Project Proposal

Course Name : Introduction to Data Science Lab
Course Code : CSE 0688 2292
Semester : 2nd Year, 2nd Semester, 2022-23
Department : Computer Science and Engineering
Institution : Shahjalal University of Science and Technology
Submission Date : 19/09/2025

Group number : 20
Registration number : 2022331033
2022331099

Project Title
Predicting Crop Yield for Agricultural Planning in Bangladesh

---

Project Overview:

Agriculture is the backbone of Bangladesh's economy, employing nearly half of the workforce and ensuring national food security. However, crop production is highly vulnerable to climatic variations, regional disparities, and seasonal changes. Predicting crop yield accurately is crucial for policymakers, agricultural agencies, and farmers to make informed decisions about resource allocation, market pricing, and food supply chain management, thereby preventing shortages and ensuring economic stability.

This project is aimed to predict the production yield of various crops across different districts and seasons in Bangladesh using ***a powerful Gradient Boosting (XGBoost) regressor***. By analyzing the comprehensive *SPAS (Seasonal Production and Area Statistics) dataset*, we need to identify key agronomic and climatic indicators to develop a machine learning model that can assist in *strategic agricultural planning* and *risk mitigation.*

Tools and Technologies:

- Jupyter Notebook
- Python
- Pandas
- Seaborn
- Scikit-learn
- Matplotlib
- XGBoost

Prerequisites:

- Data manipulation and cleaning using Pandas
- Advanced data visualization with Matplotlib and Seaborn
- Feature engineering, encoding, and scaling techniques
- Implementing machine learning regression workflows with scikit-learn

- Understanding and interpreting regression metrics (MAE, RMSE, R^2)
- Hyperparameter tuning with GridSearchCV/RandomizedSearchCV

Objectives:

- To perform exploratory data analysis (EDA) to understand the distribution of features and their relationship with crop production.
- To preprocess and clean the dataset by handling invalid entries (#DIV/0! errors) and missing values.
- To engineer meaningful features and encode categorical variables like District, Crop Name, and Season.
- To build, optimize, and evaluate an XGBoost regression model using hyperparameter tuning.
- To identify the key factors (e.g., area, district, season, average temperature) that most significantly impact crop yield.
- To translate model insights into actionable recommendations for agricultural stakeholders.

Methodology:

- Data Preprocessing and Cleaning
- Exploratory Data Analysis (EDA)
- Feature Engineering and Selection
- Model Building (Linear Regression, Decision Tree, Random Forest, XGBoost)
- Hyperparameter Optimization
- Model Evaluation and Interpretation

Expected Outcomes:

- Implementing and optimizing an advanced XGBoost regressor for a real-world economic problem.
- Evaluating model performance using regression-specific metrics in an agricultural context.
- Analyzing feature importance to understand the drivers of crop production.
- Translating machine learning results into actionable insights for agricultural planning and policy.

References:

- Scikit-learn: https://scikit-learn.org/stable/
- XGBoost Documentation: https://xgboost.readthedocs.io/
- Dataset : https://data.mendeley.com/datasets/cphdw4z5kw/2
- Any additional tools or resources during the project, will be added as required.

Team Members:

Name	Registration Number	Contributions
Shyamali Das	2022331033	Data Preprocessing, EDA, Model Building & Tuning
Afia Farzana	2022331099	Feature Engineering, Hyperparameter Optimization, Documentation