CROP RECOMMENDATION SYSTEM

Presented by Abdul Rehman MSDS-2024-S-11

PROBLEM STATEMENT

Developing a machine learning-based crop recommendation system that analyzes key environmental factors to provide accurate crop suggestions, helping farmers optimize their decisions and increase productivity.

OBJECTIVES

1. Analyze the distribution of attributes to identify patterns and ensure data quality.

2. Develop and integrate new features (e.g., Water Availability Index (WAI) and Temperature-Humidity Index (THI)) to enhance the dataset's predictive power.

3. Build machine learning model to predict crop types and evaluate their performance in terms of accuracy and interpretability.

KEY QUESTIONS

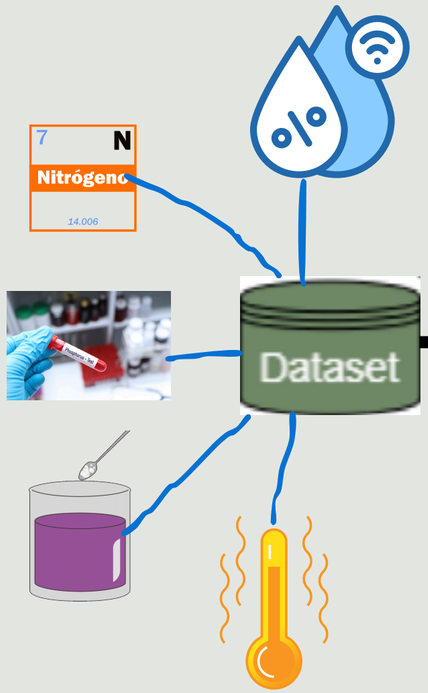
1. How do environmental factors interact and influence the classification of different crops?

2. Can a machine learning model effectively predict crop types based on environmental and soil attributes, and performs best for this dataset?

3. How does the inclusion of engineered features like THI and WAI improve the model's accuracy in predicting crop classifications?

RESEARCH METHODOLOGY

Data Collection	Data Preprocessing	EDA	Model Training	Model Evaluation



Data Preprocesing

EDA

Training data

Testing data

Decision Making

Model Evaluation

Model training

DATA COLLECTION

Data was sourced from a publicly available GitHub repository, including features like temperature, humidity, pH, and rainfall.

DATA WRANGLING

The dataset was cleaned to remove missing values, ensuring quality inputs for model training but there is no missing and dupliacte values in our dataset. During preprocessing, the dataset was normalized using min-max scaling to ensure balanced feature contributions in the model. For feature engineering, key indicators like temperature, humidity, rainfall, and pH were refined and adding new features to enhance model accuracy and performance.

FEATURE ENGINEERING

- **Temperature-Humidity Index (THI)**: A combination of temperature and humidity two critical climatic factors that directly affect crop growth. It helps quantify the level of heat stress on crops, By integrating THI, the model captures complex interactions between temperature and humidity.
- The **Water Availability Index (WAI)** combines rainfall and humidity into a single metric to measure water sufficiency for crops. This helps assess whether environmental conditions provide enough water for optimal crop growth.

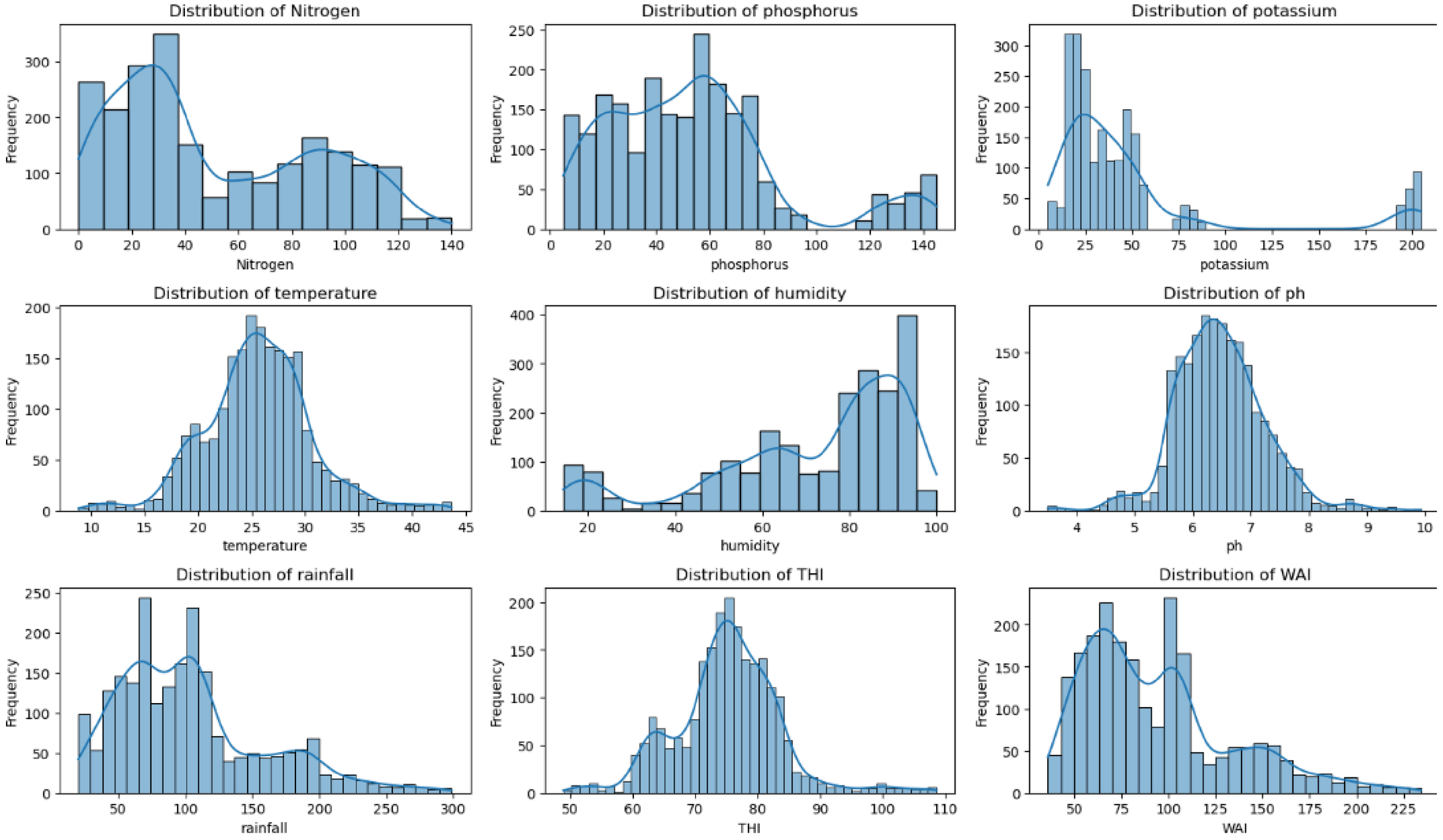
MIN MAX SCALING

MinMax Scaling is a normalization technique that transforms the data to a fixed range, typically between 0 and 1. It helps in scaling all features proportionally so that they contribute equally to the model. This scaling technique was used to ensure that all features have equal influence during model training.

EXPLORATORY DATA ANALYSIS

DESCRIPTIVE ANALYSIS

	Nitrogen	phosphorus	potassium	temperature	humidity	ph	rainfall
count	2200.000000	2200.000000	2200.000000	2200.000000	2200.000000	2200.000000	2200.000000
mean	50.551818	53.362727	48.149091	25.616244	71.481779	6.469480	103.463655
std	36.917334	32.985883	50.647931	5.063749	22.263812	0.773938	54.958389
min	0.000000	5.000000	5.000000	8.825675	14.258040	3.504752	20.211267
25%	21.000000	28.000000	20.000000	22.769375	60.261953	5.971693	64.551686
50%	37.000000	51.000000	32.000000	25.598693	80.473146	6.425045	94.867624
75%	84.250000	68.000000	49.000000	28.561654	89.948771	6.923643	124.267508
max	140.000000	145.000000	205.000000	43.675493	99.981876	9.935091	298.560117



Distribution Plot

Nitrogen : Skewed distribution with most nitrogen values concentrated between 0 and 50.

Phosphorus:Uniform distribution with peaks around 40–60, and fewer data points above 120.

Potassium : Skewed towards lower values, with most data points concentrated below 50.

Temperature : Bell-shaped distribution with most frequent values around 20–25°C, resembling a normal distribution.

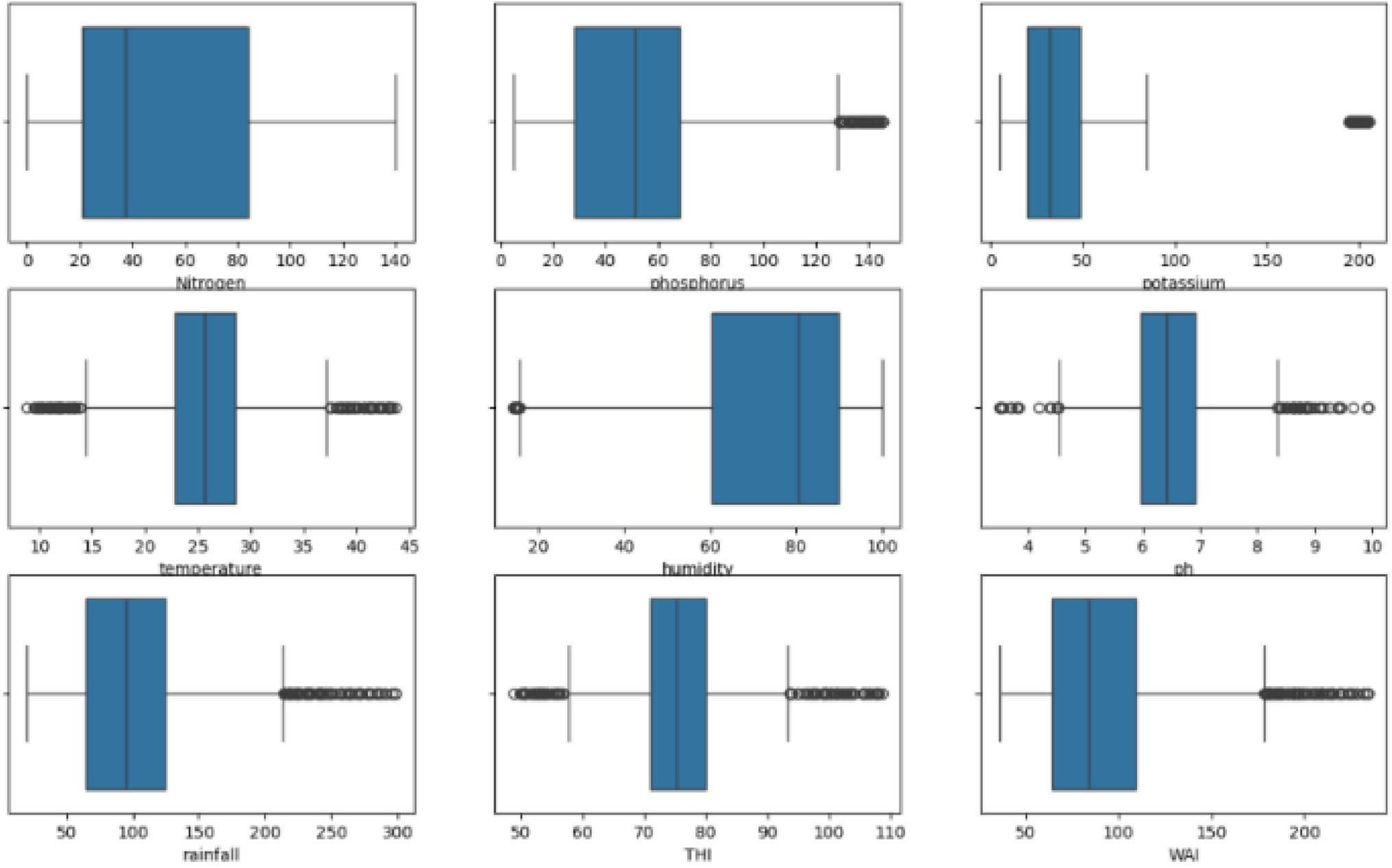
Humidity : Peaks significantly at 80–100, with fewer occurrences at moderate values (40–60).

pH : Bell-shaped distribution centered around 6–7, indicating predominantly near-neutral soil pH.

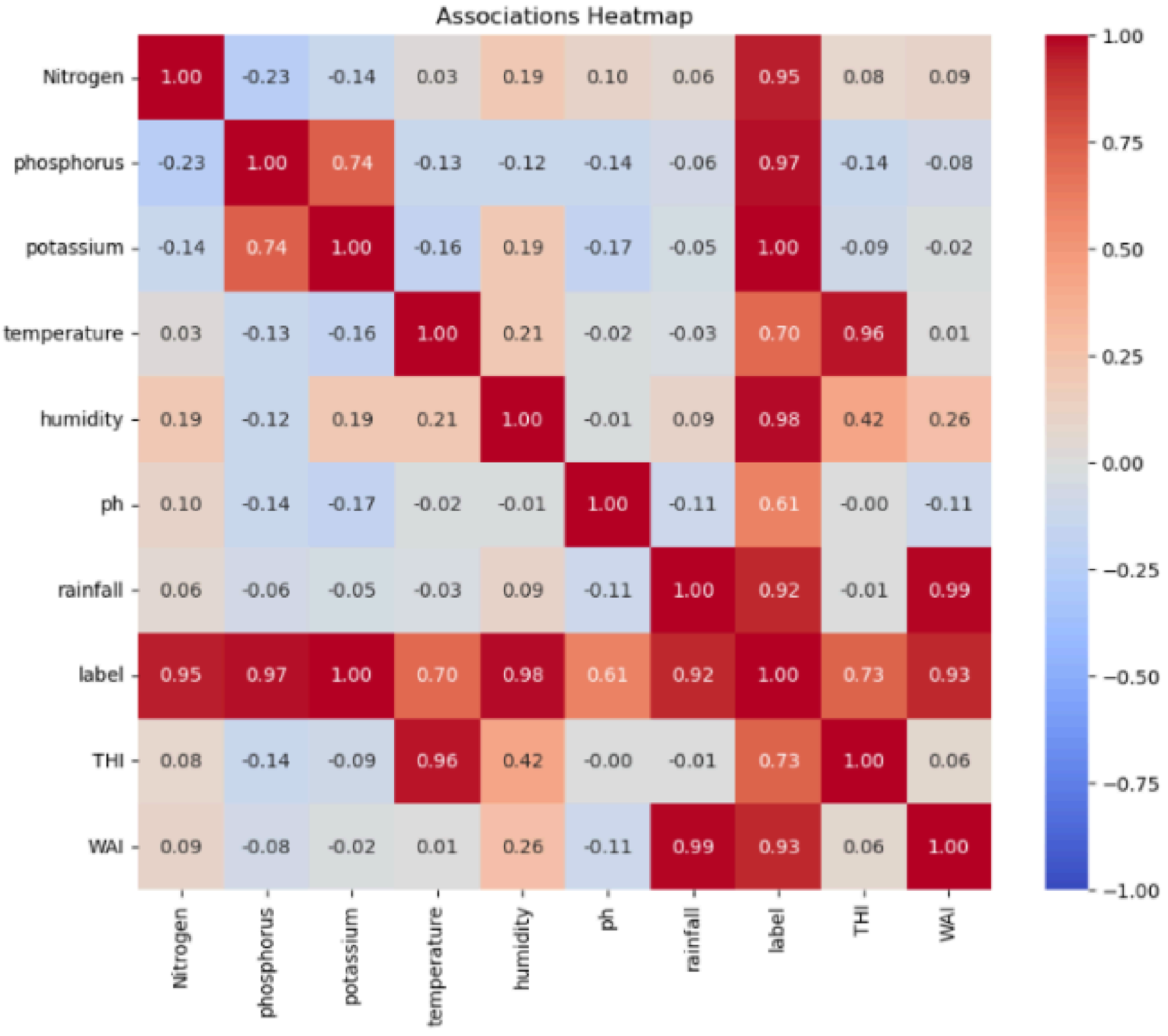
Rainfall :Bi-modal distribution with peaks at 100 and 250, with few data points at very low or high rainfall levels.

Temperature-Humidity Index (THI): Slightly bell-shaped distribution, with most values between 70–90, indicating balanced heat stress conditions.

Water Availability Index (WAI): Peak near 150, with most values between 100–200, indicating moderate water availability



Outlier Detection


Associations Heatmap
Nitrogen 1.00 -0.23 -0.14 0.03 0.19 0.10 0.06 0.95 0.08 0.09
phosphorus -0.23 1.00 0.74 -0.13 -0.12 -0.14 -0.06 0.97 -0.14 -0.08
potassium -0.14 0.74 1.00 -0.16 0.19 -0.17 -0.05 1.00 -0.09 -0.02
temperature 0.03 -0.13 -0.16 1.00 0.21 -0.02 -0.03 0.70 0.96 0.01
humidity 0.19 -0.12 0.19 0.21 1.00 -0.01 0.09 0.98 0.42 0.26
ph 0.10 -0.14 -0.17 -0.02 -0.01 1.00 -0.11 0.61 -0.00 -0.11
rainfall 0.06 -0.06 -0.05 -0.03 0.09 -0.11 1.00 0.92 -0.01 0.99
label 0.95 0.97 1.00 0.70 0.98 0.61 0.92 1.00 0.73 0.93
THI 0.08 -0.14 -0.09 0.96 0.42 -0.00 -0.01 0.73 1.00 0.06
WAI 0.09 -0.08 -0.02 0.01 0.26 -0.11 0.99 0.93 0.06 1.00
Nitrogen
phosphorus
potassium
temperature
humidity
ph
rainfall
label
THI
WAI
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00

PREDICTIVE ANALYSIS

MODEL TRAINING

- The dataset is split into training and testing sets using train_test_split.
- The classifier (DecisionTreeClassifier()) is trained on the training data using model.fit(x_train, y_train).
- Decision Trees in crop recommendation systems offer clear interpretability, handling both linear and non-linear relationships between environmental factors. They also identify key features influencing crop choices, are robust to outliers, and require minimal data preprocessing.

EVALUATION

- Confusion Matrix: It displays how well the model predicted each class
- The Classification Report includes precision (relevance of selected items), recall (ability to identify relevant items), and F1-score (balance between precision and recall). It also shows support, the actual number of occurrences for each class in the dataset.

RESULTS

- The model performs very well, achieving near-perfect accuracy (98%).
- Precision, recall, and F1-scores for each class are also high, indicating good classification for each fruit type (e.g., apple, banana, orange).

Thank You

For your attention