Project Report

1️⃣ Problem Understanding

Phase 1 – Exploratory Data Analysis (EDA):

A fertilizer recommendation dataset was analyzed to understand feature distribution, class balance, and relationships between environmental conditions and fertilizer usage. The dataset was found to be clean, balanced, and free from significant outliers. Correlation and distribution analysis confirmed suitability for modeling.

Phase 2

Part A – Crop Recommendation (Classification)

The goal was to build a classification model capable of recommending the most suitable crop based on:

- Soil nutrients (N, P, K)
- Environmental conditions (temperature, humidity, rainfall, pH)

This is a multiclass classification problem where the target variable is crop type.

Part B – Yield Prediction (Regression)

The goal was to build a regression model to predict crop yield (tons per hectare) using:

- Rainfall
- Temperature
- Soil type
- Crop type
- Fertilizer usage
- Irrigation usage
- Weather conditions
- Days to harvest

This is a supervised regression problem where the target variable is continuous.

2️⃣ Data Preprocessing Steps

The following preprocessing steps were performed:

✔ Data Cleaning

- Verified dataset structure using `.info()`
- Confirmed no missing values
- Checked for duplicate entries

- Standardized column names

✔ Outlier Detection

- Used IQR method and visual boxplots
- No significant outliers detected

✔ Encoding Categorical Variables

- Label Encoding was applied to categorical columns such as:
  - Crop
  - Soil_Type
  - Region
  - Weather_Condition
- Boolean variables (Fertilizer_Used, Irrigation_Used) were converted to integer

✔ Feature Scaling (Classification Only)

- StandardScaler was applied for Logistic Regression
- Scaling was performed after train-test split to prevent data leakage

✔ Train-Test Split

- Dataset was split using 80% training and 20% testing data
- Random state fixed for reproducibility

3️⃣ Model Comparison Table

📌 Crop Recommendation (Classification)

Model	Test Accuracy	Precision	Recall	F1 Score
Logistic Regression	96.36%	0.964	0.964	0.964
Random Forest	99.31%	0.994	0.993	0.993
XGBoost	98.41%	0.985	0.984	0.984

📌 Yield Prediction (Regression)

Model	RMSE	MAE	R^2
Linear Regression	0.501	0.399	0.9138
Random Forest	0.522	0.416	0.9064
XGBoost	0.509	0.406	0.9109

4️⃣ Final Model Selection Reasoning

🌾 Crop Recommendation

Random Forest achieved the highest accuracy (99.31%) and F1-score (0.993). Although it achieved perfect training accuracy, the small difference between training and testing performance indicated acceptable generalization.

Random Forest was selected because:

- It captures nonlinear relationships effectively
- It handles multiclass classification well
- It provided the highest predictive performance

🌾 Yield Prediction

Linear Regression achieved:

- Highest R^2 score (0.9138)
- Lowest RMSE (0.501)
- Lowest MAE (0.399)
- Almost identical train and test R^2 (no overfitting)

The dataset exhibited strong linear relationships (Rainfall correlation = 0.76), making Linear Regression sufficient and more stable than complex tree models.

Therefore, Linear Regression was selected as the final regression model.

5️⃣ Challenges Faced

1. Handling version differences in scikit-learn (RMSE calculation issue).
2. Interpreting correlation results for encoded categorical variables.
3. Preventing data leakage during scaling.
4. Managing computational cost due to large dataset size
5. Identifying overfitting in tree-based models.

6️⃣ Suggestions for Real-World Deployment

To deploy this system in real-world agricultural settings:

- Integrate with real-time weather APIs.
- Connect with IoT soil sensors for live NPK and moisture data.
- Build a web or mobile interface for farmers.
- Implement periodic retraining using new seasonal data.
- Validate model performance using real farm datasets.

Additionally, a cloud-based deployment using Flask/FastAPI with a REST API can allow scalable real-time recommendations.

🎯 Final Conclusion

This project demonstrates that model performance depends heavily on data characteristics. While complex ensemble methods performed best for classification, simpler linear models proved superior for regression due to strong linear relationships in the dataset.

Careful evaluation using appropriate metrics and overfitting analysis led to informed and justified model selection.