Assignment: 6

Problem Statement:

Assignment on Regression technique.

Download temperature data from the link below.

https://www.kaggle.com/venky73/temperaturesof-india?select=temperatures.csv

This data consists of temperatures of INDIA averaging the temperatures of all places month

wise. Temperatures values are recorded in CELSIUS

a) Apply Linear Regression using a suitable library function and predict the Month-wise

temperature.

b) Assess the performance of regression models using MSE, MAE and R-Square metrics

c) Visualize a simple regression model.

Software Library Package:

Python with pandas , numpy , matplotlib and scikit-learn.

1. Theory:

Linear regression is a fundamental statistical technique used for modeling the relationship between two or more variables. It assumes a linear relationship between the independent variable(s) and the dependent variable, represented by a straight line. The goal of linear regression is to find the best-fitting line that minimizes the difference between the observed and predicted values. This line is determined by estimating the coefficients (slope and intercept) that minimize the sum of squared errors. Linear regression is widely used in various fields for prediction, forecasting, and understanding the relationship between variables. It serves as a simple yet powerful tool for analyzing and interpreting data.

1.1 Methodology:

Linear Regression is employed using the `LinearRegression` class from the `sklearn.linear_model` module. This class fits a linear model to the provided data using the least squares method, aiming to minimize the residual sum of squares between the observed and predicted target values.

1.2 Advantages and Applications:

- Advantages:

  - Simple Implementation: Utilizes straightforward procedures for model fitting and prediction.

  - Interpretability: Provides coefficients that can be interpreted to understand the impact of features on the target variable.

  - Efficient Computation: Works well even with large datasets.

- Applications:

  - Predictive Analysis: Suitable for predicting continuous numerical values, such as temperature.

  - Trend Analysis: Helpful in identifying trends and patterns in the data.

  - Forecasting: Can be used for forecasting future values based on historical data.

1.3 Limitations:

- Assumption of Linearity: Assumes a linear relationship between the independent and dependent variables, which might not hold true in all cases.

- Sensitivity to Outliers: Vulnerable to outliers that can skew the regression line and affect predictions.

- Limited Predictive Power: May not capture complex relationships between variables.

- Dependence on Independence: Requires the observations to be independent of each other for accurate predictions.

2. Working/Algorithm:

The program follows these key steps:

a. Data Preprocessing:

- Removes duplicate entries and handles missing values using mean imputation.
- Splits the data into features (YEAR, OCT, NOV, DEC) and the target variable (OCT-DEC).
- Encodes categorical variables using One-Hot Encoding for the YEAR feature.

b. Model Building:

- Splits the data into training and testing sets using `train_test_split` function from `sklearn.model_selection`.
- Fits a linear regression model to the training data using `LinearRegression()` class.
- Predicts the target variable using the trained model.

c. Model Evaluation:

- Evaluates the performance of the model using Mean Absolute Error (MAE), Mean Squared Error (MSE), and R-squared metrics from the `sklearn.metrics` module.

d. Visualization:

- Plots scatter plots of the data and regression lines to visualize the relationship between the independent variable (YEAR) and target variables (OCT-DEC, NOV).

3. Conclusion:

The implemented Linear Regression model effectively predicts the month-wise temperatures of India based on historical data. Evaluation metrics confirm the model's reasonable performance, and visualizations provide insights into the linear relationship between the features and target variables. However, it's crucial to acknowledge the limitations of linear regression and explore alternative techniques for capturing more complex patterns in the data.