Ahmedabad University

CSE 623 Machine Learning

Weekly Report - 1

Group- Vector Minds

Team

Member	**Enrollment No.**
Adyan Shamim	AU2120040
R. Priscilla	AU2340001
Shreya Dhumal	AU2340124
Vishwa Bharoliya	AU2340002

Project Title: Crop Yield Prediction Using Classical Machine Learning Models and Climatic Factors

Project Definition: Agriculture plays an important role in our lives, especially in a well-known agricultural country like India. It provides food security to a larger population; however, its productivity depends heavily on factors such as rainfall, temperature, and seasonal weather conditions. Due to increasing climate variability, predicting crop yields often becomes difficult. The aim of this project is to develop a predictive system that estimates crop yield based on provided agricultural and historical data. The project applies a classical machine learning regression model to analyze the effect of environmental factors on agricultural productivity.

The project treats crop yield prediction as a supervised regression problem.

Where:

- Input variables (Features): Rainfall, temperature, state, crop type, season, fertilizer usage, area of cultivation, etc.
- Output variable (Target): Crop yield which is production per unit area.

Objectives for Week 1

1. Identification of crop yield prediction as a supervised regression model and determining input features (climatic and agricultural factors) and the target variable (crop yield).
2. Review of research papers, textbooks and Scikit-learn documentation to understand regression models. Importance of evaluation metrics like MAE, RMSE and R^2 in a regression model. Understanding of preprocessing for model data.
3. Exploration of the provided dataset and initialising the workflow for projects such as data cleaning and selecting appropriate regression models for further project work.

Work Completed:

During the first week, the initial groundwork for the crop yield prediction project was completed. The main focus was on understanding the problem statement, collecting datasets, and beginning the preprocessing stage and understanding the project pipeline.

The project aims to develop a predictive system that estimates crop yield based on climatic and agricultural factors. The input consists of climatic and agricultural features and the output is the crop yield. The project emphasizes model comparison, performance evaluation, feature importance analysis, and interpretability of results. Understanding the agricultural significance of yield prediction highlighted its role in agricultural planning, food security, resource optimization, and policy formulation.

A literature review was also done which covered the importance of crop yield prediction in agriculture, the influence of factors and the differences between traditional statistical approaches and machine learning methods. Fundamental concepts of regression models and commonly used evaluation metrics such as Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R^2 Score were reviewed.

Key references included *Data Mining: Concepts and Techniques* by Han et al., Scikit-learn documentation, and Kaggle notebooks related to crop yield prediction. This review reinforced the importance of proper preprocessing, feature scaling, and encoding in regression-based machine learning models.

Dataset Collection:

We explored two Kaggle datasets : the Crop Yield in Indian States Dataset and the Agriculture Crop Yield Dataset. Important features identified in these datasets include state, crop, season, rainfall, temperature, area, production, and yield. Initial data exploration was performed where the datasets were loaded and examined by checking shape, column names, and data types.

Dataset: https://www.kaggle.com/datasets/akshatgupta7/crop-yield-in-indian-states-dataset

Preprocessing Setup:

Preprocessing setup was initiated during this week. Preliminary steps included handling null or inconsistent entries and removing duplicate records. Plans were made to apply encoding for categorical features such as crop name and season. The need for scaling or normalization of numerical features was also identified to ensure better performance of regression models.

The project environment was finalized using Python along with NumPy, Pandas, Scikit-learn, Matplotlib, and Jupyter Notebook.

Tasks Planned for the Coming Week:

For the coming week, we aim to start data preprocessing and implementing the first set of regression models. Planned tasks include completing missing value treatment, performing outlier detection, applying encoding techniques for categorical variables, and scaling numerical features where required and producing a cleaned dataset ready for model training and testing.

Linear Regression, Ridge Regression, and Lasso Regression models will be implemented using a proper train-test split strategy. Evaluation metrics such as MAE, RMSE, and R^2 Score will be used for performance comparison. Additionally, correlation heatmaps and feature distribution plots will be generated to support analysis, and preliminary feature importance analysis will begin.

By the end of next week, the expected deliverable is a fully cleaned and preprocessed dataset ready for model training and testing, along with baseline regression models implemented and evaluated.