Crop Yield Prediction Using Machine Learning

AI / ML Engineer – Problem Statement

Build a machine learning model that predicts an outcome using a real or public dataset.

Scenario:
Given a dataset related to crop yield, sales, energy usage, or health metrics, build a model to predict a future value and explain the results.

Google Colab Link:
https://colab.research.google.com/drive/1B-bNgzs98LgPA07joGYF5gxH-DSAYyoM?usp=sharing

GitHub Repository:
https://github.com/keerthana123987/crop_yield_prediction

1. Problem Understanding

Agriculture plays a vital role in ensuring food security, supporting rural livelihoods, and contributing significantly to economic growth. In many countries, especially developing nations, a large portion of the population depends on agriculture as their primary source of income.

Crop yield is influenced by various environmental and management factors such as rainfall, temperature, soil quality, humidity, fertilizer usage, pesticide application, and land area.

This project aims to build a machine learning–based predictive model that estimates crop yield using historical agricultural and climatic data.

Accurate crop yield prediction helps farmers and policymakers in planning crop selection, irrigation, storage, and distribution. The adoption of data-driven approaches supports smart and sustainable agricultural practices.

2. Dataset Description

The dataset contains 50,000 records with 10 attributes.

Features:
- Region
- Crop Type
- Rainfall
- Temperature
- Humidity
- Soil pH
- Fertilizer Usage
- Pesticide Usage
- Area

Target:

- Crop Yield (tons per hectare)

3. Data Preprocessing

- Checked missing values
- Encoded categorical variables
- Normalized numerical features
- Split data into training and testing sets (80:20)
- Applied StandardScaler for feature scaling

4. Model Pipeline

Two models were implemented:
- Linear Regression
- Random Forest Regressor

Pipeline:
Data → Preprocessing → Training → Evaluation → Prediction

Random Forest was selected as the final model due to better performance.

5. Results and Evaluation

Evaluation Metrics:
- RMSE (Root Mean Square Error)
- MAE (Mean Absolute Error)
- R^2 Score

Random Forest achieved better accuracy and lower error compared to Linear Regression.
Actual vs predicted graphs were used for visualization.

6. Inference and Explanation

Feature importance analysis showed that rainfall, temperature, and fertilizer usage are major factors affecting crop yield.

Higher rainfall and proper fertilizer usage improved productivity, while extreme soil pH reduced yield.

The model learns nonlinear relationships between environmental factors and yield.

7. Conclusion

This project successfully built a predictive model for crop yield estimation using machine learning techniques.

Random Forest provided reliable predictions.

Future improvements include integrating real-time weather data and deep learning models.