Assignment 6

Name: Srinidhi Soundarrajan

Roll No: 281064

Batch: A3

Statement

In this assignment, we aim to:
a) Apply Linear Regression using suitable library functions to predict month-wise temperatures.
b) Compute and display summary statistics for the dataset, including minimum, maximum, mean, range, standard deviation, variance, and percentiles.
c) Assess the performance of the regression model using MSE, MAE, and R-Square metrics.
d) Visualize the simple regression model and feature distributions using histograms.

Objective

1. Understand how to compute summary statistics for a dataset using Python.
2. Learn how to visualize data distributions using histograms.
3. Develop skills in data preprocessing, transformation, and integration for regression modeling.
4. Build and evaluate a Linear Regression model for temperature prediction.

Resources Used

- **Software**: Visual Studio Code
- **Libraries**: Pandas, Matplotlib, Seaborn, Scikit-Learn

Introduction to Pandas and Regression Analysis

Pandas is an open-source Python library designed for data manipulation and analysis. It provides data structures like Series (1D) and DataFrame (2D), making it easier to analyze and process structured data efficiently.

Key Capabilities

- Importing data from various formats (CSV, Excel, SQL).
- Data preprocessing (cleaning, transforming, handling missing values).
- Statistical and analytical operations (descriptive statistics, visualization).
- Machine learning model building (classification, regression).

Basic Functions Used

1. pd.read_csv() – Reads data from a CSV file into a DataFrame.
2. describe() – Generates descriptive statistics for numerical columns.
3. hist() – Creates histograms to visualize data distributions.
4. fillna() – Handles missing values using strategies like mean, median, or mode imputation.
5. train_test_split() – Splits data into training and testing sets.
6. LinearRegression() – Builds a regression model.
7. mean_squared_error(), mean_absolute_error(), r2_score() – Evaluate the regression model's performance.

Methodology

1. Data Collection and Exploration

- Load the temperature dataset into a Pandas DataFrame.
- Examine its structure, including feature types, missing values, and unique categories.

2. Data Preprocessing

- Handle missing values using mean or median imputation.
- Perform feature selection and transformation as necessary.

3. Summary Statistics Computation

- Compute statistics like minimum, maximum, mean, range, standard deviation, variance, and percentiles using describe().

4. Feature Visualization using Histograms

- Use hist() and sns.histplot() to visualize the distributions of different features.

5. Model Building (Linear Regression)

- Split the dataset into training and testing sets using train_test_split().
- Train a Linear Regression model using LinearRegression().
- Make predictions using the test data.
- Evaluate model performance using MSE, MAE, and R-Square metrics.

6. Model Visualization

- Plot the regression model to understand its accuracy.
- Visualize residuals to analyze prediction errors.

Advantages of Pandas and Regression Analysis

1. Pandas simplifies data handling and provides powerful functions for data cleaning and transformation.
2. Visualization tools help in understanding feature distributions and data patterns.
3. Regression analysis allows for effective prediction of numerical variables.

Disadvantages of Pandas and Data Processing

1. Large datasets can be memory-intensive and slow down processing.
2. Data preprocessing can be complex when dealing with unstructured data.

Conclusion

This assignment demonstrated the application of Linear Regression for temperature prediction. We explored data distributions, computed summary statistics, and evaluated model performance using regression metrics. The use of Pandas, Seaborn, and Scikit-Learn enabled efficient data handling and predictive analysis.