**Abdul Rehman**

**2024(S)-MSDS-11**

**Advance Techniques**

**Crop Recommendation System**

**Catalog**

[Problem Statement 2](#_Toc184655687)

[Significance 2](#_Toc184655688)

[Objectives 2](#_Toc184655689)

[Potential Impact 3](#_Toc184655690)

[Key Questions 3](#_Toc184655691)

[Data Collection 4](#_Toc184655692)

[Data pre-processing 5](#_Toc184655693)

[Data Cleaning: 5](#_Toc184655694)

[Feature Engineering: 5](#_Toc184655695)

[Exploratory Data Analysis (EDA) 6](#_Toc184655696)

[Descriptive analysis 6](#_Toc184655697)

[Distribution Plots 7](#_Toc184655698)

[Outlier Detection 8](#_Toc184655699)

[Correlation Analysis 10](#_Toc184655700)

[Predictive Analysis 11](#_Toc184655701)

The agricultural sector plays a vital role in the global economy, providing food and raw materials for various industries.

# Problem Statement

Farmers often face challenges in choosing the best crops for their land due to the complexity of environmental factors like soil conditions, climate variations, and market demands. These difficulties are further compounded by unpredictable weather patterns and limited access to expert advice, especially for smallholder farmers in developing countries. As a result, they may experience poor crop yields, inefficient resource use, and higher susceptibility to crop failures caused by climate change.

# Significance

The development of a **data-driven crop recommendation system** is crucial for addressing these challenges in modern agriculture. Such systems:

* **Improved Productivity:** By analyzing soil and environmental data, farmers can make well-informed decisions that result in better yields and profits.
* **Climate Adaptation:** It helps farmers adjust to changing weather patterns by providing insights into which crops are more likely to thrive in specific conditions.
* **Sustainability:** This approach reduces waste, whether it’s water, fertilizers, or other resources, while minimizing environmental harm.
* **Support for Farmers:** Small-scale farmers gain access to affordable tools that can guide them in optimizing their crop choices, enhancing their livelihoods and food security.

# Objectives

The primary objectives of this project are:

1. To develop a model capable of accurately recommending crops based on environmental conditions such as soil pH, temperature, rainfall, and humidity.
2. To enhance prediction accuracy by engineering meaningful features like Thermal Humidity Index (THI) and Water Availability Index (WAI).
3. To explore and compare machine learning algorithms like Decision Trees to identify the most effective method for crop recommendation.
4. To empower smallholder farmers with actionable and accessible insights, enabling them to improve their crop selection and productivity.

# Potential Impact

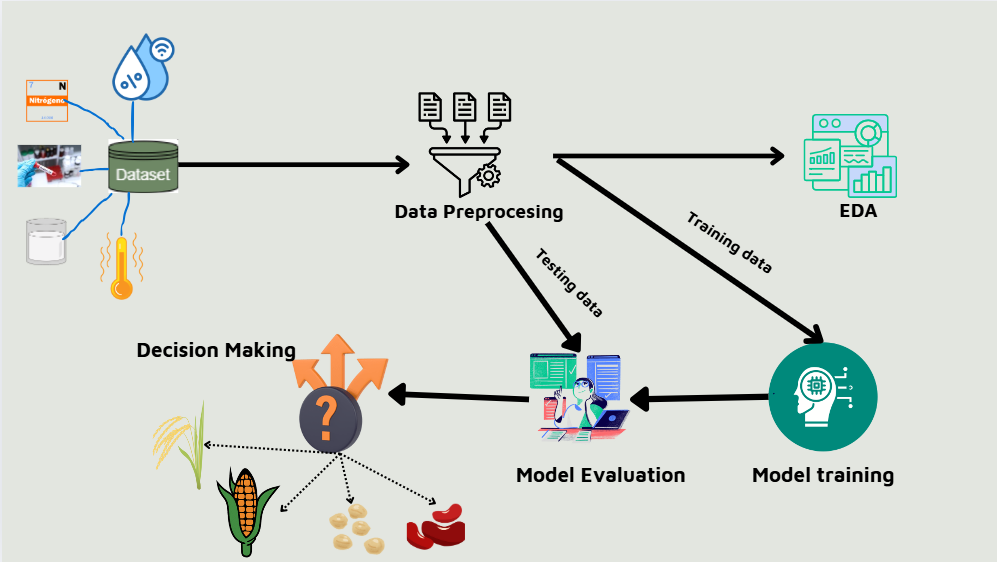
This system has the potential to create a significant positive impact:

1. **Ensuring Food Security:** By increasing crop yields, it helps meet growing food demands and reduce hunger.
2. **Economic Benefits:** Higher productivity can lead to greater profits for farmers, stimulating rural economies.
3. **Promoting Sustainability:** Efficient resource use and reduced environmental impact support long-term agricultural practices.
4. **Building Resilience:** By helping farmers adapt to changing climates, the system reduces the risks associated with crop failures.
5. **Knowledge Accessibility:** It bridges the gap in agricultural expertise, offering farmers data-driven recommendations that are affordable and easy to use.

This project addresses the critical intersection of technology, sustainability, and food security, paving the way for resilient and resource-efficient agricultural practices.

# Key Questions

* How can environmental factors like soil pH, nitrogen levels, temperature, and rainfall be integrated to accurately recommend crops?
* Which machine learning algorithm provides the most reliable predictions for crop suitability?
* How do new features like Thermal Humidity Index (THI) and Water Availability Index (WAI) improve the accuracy of crop recommendations?
* To what extent can the system improve crop yields and profitability for smallholder farmers?



# Data Collection

To address the problem of crop recommendation, a dataset relevant to soil, climate, and crop conditions was obtained from publicly available source Github. The dataset comprises a comprehensive set of features essential for meaningful analysis, including:

* **Soil Properties:** Nitrogen, phosphorus, potassium, and pH levels.
* **Environmental Factors:** Temperature, rainfall, and humidity.
* **Crop Labels:** A classification of crops based on suitability for given conditions.

The dataset contains sufficient observations to ensure robust training and validation of machine learning models. It has undergone preprocessing to handle missing values, duplicates, and inconsistencies, ensuring the integrity and quality of the analysis.

# Data pre-processing

## Data Cleaning:

There are no missing and duplicate values in the dataset.

## Feature Engineering:

Feature engineering is one of the most important steps in operationalizing a machine learning model. Thus, new variables derived from existing variables make it possible to extend the possibilities of the model to capture essential patterns and dependencies. For the crop recommendation system, two key features were added: Relative Humidity and Temperature Humidity Index (THI) and Water Balance Index (WAI).

**Thermal Humidity Index (THI):** The Thermal Humidity Index is a measure that captures the combined effect of temperature and humidity on a system, particularly how they influence the growing conditions for crops.

* Crops are highly sensitive to the combined effects of heat and moisture. THI reflects these interactions, which are often overlooked when considering temperature and humidity separately.
* Including THI helps the model better predict crop suitability under varying climatic conditions.

#### ****Water Availability Index (WAI):**** The Water Availability Index combines two variables rainfall and humidity into a single metric that reflects the availability of water resources for crop growth.

* Water is a critical resource for agriculture. Combining rainfall and humidity provides a comprehensive view of water availability for the soil and atmosphere.
* WAI enhances the model's ability to distinguish crop suitability in regions with varying water resources.

Min-Max Scaling is a data normalization technique that transforms numerical values into a fixed range, typically between 0 and 1. It is used to ensure that all features contribute equally to the model's learning process, preventing features with larger ranges from dominating others. Features like temperature (e.g., ranging from 0°C to 50°C) might dominate smaller-scale features like pH (e.g., ranging from 0 to 14) if left unscaled. Many machine learning algorithms, like gradient descent-based models or distance-based algorithms (e.g., KNN), perform better when input data is on a similar scale. Normalization ensures that the model treats all features equally, rather than giving undue weight to features with larger magnitudes.

# Exploratory Data Analysis (EDA)

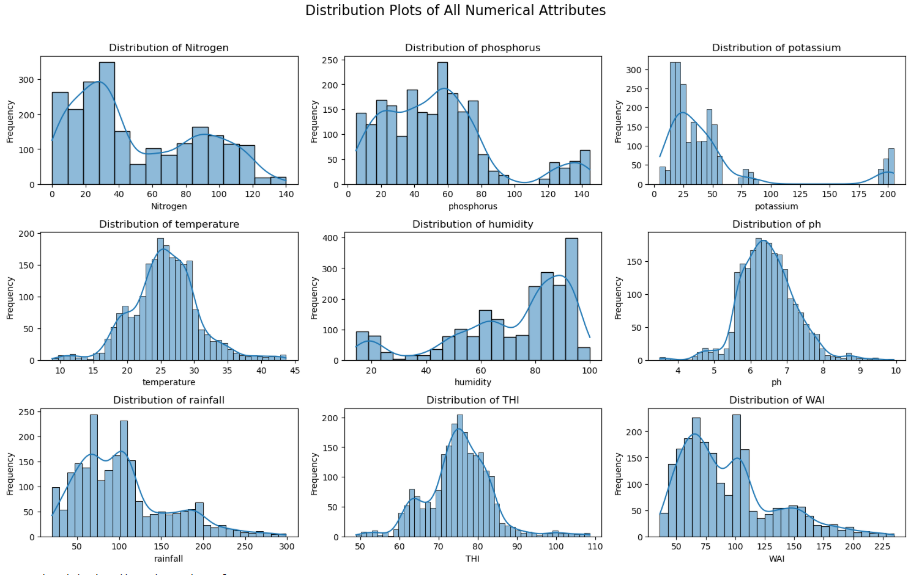
## Descriptive analysis

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **count** | **mean** | **std** | **min** | **25%** | **50%** | **75%** | **max** |
| **Nitrogen** | 2200 | 50.55 | 36.91 | 0 | 21.00 | 37.00 | 84.25 | 140 |
| **phosphorus** | 2200 | 53.36 | 32.98 | 5 | 28.00 | 51.00 | 68.00 | 145 |
| **potassium** | 2200 | 48.14 | 50.64 | 5 | 20.00 | 32.00 | 49.00 | 205 |
| **temperature** | 2200 | 25.61 | 5.063 | 8.82 | 22.76 | 25.59 | 28.561 | 43.67 |
| **humidity** | 2200 | 71.48 | 22.26 | 14.25 | 60.26 | 80.47 | 89.94 | 99.98 |
| **ph** | 2200 | 6.46 | 0.773 | 3.50 | 5.971 | 6.425 | 6.92 | 9.935 |
| **rainfall** | 2200 | 103.46 | 54.95 | 20.21 | 64.55 | 94.867 | 124.26 | 298.56 |
| **THI** | 2200 | 75.14 | 8.393 | 48.83 | 70.95 | 75.318 | 79.99 | 108.59 |
| **WAI** | 2200 | 93.86 | 39.66 | 36.01 | 63.79 | 83.758 | 109.61 | 234.43 |

The Table offers descriptive analysis of important environmental and nutrient indicators. The coefficients of variance of N, P, K are moderate to high, but the highest CV was recorded for potassium. Less variation was observed in temperature, pH as it slightly deviates from neutral pH and temperature fluctuated around 25 degree Celsius. The distribution of humidity and rainfall show moderate fluctuations due to variation in environments. This analysis indicates stability of the Temperature Humidity Index CCT around 75, with fluctuations revealing the differences of heat stress. The Water Availability Index (WAI) shows fluctuation which informs the variation of water conditions in the samples collected. The values obtained capture many aspects of environmental stability food resource availability and variability of indices some of these findings are informative of ecological and agricultural conditions.

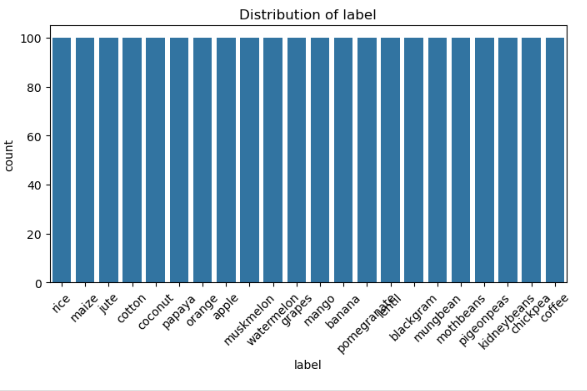
## Distribution Plots

The distribution plots show the normal distribution or skewed distribution of features and provide information about data preprocessing like data transformation or data handling of outliers. The following diagram display the histogram with kernel density estimate plot of all numerical attributes.



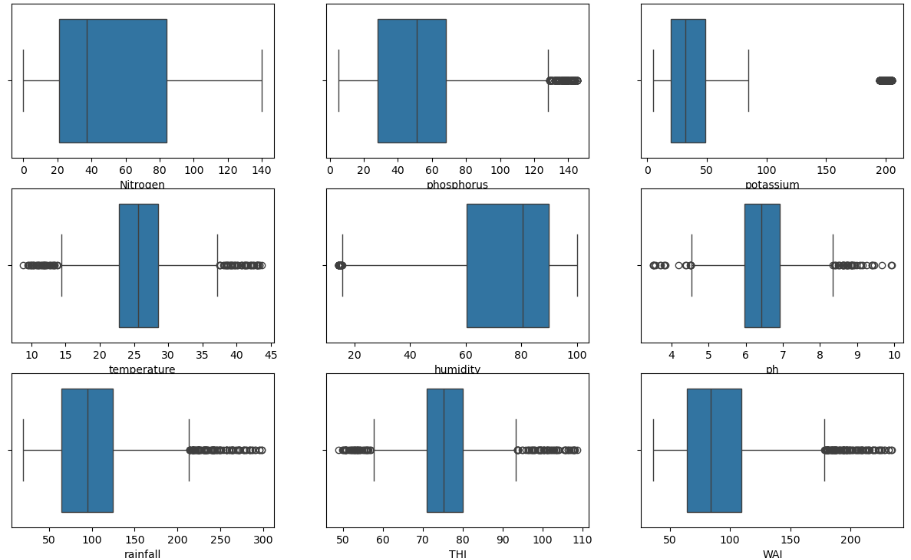
* **Nitrogen** appears to have a right-skewed distribution, meaning there are fewer observations with very high nitrogen values.
* **Phosphorus** and **Potassium** distributions seem more uniform but still exhibit some skewness, with peaks at certain values.
* **Temperature** has a relatively normal distribution, peaking around 25°C, which is common for agricultural data.
* **Humidity** shows a somewhat skewed distribution with a higher concentration of values in the lower range.
* **pH** and **WAI (Water Availability Index)** distributions are more symmetric, with peaks around the middle values.
* **Rainfall** and **THI (Temperature-Humidity Index)** have complex distributions, with spikes at specific values, suggesting that extreme values in weather conditions may be significant.

The bar **chart** displays the distribution of the target variable (label), which represents different crop types (e.g., rice, maize, banana, etc.). The distribution is **even across all crop types**, with each category having roughly the same number of samples (100 for each type). This suggests that the dataset is **balanced** in terms of crop class distribution, which is beneficial for machine learning models, as it helps avoid bias toward any particular class. A balanced dataset helps ensure that the model does not become biased towards any specific crop type, improving its ability to generalize across all classes.



## Outlier Detection

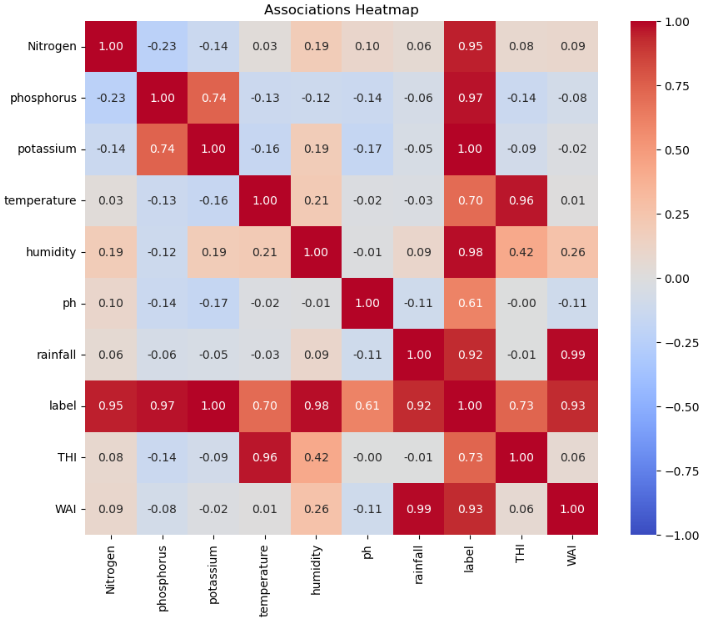
The boxplots provide a summary of the distributions of several numerical features:



* Nitrogen and Phosphorus show distributions with a small number of outliers above the upper whisker, indicating values significantly higher than the bulk of the data.
* Potassium has a more prominent outlier spread, suggesting some extreme high values that deviate from the majority.
* **Temperature** is relatively tightly distributed, with a few outliers on the higher end. Most data points are concentrated between approximately 20–35°C.
* The distribution of humidity shows a wide range of values (from 20–100%), with fewer outliers compared to other variables.
* The pH values are clustered around the middle range (likely neutral to slightly acidic/alkaline), with a few outliers on both ends.
* Rainfall values are spread over a wide range, with significant outliers on the higher end, indicating heavy rainfall data points
* **THI (Temperature-Humidity Index)** has a concentrated range but includes outliers on the higher end, suggesting extreme heat-stress conditions.
* **WAI (Water Availability Index)** shows a compact distribution but highlights some outliers at the upper extreme.

## Correlation Analysis

The heatmap visualizes the correlations between all the features and the target variable (label), using a color scale to represent correlation values.



* **Nitrogen and label (crop type):** Correlation of **0.95**, which indicates a very strong positive relationship between nitrogen levels and the target crop class. This suggests that nitrogen might be a significant factor in predicting the crop type.
* **Phosphorus and label:** Strong positive correlation (**0.97**), indicating its importance as a predictor for the crop type.
* **Humidity and label:** A moderate correlation (**0.98**) indicates that humidity also plays a role in determining the crop type.
* **Rainfall and WAI** show a very strong positive correlation (**0.99**), meaning they are likely representing similar environmental factors related to water availability.
* **THI (Temperature-Humidity Index)** has a moderate correlation with rainfall (**0.73**) and label (**0.73**), suggesting that temperature and humidity combined influence rainfall and, subsequently, the crops.
* There are a few negative correlations, such as between **Nitrogen and Potassium (-0.14)**, **Potassium and Temperature (-0.16)**, and **WAI and Temperature (-0.11)**.

This heatmap reveals key relationships between the features. Features like nitrogen, phosphorus, and humidity show strong positive correlations with the label, making them potentially very important for prediction. The negative correlations suggest areas where changes in one feature could influence others inversely.

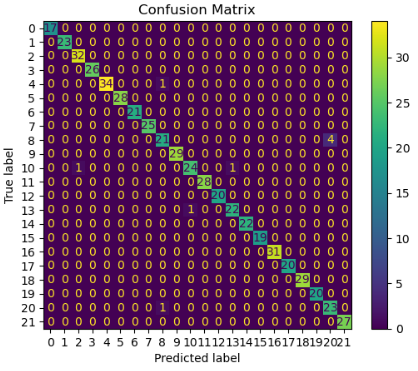
# Predictive Analysis

Decision tree classifier was used for this project

| **Metric** | **Score** |
| --- | --- |
| Accuracy | 98.36% |
| Precision | 98% |
| Recall | 98% |
| F1-Score | 98% |

The classification report displays the evaluation metrics for a multiclass classification model. Here's the interpretation of the metrics:

1. **Precision**: Measures the accuracy of positive predictions for each class. For example, the precision for "apple" is 1.00, which means all predictions for the "apple" class were correct.
2. **Recall**: Measures the model's ability to capture all instances of a particular class. For example, the recall for "jute" is 0.84, indicating the model correctly identified 84% of the "jute" samples.
3. **F1-Score**: The harmonic mean of precision and recall, balancing the trade-off between false positives and false negatives. A high F1-score, such as 1.00 for "apple," indicates excellent performance for that class.
4. **Support**: The number of true instances for each class in the test dataset. For example, there are 25 instances of "jute."
5. The model achieved 98.36% accuracy, indicating that most predictions across all classes were correct.



**Presentation Video Link:**

https://www.loom.com/share/4aaf2585299b4802b7d307e99ddb4b03?sid=ceb3f242-59ee-4df0-be17-3ebda469b73d