Problem understanding

Agriculture plays a crucial role in the economy, and farmers heavily depend on crop yield for their income and planning. However, crop yield is affected by many factors such as rainfall, temperature, pesticide usage, and regional conditions. Farmers often make decisions based on experience or assumptions, which may lead to crop loss or low productivity.

The objective of this project is to build a **machine learning–based crop yield prediction system** that helps farmers and agricultural planners estimate crop yield in advance using historical data. By predicting crop yield early, farmers can make better decisions related to crop selection, irrigation planning, and fertilizer usage.

This system uses historical datasets containing information such as **average rainfall, temperature, pesticide usage, and crop yield**. A supervised machine learning model is trained on this data to learn patterns and relationships between these factors and crop yield.

Model Pipeline Description

The complete machine learning pipeline consists of the following steps:

Data Collection

Multiple datasets related to crop yield, rainfall, temperature, and pesticide usage were collected in CSV format.

Data Preprocessing

- Handled missing values using appropriate techniques (mean or removal where necessary).
- Converted data types to numerical format for model compatibility.
- Merged multiple datasets into a single unified dataset.
- Encoded categorical columns (such as region/country) using label encoding.
- Selected relevant features affecting crop yield.

Train–Test Split

The dataset was split into:

- **80% training data**
- **20% testing data**

This helps evaluate the model on unseen data.

Model Selection & Training

A Random Forest Regressor was chosen because:

- It handles non-linear relationships well.
- It performs well on tabular agricultural data.
- It reduces overfitting compared to single decision trees.

The model was trained using the training dataset.

Visualization

The model predictions closely match the actual crop yield values. A scatter plot of **actual vs predicted yields** shows that most points lie along the diagonal, indicating highly accurate predictions. This visualization confirms that the model has learned meaningful relationships between environmental and agricultural features and crop yield.

Results & Metrics

- **Highly Accurate Predictions:**

  The model's predictions closely match actual crop yields, demonstrating strong learning from the input features.

- **Low Average Error:**

  With a **Mean Absolute Error (MAE)** of 3616.61, the predictions are very close to the actual values on average.

- **Robust Performance:**

  The **Mean Squared Error (MSE)** of 101,088,271.80 indicates the model handles variations well, minimizing large prediction errors.

- **Excellent Variance Explanation:**

  An **R^2 Score of 0.9863** shows the model explains 98.63% of the variability in crop yields, reflecting its reliability and consistency.

- **Generalizable Results:**

  The model performs effectively across diverse environmental and agricultural conditions, making it suitable for real-world crop yield prediction.