

*Course Project Report*

**MUHAMMAD SHARJEEL NAEEM.**

2024-MS-DS-102.

Advance Techniques in Data Science

26-12-2024

MSc. Data Science

**UNIVERSITY OF ENGINEERING AND TECHNOLOGY, LAHORE.**

**Report for Crop Yield Dataset Analysis**

This report analyzes a crop yield dataset through multiple stages: uploading, data cleaning, exploratory data analysis (EDA), feature engineering, and predictive modeling. The process involves various steps such as handling missing data, visualizing data, detecting outliers, and building a machine learning model to predict crop yields.

**Problem**

Predicting Crop Yield Based on Environmental and Agronomic Factors

**Problem Statement**

Accurate prediction of crop yield is crucial for improving agricultural efficiency and ensuring food security. By analyzing environmental conditions, soil health, irrigation practices, and other factors, we can predict how much yield a farmer is likely to get from a particular crop in a given season.

**Significance**

This problem holds high relevance for farmers, agricultural planners, and policymakers. Understanding crop yield patterns can help optimize resource allocation, reduce waste, and improve food production efficiency, ultimately leading to enhanced food security and sustainable agricultural practices.

**Objectives**

* 1. Build a predictive model to forecast crop yields.
  2. Identify the main factors influencing crop yield.
  3. Analyze the potential impact of environmental and agronomic factors on yield variability.

**Potential Impact**

Improved predictions of crop yield can help farmers plan better, optimize inputs (e.g., water, fertilizers), and enhance productivity, while also providing valuable insights for agricultural research and policy planning.

Asking the Right Questions

Key Questions

* 1. What are the primary environmental (e.g., temperature, rainfall) and agronomic factors (e.g., soil type, crop variety) influencing crop yield?
  2. How accurately can we predict crop yield using these features?
  3. Are there any seasonal trends or patterns in crop yield that could help improve predictions?
  4. How do irrigation practices and weather conditions correlate with crop yield in different regions?
  5. What are the most important features influencing crop yield?
  6. Are there any strong correlations between variables (e.g., temperature, humidity, precipitation)?
  7. Which crop types or seasons have the highest yields?

**Data Collection**

Crop\_yield\_dataset.csv

**Uploading and Loading the Dataset**

The script begins by using the Google Colab files. Upload () function to upload a CSV file (assumed to be named "Crop yield dataset.csv"). After the file is uploaded, it is read into a panda DataFrame (df) using pd. read\_ csv (). The first few rows and a summary of the dataset are displayed with df. Head () and df. Describe ().

**Data Cleaning**

2.1 Missing Data Handling

The dataset might contain missing values, which are handled in this section:

* Numeric columns: Missing values are filled with the mean of the respective column.
* Categorical columns: Missing values are filled with the most frequent value (mode) of the column.

2.2 Outlier Detection and Removal

Outliers can distort statistical analysis. Two methods are provided to detect and remove outliers from the "Crop\_ Yield" column:

* IQR method: Outliers are defined as values outside 1.5 times the interquartile range (IQR).
* Z-score method: Outliers are defined as values with a Z-score greater than 3.

**Feature Engineering**

3.1 Data Standardization and Normalization

* Standardization: The "Crop\_ Yield" column is standardized (mean = 0, standard deviation = 1) using Standard Scaler.
* Normalization: The "Crop\_ Yield" column is normalized to a range between 0 and 1 using Min Max Scaler.

**Exploratory Data Analysis (EDA)**

4.1 Descriptive Statistics

Summary statistics are displayed using df. Describe ().

4.2 Visualizations

Various visualizations are created to understand the data distribution and relationships:

* Histogram: Distribution of the "Precipitation" column.
* Box plot: Outliers in the "Precipitation" column are visualized.
* Correlation matrix: Displays correlations between numeric variables.
* Pair plot: Visualizes relationships between variables.
* Scatter plots: Examines the relationship between "Precipitation" and "Crop\_ Yield"

**Predictive Analysis**

5.1 Data Preprocessing

* The Crop-Type column is one-hot encoded using One Hot Encoder, converting categorical values into binary columns.
* The original "Crop-Type" column is removed, and features (X) and the target variable (y, "Yield") are defined.

5.2 Train-Test Split

The data is split into training and testing sets (80% for training and 20% for testing) using train\_ \_split.

5.3 Model Training and Evaluation

A Random Forest Regressor is trained on the training data. The model’s performance is evaluated using the R-squared score on the test data. R-squared indicates the proportion of variance in the target variable (crop yield) explained by the model.

**Summary of Key Steps**

1. Data Uploading and Exploration: Load and explore the dataset.
2. Data Cleaning: Handle missing data by filling numeric columns with the mean and categorical columns with the mode.
3. Outlier Detection and Removal: Identify and remove outliers using IQR and Z-scores.
4. Feature Engineering: Standardize and normalize the "Crop\_ Yield" feature.
5. Exploratory Data Analysis (EDA): Visualize data distributions, relationships between variables, and correlations.
6. Predictive Modeling: Build a Random Forest Regressor to predict crop yield and evaluate its performance using R-squared.

**Project code**

https://colab.research.google.com/drive/17hCtGVdI3SygVKDZKjl9cwZh0f3Z7h4i?usp=sharing