Ahmedabad University

CSE 623 Machine Learning

Weekly Report - 2

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 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): State, Crop type, Season, Crop Year, Area of cultivation, Annual Rainfall, Fertilizer usage, and Pesticide usage

- Output variable (Target): Crop yield, which is production per unit area.

Objectives for Week 3

1. Complete data preprocessing by performing three tasks: treating missing values, encoding categorical variables, and scaling numerical features.
2. Conduct Exploratory Data Analysis (EDA) to examine how features in the dataset distribute values and relate to crop yield.
3. Build three baseline regression models using Linear Regression, Ridge Regression, and Lasso Regression.
4. Measure and analyze model performance with three metrics: MAE, RMSE, and R^2 Score.
5. Conduct an initial evaluation of feature importance and explain the findings.

Work Completed

Work Completed During this phase, the project progressed from theory to practice. We cleaned and prepared the dataset for model training.

Data Preprocessing

1. Removed duplicate records and addressed missing values to maintain data quality.
2. Encoded categorical variables, including State, Crop, and Season, using suitable encoding techniques.
3. Standardized numerical features such as Crop_Year, Area, Annual_Rainfall, Fertilizer, and Pesticide to improve regression performance.
4. Outlier detection used boxplots, which identified extreme values in Area, Fertilizer, and Pesticide usage, and later applied regularization methods to lessen their impact.
5. Next, we divided the dataset into training and testing sets using an appropriate train-test split strategy.

Model Implementation

We implemented three classical regression models using Scikit-learn:

1. Linear Regression. This served as a baseline model to understand linear relationships between features and crop yield.
2. Ridge Regression. This model addressed multicollinearity and reduced overfitting using L2 regularization.
3. Lasso Regression. This model focused on feature selection and regularization using L1

Exploratory Data Analysis (EDA)

- Exploratory Data Analysis was performed after preprocessing to know about data and explore the relationships between variables and crop yield.
- Distribution plots were created for various numerical variables, including Annual_Rainfall, Area, Fertilizer, Pesticide, Crop_Year, and Yield. Then the plots actually provide the spread and skewness for the data. Some variables mainly, fertilizer and pesticide use were found to have right-skewed distributions, which reinforced the need for feature scaling.
- Box plots were included to identify the outliers. Some outliers were found in variables Area, fertilizer and pesticide have reinforced the regularization techniques such as Ridge and Lasso regressions and have accounted for their effect.
- A correlation heatmap explores relationships between variables. Area and fertilizer were found to be relatively stronger positive correlation with yield, while Annual_Rainfall had a moderate effect. Some correlations were found to be independent variables, which suggests the presence of multicollinearity.
- Scatter plots further revealed positive correlations between Area, Fertilizer use, and Yield. Categorical variables such as State, Crop, and Season were also explored to reveal differences in productivity levels across states and seasons.
- In summary, EDA was employed to determine the factors that influence crop yield. Each model was trained using the training dataset and evaluated on the test dataset.

Model Evaluation

The models were evaluated using the following metrics given :

- **Mean Absolute Error (MAE)** – Measures average prediction error

- **Root Mean Squared Error (RMSE)** – Penalizes larger errors more better

- **R² Score** – Indicates the ratio of variance that is explained by the model

By comparing the performance, differences in generalization capability between models was observed, both ridge and lasso regression showed improved stability compared to simple inlear regression in handling correlated features.

Initial Findings

- Rainfall and area of cultivation showed significant correlation with crop yield.

- Regularization techniques helped decrease the amount of model overfitting

- Feature scaling had a positive effect on both model convergence and performance.

The results we got show the importance of proper preprocessing and model selection in regression-based agricultural prediction systems.

Tasks Planned for Week 3

For the upcoming week, we plan to do the following tasks for optimization:

- Hyperparameter tuning of Ridge and Lasso models using GridSearchCV

- Implementation of more regression models such as Decision Tree Regressor and Random Forest Regressor for comparison in the degree of performance

- Cross-validation to ensure model robustness

- Detailed feature importance analysis and visualization

- Preparation of performance comparison tables and result interpretation

By the end of Week 3, the expected deliverable is an optimized regression model with improved prediction accuracy, along with comparative analysis across multiple machine learning approaches.