CAPSTONE PROJECT REPORT

SUSTAINABLE CROP YIELD PREDICTION USING MACHINE LEARNING

PROJECT MEMBERS:

- Sakshi Morey
- Ankita Godghate
- Kashish Walke
- Rutuja Saharkar

COMPANY NAME: Edunet Foundation

1. INTRODUCTION

Agriculture plays a vital role in the global economy and food security, yet it faces numerous challenges due to climate change, soil degradation, and unpredictable weather patterns. Farmers often struggle with yield estimation, leading to inefficient resource allocation and potential financial losses. **Sustainable Crop Yield Prediction using Machine Learning** aims to address these challenges by leveraging **data-driven insights** for optimized agricultural productivity.

Our project aligns with the **United Nations' 17 Sustainable Development Goals (SDGs)**, specifically:

(1) Zero Hunger (SDG 2) – By improving crop yield predictions, farmers can enhance food security and reduce the risk of shortages.
(2) Climate Action (SDG 13) – Helps in sustainable farming practices by considering weather patterns and reducing environmental impact.
(3) Responsible Consumption & Production (SDG 12) – Optimized yield predictions reduce wastage of agricultural resources.
(4) Industry, Innovation & Infrastructure (SDG 9) – Uses Machine Learning & AI to bring technological advancement in agriculture.

By utilizing Machine Learning algorithms, specifically Linear Regression, we aim to predict crop yield efficiently based on soil quality, weather conditions, and past agricultural data. Our approach ensures sustainable farming, reduces environmental impact, and enhances decision-making for farmers.

2. PROBLEM STATEMENT

To build a **predictive model** for crop yield estimation using **Soil Quality (pH), Weather Conditions, and Past Agricultural Data (past_crop_yield)** to help in agricultural decision-making.

3. DATASET DESCRIPTION

The dataset used in this project consists of agricultural records with the following attributes:

Feature Name	Description
Soil_Quality	A scale (1-10) indicating soil fertility
Temperature (°C)	Average temperature of the region
Rainfall (mm)	Rainfall received in the area
Past_Yield (kg/hectare)	Historical crop yield data
Crop_Yield (kg/hectare)	Target variable (Actual crop yield)

4. METHODOLOGY

4.1 Dataset Description

The dataset used in this project consists of multiple features related to soil conditions, weather data, and past yield records. The dataset underwent preprocessing to remove missing values and ensure proper data formatting.

4.2 Data Preprocessing

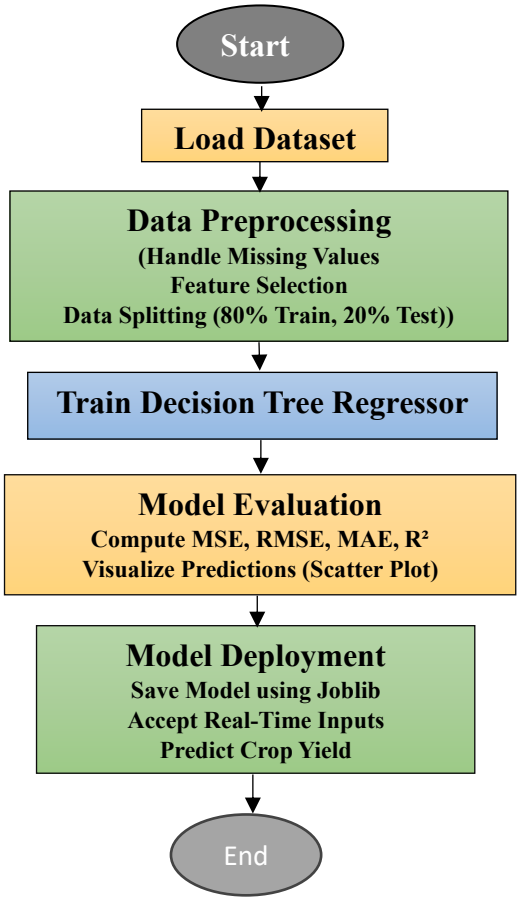
- Handling Missing Values: Checked for null values and removed them.
- Feature Selection: Used relevant attributes for training the model.
- Data Splitting: The dataset was split into 80% training and 20% testing.

4.3 Machine Learning Models Used & Flow Chart

To ensure the best possible accuracy, multiple machine learning models were trained and compared:

1. Decision Tree Regressor
2. Random Forest Regressor
3. XGBoost Regressor
4. Linear Regression
5. Gradient Boosting Regressor
6. Support Vector Regressor (SVR) (without tuning)

Flowchart:



4.4 Model Training & Evaluation Metrics

Evaluation Metrics Used:

1. Mean Squared Error (MSE)
2. Root Mean Squared Error (RMSE)
3. Mean Absolute Error (MAE)
4. R-squared (R^2 Score)

Model Comparison Table

Model	R^2 Score (Accuracy)	RMSE	MAE
Linear Regression	82.44%	92.88	74.87
Random Forest	99.30%	18.54	8.05
Decision Tree	**99.62%**	**13.65**	**1.15**
Gradient Boosting	87.32%	78.92	64.69
XGBoost	99.33%	18.10	10.91
SVR (without tuning)	82.31%	93.21	74.50

After training the Decision Tree Regression model, we evaluated its performance using the following metrics: **Higher R-squared value (99.62%)** indicates that the model explains most of the variance in crop yield. **Lower RMSE (13.65) and MAE (1.15)** show that the model's predictions are close to actual values.

4.5 Best Suited Model: Decision Tree Regressor : Among all tested models, **Decision Tree Regressor** performed the best with an **R^2 Score of 99.62%**. This model is best suited for the crop yield prediction task due to:

- **High Accuracy:** The Decision Tree model captures non-linear relationships between features, making it well-suited for agricultural datasets.
- **Interpretability:** It provides clear decision-making steps, making it easier to understand the impact of each feature on the yield.
- **Efficiency:** The model runs quickly and effectively with structured numerical data.
- **Low Bias and High Variance:** Since Decision Trees can overfit, careful parameter tuning can further enhance performance while maintaining accuracy.

4.6 Visualization: Actual vs Predicted Crop Yield : A scatter plot was generated to compare actual vs predicted crop yield values. The red dashed line represents the ideal prediction line, indicating where the predicted values should align perfectly with the actual values. The blue points in the scatter plot represent the actual vs predicted values for crop yield. A strong alignment of blue points along the red dashed line signifies that the model is making highly accurate predictions. Minimal deviation from the line indicates that the model generalizes well to unseen data.

4.7 Model Deployment: To deploy the trained model, it was saved using the Joblib library, allowing real-time predictions. Model (crop_yield_data.pkl). A user can input real-time data, and the model will predict the crop yield.

```
[699]: print("\nEnter the following values to predict Crop Yield ")
soil_quality = float(input("Enter Soil Quality (1-10 scale): "))
temperature = float(input("Enter Temperature (°C): "))
rainfall = float(input("Enter Rainfall (mm): "))
past_yield = float(input("Enter Past Crop Yield (kilograms per hectare): "))
# Convert input into a DataFrame
input_data = pd.DataFrame([[soil_quality, temperature, rainfall, past_yield]],
                          columns=['Soil_Quality', 'Temperature', 'Rainfall', 'Past_Yield'])
# Predict crop yield
predicted_yield = model.predict(input_data)[0]
print(f"\n * Predicted Crop Yield: {predicted_yield:.2f} kilograms per hectare")
```

5. RESULT

Among all tested models, the **Decision Tree Regressor** performed the best with an **R^2 Score of 99.62%**, making it the optimal choice for crop yield prediction.

```
Enter the following values to predict Crop Yield
Enter Soil Quality (1-10 scale):  6
Enter Temperature (°C):  23
Enter Rainfall (mm):  400
Enter Past Crop Yield (kilograms per hectare):  500

 * Predicted Crop Yield: 2072.35 kilograms per hectare
```

6. CONCLUSION & FUTURE SCOPE

This project successfully developed a Machine Learning-based Crop Yield Prediction System that:

- Provides **highly accurate** predictions for crop yield.
- Helps **farmers optimize agricultural productivity** and resource management.
- Supports **sustainable agriculture and food security initiatives**.

Future Scope:

- **Enhancing Features:** Adding soil nutrient levels, fertilizer usage, and humidity.
- **Testing Advanced Models:** Exploring Deep Learning & Hyperparameter Tuning for better accuracy.
- **Deploying as a Web-Based Tool:** Providing real-time predictions through a web or mobile application for farmers and agricultural researchers.