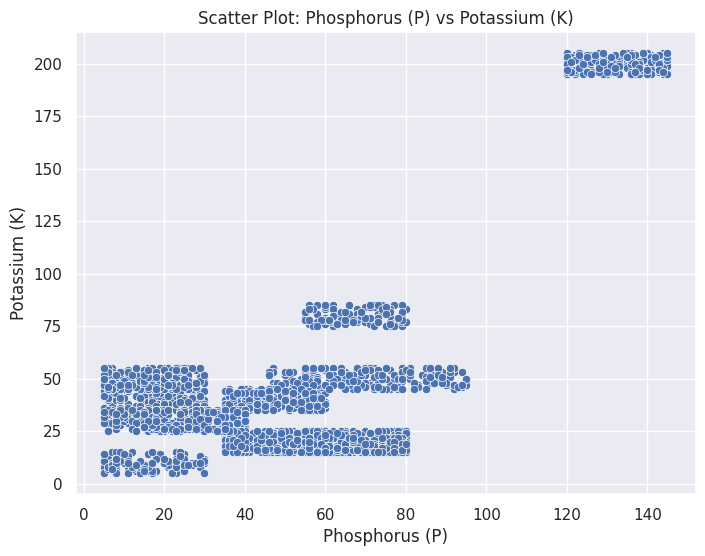
Let's explore the data to discover relationships and insights. We will:

* Visualize the distribution of environmental factors (temperature, pH, etc.).
* Investigate correlations between these factors.
* Explore the crops concerning different environmental conditions.



**1.** **Scatter Plot: Phosphorus (P) vs Potassium (K):**

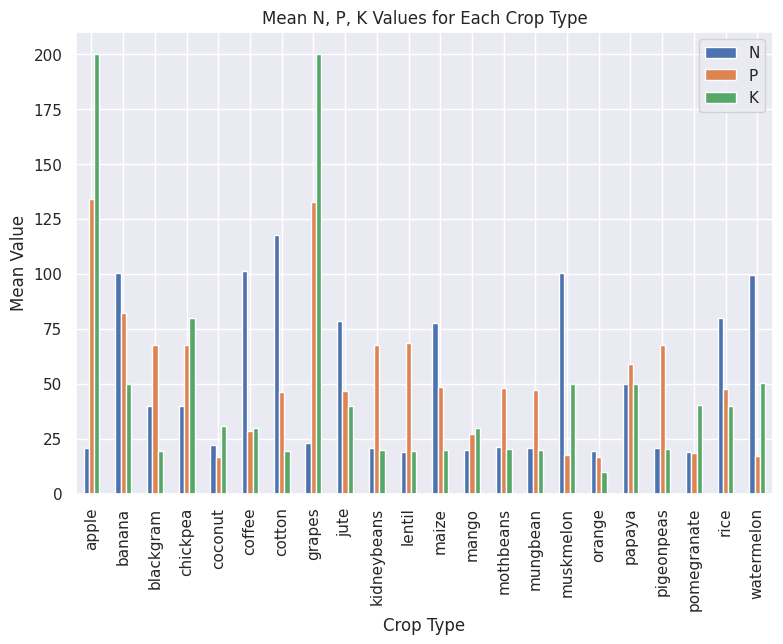
 Phosphorus (P) and Potassium (K) show a strong positive correlation (0.74), indicating they tend to increase together.



The scatter plot of Phosphorus (P) versus Potassium (K) reveals distinct clustering patterns. Most data points are concentrated in two major clusters. The first and largest cluster occurs when Phosphorus values range from 0 to 50, with Potassium values varying between 0 and 100. The second, smaller cluster is centered around a Phosphorus value of approximately 140, with Potassium values between 75 and 100. This suggests that there may be distinct groupings or categories within the data where specific combinations of P and K are common. Additionally, the Potassium values largely remain under 100, while Phosphorus has a more widely dispersed range, indicating some variability between the two elements.

1. **Bar chart for mean N, P, K values by crop type**:

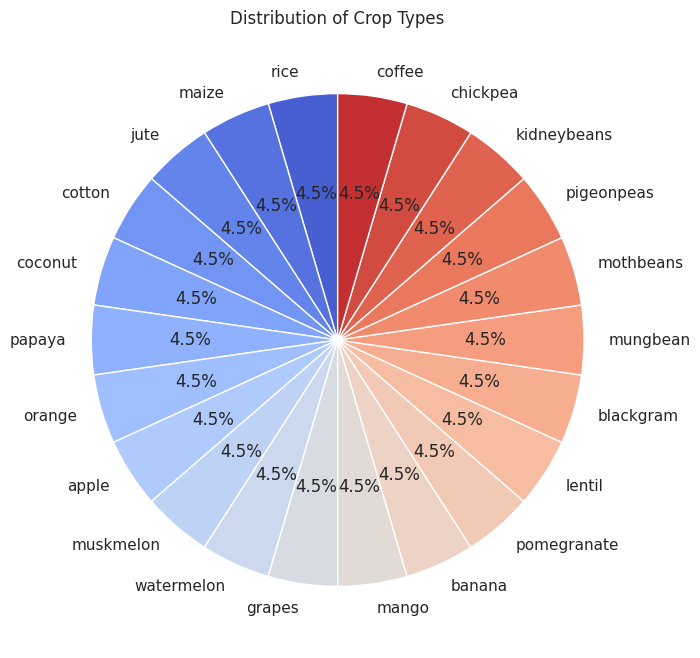
A bar chart for the mean **N, P, K** values grouped by **label (crop type)**, to show which crops prefer specific nutrient levels.



The bar chart compares the mean values of Nitrogen (N), Phosphorus (P), and Potassium (K) across various crop types. The chart shows that the levels of N, P, and K vary significantly by crop. For most crops, Nitrogen levels tend to be higher compared to Phosphorus and Potassium. Notably, crops like grapes and cotton have particularly high Nitrogen values, while watermelon exhibits elevated Potassium levels. Phosphorus shows more moderate and consistent values across crops. This indicates that different crops have varying nutrient needs, with some requiring substantially more of certain nutrients than others.

1. **Pie chart for crop type distribution**:

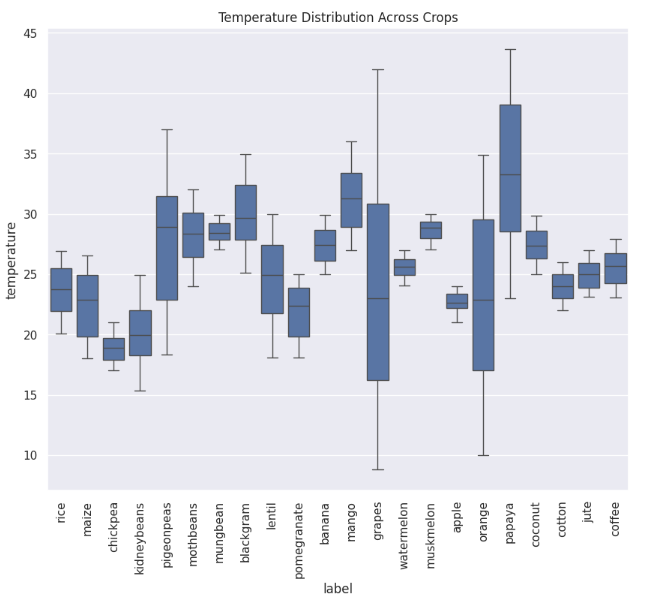
* A pie chart showing the distribution of different crop types in the dataset.



The pie chart displays an equal distribution of crop types, with each of the 22 crops represented occupying 4.5% of the total.

1. **Box plot for temperature across crop types**:

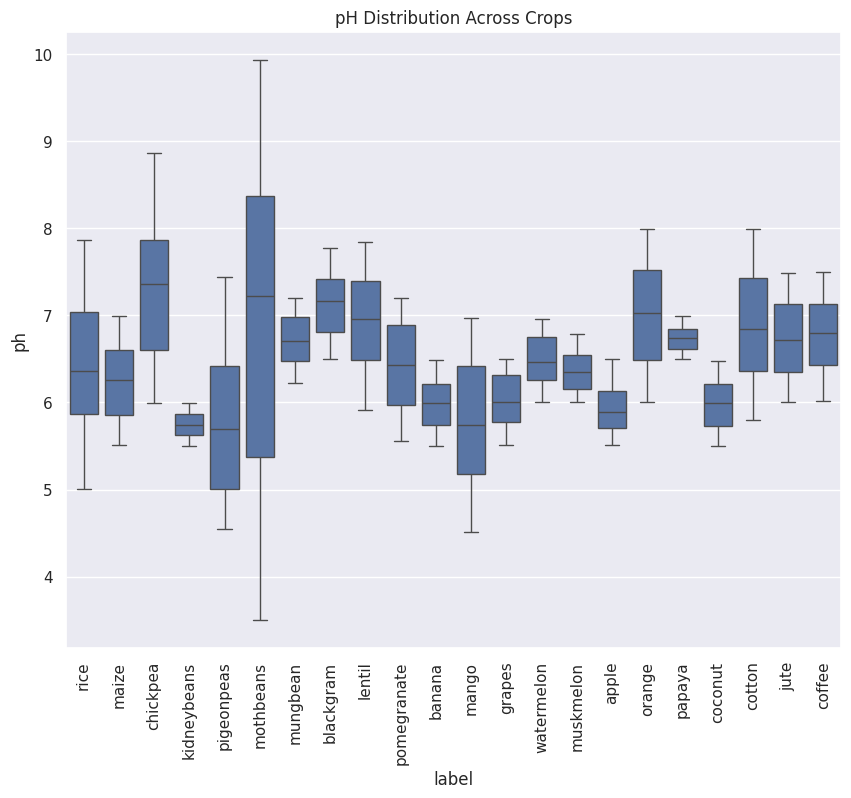
Box plots for variables like **temperature** and **label** across different crops



This box plot illustrates the temperature distribution across various crop types. The crops show significant variation in their temperature ranges, with some like watermelon and coconut displaying wide temperature spans. Others, such as muskmelon and apple, have narrower ranges, indicating more specific temperature requirements. Most crops fall within a temperature range of 20-35°C, with median temperatures clustered around 25-30°C.

1. **Box plot for pH across crop types**:

Box plots for variables like **pH** and **label** across different crops

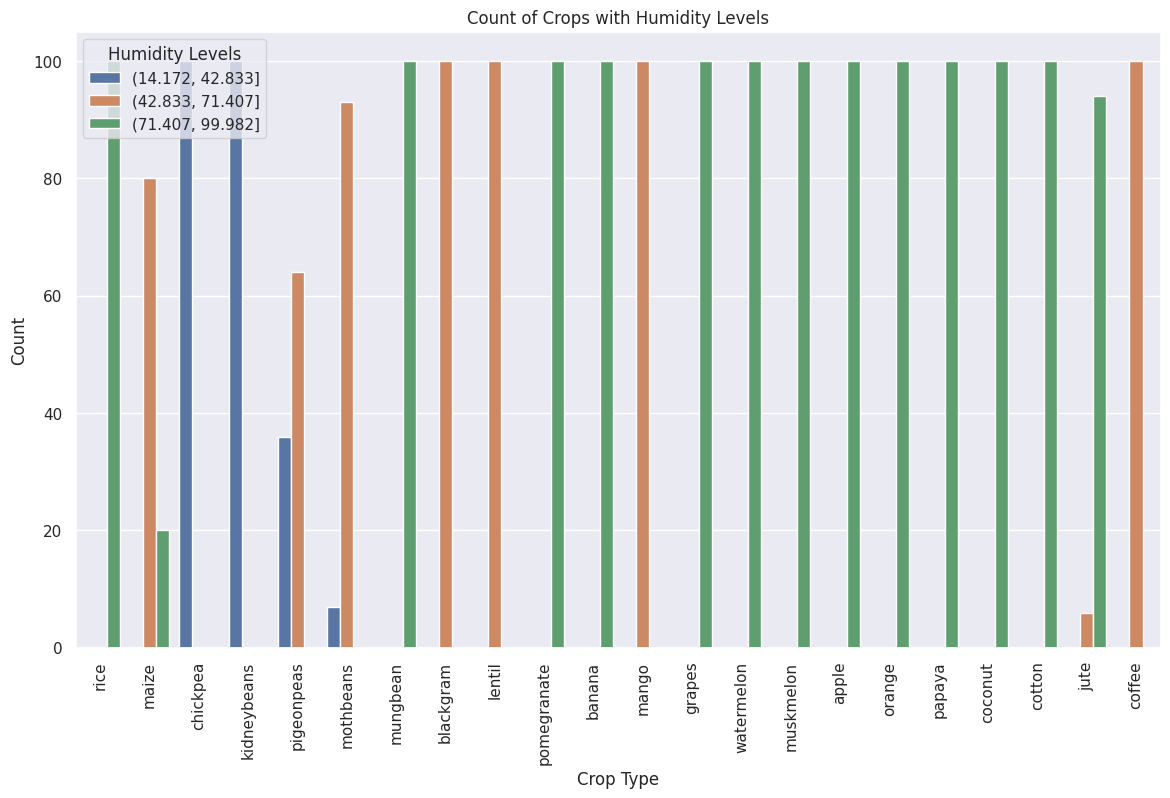


The box plot illustrates pH distribution across various crop types, revealing diverse soil preferences. Most crops favor slightly acidic to neutral conditions, with median pH values between 6 and 7. Some crops like pigeonpeas and mothbeans demonstrate wide pH tolerances, while others such as muskmelon and apple show narrower ranges. Notably, coffee, jute, and coconut lean towards higher pH levels. Conversely, rice and maize have lower median pH values. The overall pH range spans from about 4.5 to 10, highlighting the varied soil requirements in agriculture.

**6. Generate the count plot of humidity levels:**

We divide the humidity into three ranges, which helps to differentiate how many crops fall within each humidity level.

The x-axis represents the crop types, and the y-axis shows the count of observations for each crop.



This count plot illustrates the distribution of humidity levels across various crop types. The humidity levels are categorized into three ranges: low (14.172-42.833), medium (42.833-71.407), and high (71.407-99.982). Most crops show a preference for medium to high humidity levels, with many exclusively falling in the highest humidity range. Rice and maize exhibit the most diverse humidity requirements, spanning all three categories. Crops like watermelon, muskmelon, apple, and several others consistently prefer high humidity environments. Interestingly, some crops such as chickpea and kidneybean show a preference for lower humidity levels.

# Machine Learning Models

In this project, we used the **Logistic Regression** model to predict our target variable (label).

**Purpose of the Logistic Regression Model:**

The Logistic Regression model is used primarily for binary classification tasks, where the goal is to predict one of two possible outcomes. It estimates the probability that a given input belongs to a specific class (usually coded as 0 or 1) using a logistic function (sigmoid curve). Since our target class values are categorical, i.e., not discreet, for that we chose this model.

**Advantages of Logistic Regression:**

* Does not require scaling for binary features
* Provides probabilistic outputs
* Efficient for small to medium-sized datasets

**Disadvantages of Logistic Regression:**

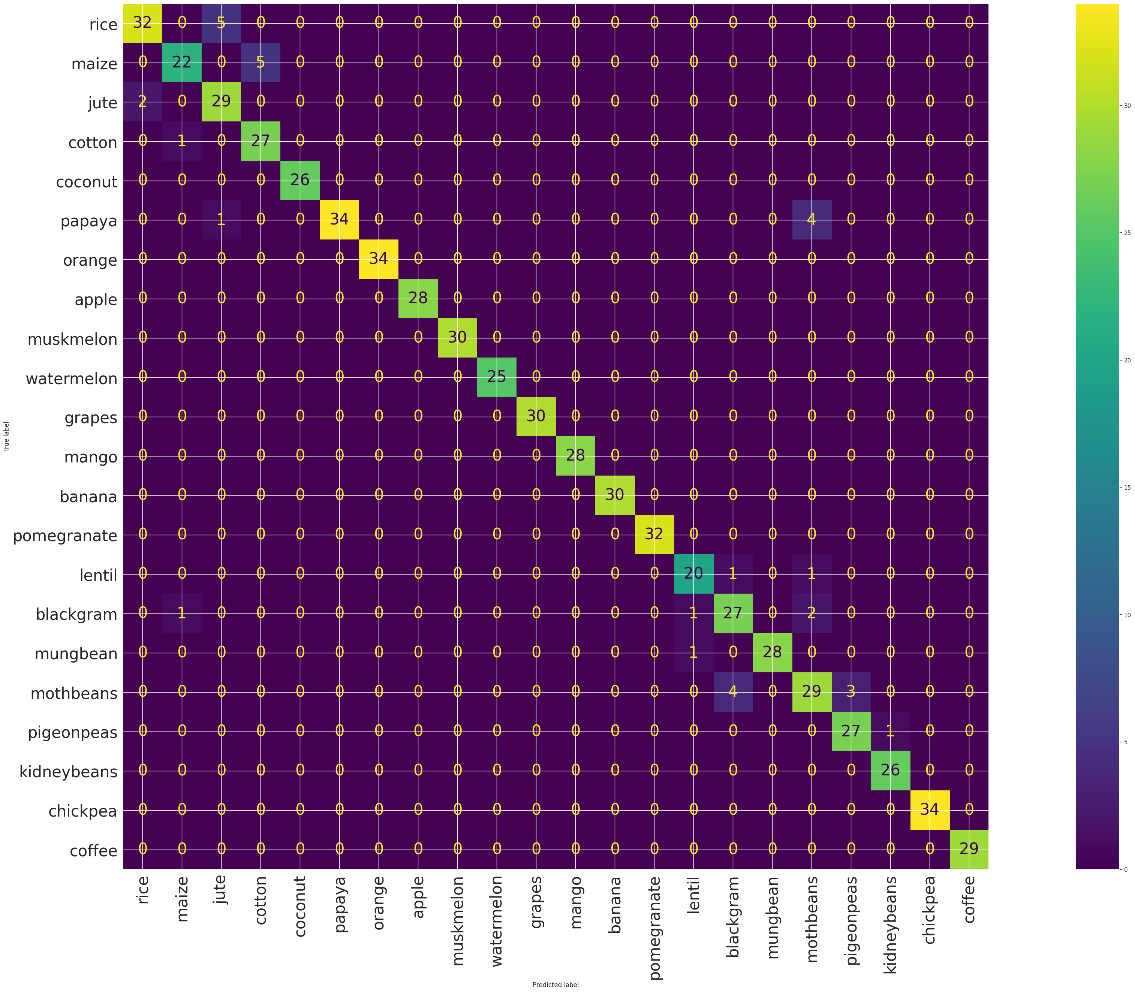
* Limited ability to capture non-linear relationships
* Sensitive to outliers

**Results:**

By running this model in our dataset, the prediction was well performed as we get the predictions accuracy 0.95 that indicates 95% prediction results are correct.

**F1 score result:** After training this model to evaluate its performance we get same result of 95% that previously we got in model accuracy.

**Confusion Matrix:**



In this confusion matrix, The diagonal values indecate that the model predicts correctly and the Off-diagonal values indicate misclassifications. For example, 1 instance of maize was misclassified as jute. Overall, the model exhibits strong performance in classifying all the crops.

5. Conclusion:

The crop recommendation dataset shows that the success of a crop is strongly linked to key factors like soil nutrients (nitrogen, phosphorus, and potassium), temperature, humidity, soil pH, and rainfall. Each crop grows best under specific conditions, meaning these factors vary depending on the type of crop. For example, rice grows well in areas with high humidity and heavy rainfall, while other crops may need different nutrient levels and weather conditions. Using this data helps make better crop recommendations, supporting sustainable farming by matching crops to the environment.