Crop Yield Estimation

Milestone 1

Presented by: Angelica, Arunima, and Ethel

Date: 08/10/2025

Dataset Overview

Dataset:

- Synthetically generated dataset from **Kaggle** by Samuel Ott Attakorah simulating real-world agricultural scenarios. It contains 1 million samples derived from established agricultural factors such as soil, weather and farming practices.

Purpose:

- Created for practicing machine learning models, specifically for specifically for predicting crop yield.

Learning Task:

- Supervised Regression

Features

Categorical Features

- Region (4 distinct regions)
- Crop Type (6 crops)
- Soil Type (6 major types)
- Weather Condition (3 conditions)
- Fertilizer Used (True/False)
- Irrigation Used (True/False)

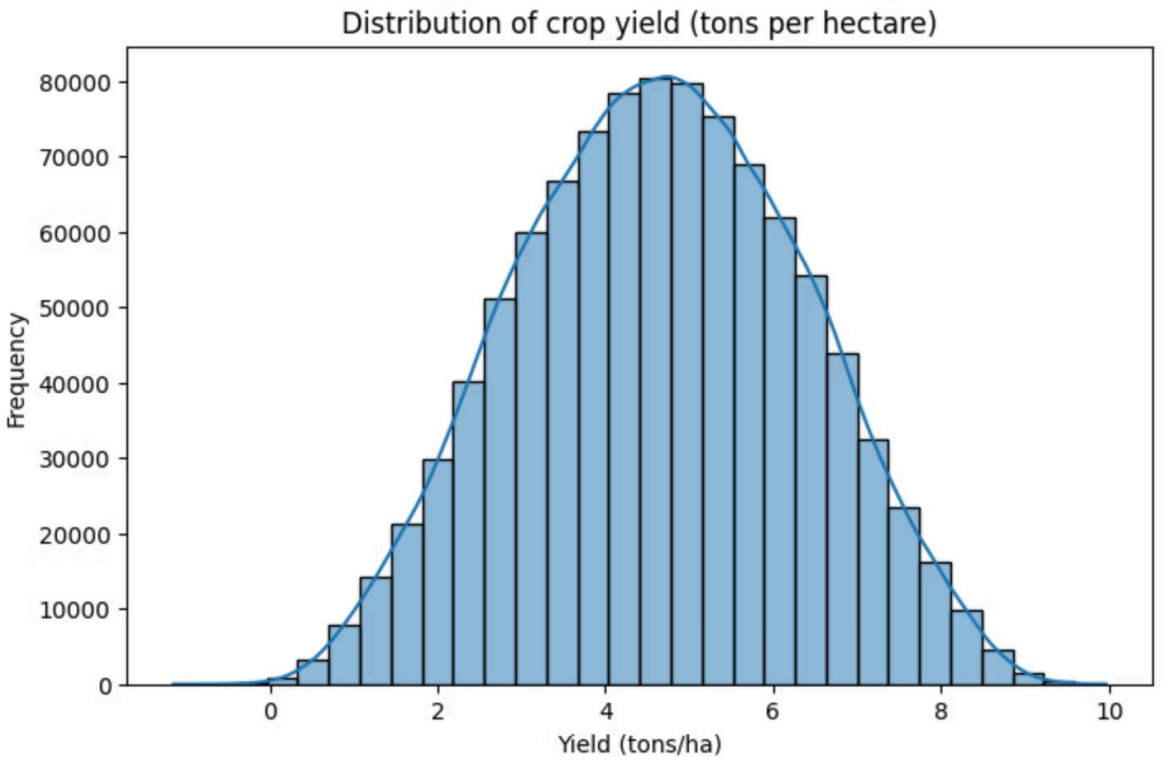
Numerical Features

- Rainfall (mm)
- Temperature (°C)
- Days to Harvest (days) - number of days taken to harvest crop after planting

Output Variable:

Yield (tons/hectare)- A continuous numerical variable of the total crop yield produced

EDA – Crop Yield Distribution

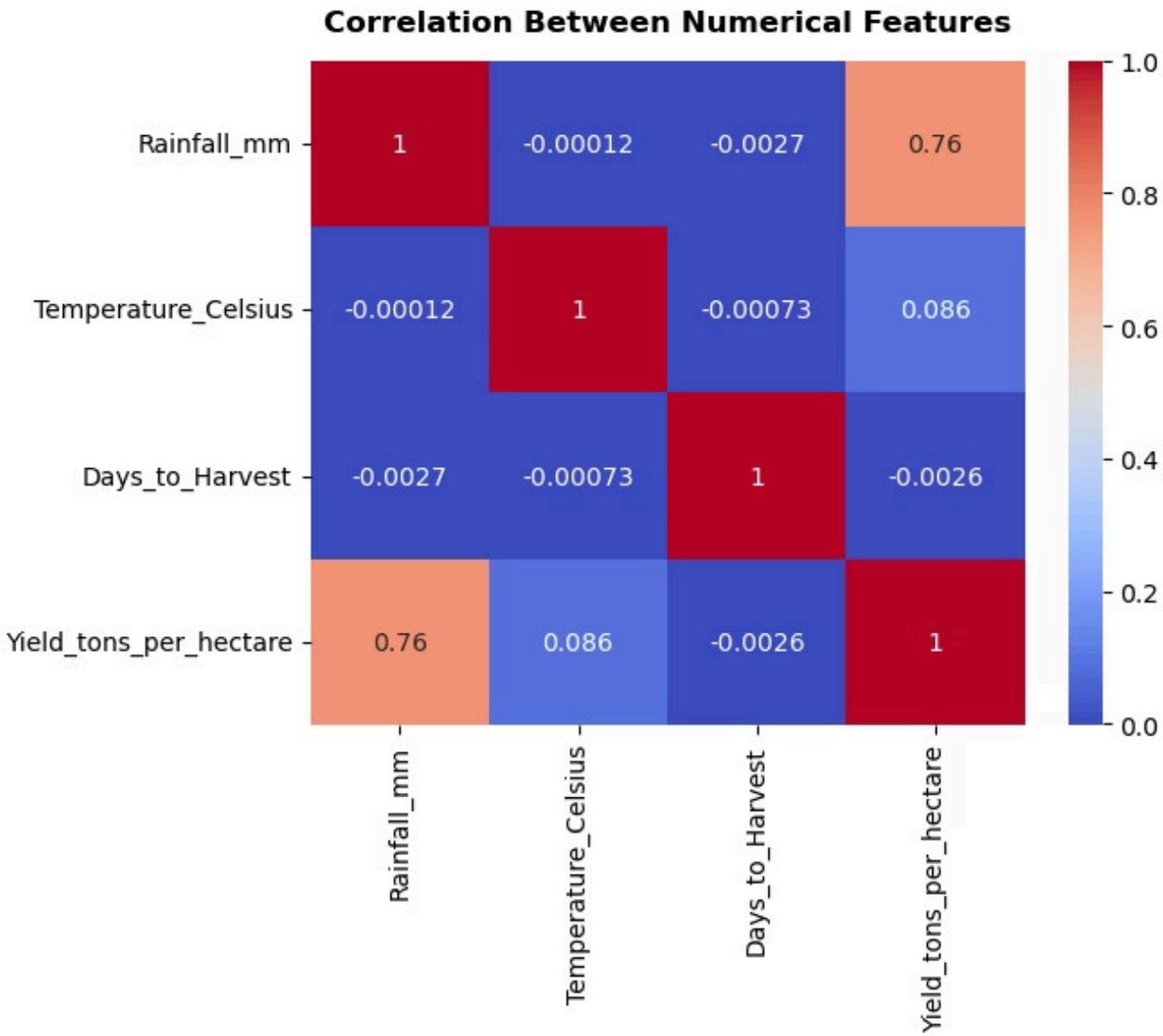


- Approximately a normal Distribution.
- The minimum yield value is **negative** which shows presence of outliers.

	Rainfall_mm	Temperature_Celsius	Days_to_Harvest	Yield_tons_per_hectare
count	1000000.000000	1000000.000000	1000000.000000	1000000.000000
mean	549.981901	27.504965	104.495025	4.649472
std	259.851320	7.220608	25.953412	1.696572
min	100.000896	15.000034	60.000000	-1.147613
25%	324.891090	21.254502	82.000000	3.417637
50%	550.124061	27.507365	104.000000	4.651808
75%	774.738520	33.753267	127.000000	5.879200
max	999.998098	39.999997	149.000000	9.963372

Table 1: descriptive statistics (numerical features)

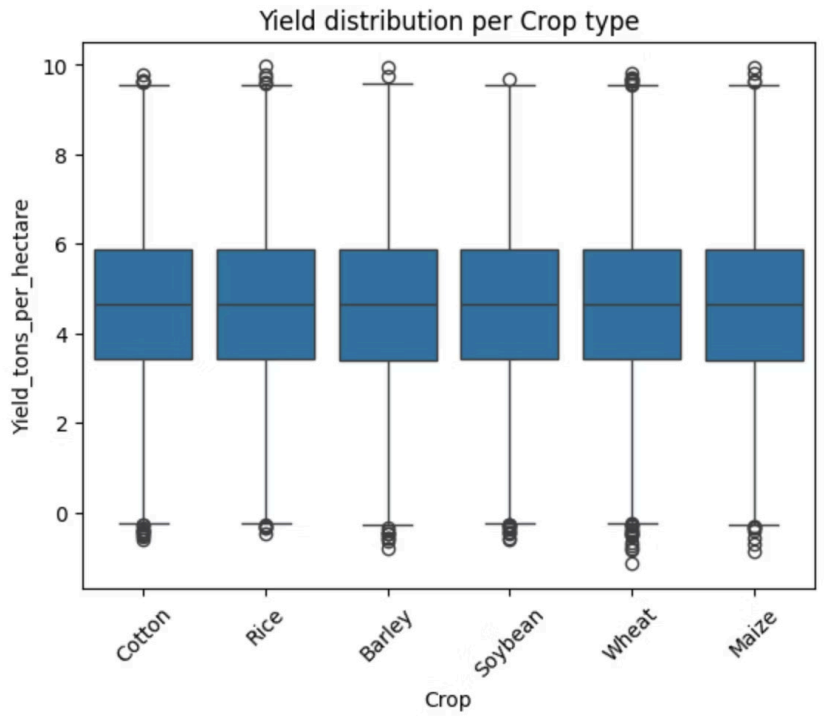
EDA – Numerical Feature Correlation

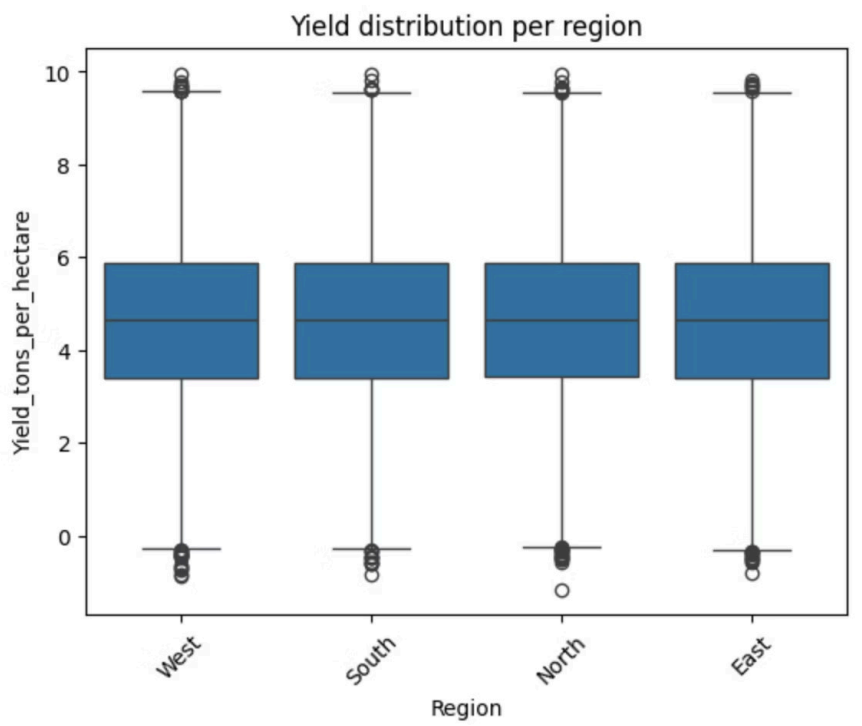


Rainfall has strong positive correlation of 0.76. This suggests that higher rainfall tends to be associated with a higher yield.

Temperature and days to harvest have a weak correlation to crop yield.

EDA – Categorical Feature Impact





Preprocessing

1. **Data Cleaning:**
   - No missing values.
   - Identified and removed 231 observations with negative yield values.
2. **Feature Scaling:**
   - All numerical features normalized because they have varying scales and units.
3. **Encoding categorical features:**
   - Nominal features, i.e. crop type or soil type , encoded using OneHotEncoder.
   - Binary features, i.e. fertilizer used, encoded manually as 0 or 1.

Evaluation Protocol

1

Initial data Split

Train/Test Split: 80%/20%

Why: 1 million data points hence 200k test set is enough for model to generalize fairly

2

Cross Validation

Nested GroupKFold Cross Validation

Why: to separate hyperparameter tuning (Inner Loop) from final model performance estimation (Outer Loop). Both loops will group by Region since crop yields in the same region may share similarities preventing data leakage.

3

Final model Training

Final check on 20% Hold-Out Test Set

Why: verify the model's generalization performance on unseen data.

Metrics

In this study, we will evaluate the performance of our crop yield estimation model using two commonly used regression metrics:

- Root Mean Squared Error (**RMSE**) : provides an indication of the model's prediction error in the same units as the target variable and is particularly sensitive to large errors.
- Mean Absolute Error (**MAE**) : measures the average magnitude of the errors regardless of their direction and is less influenced by outliers.