Name: Jiya Sharma
Roll no: 281056
PRN: 22311741

Assignment 6

Problem Statement:

Apply Linear Regression using a suitable library function to predict month-wise temperature and evaluate the model using performance metrics.

Objective:

- To apply regression techniques for predicting temperature trends.
- To preprocess and analyze temperature data for better model performance.
- To evaluate model performance using MSE, MAE, and R-Square metrics.
- To visualize the regression model and interpret the results.

Resources used:

- Software used: Visual Studio Code
- Libraries used: Pandas, Matplotlib, Seaborn, SKLearn

Theory:

Regression is a supervised learning technique used to model relationships between a dependent variable (temperature) and one or more independent variables (month). Linear Regression assumes a linear relationship between these variables and fits a straight line that minimizes errors.

Linear Regression Formula:

$y=mx+c$

Where:

- y is the dependent variable (Temperature).
- x is the independent variable (Month).
- m is the slope (rate of change).
- c is the intercept (baseline value)

a) Mean Squared Error (MSE): Measures the average squared differences between actual and predicted values. Lower values indicate better performance.
b) Mean Absolute Error (MAE): Measures the average absolute differences between actual and predicted values.
c) R-Square (R^2): Represents the proportion of variance explained by the model. Closer to 1 indicates a better fit.

Confusion Matrix:

A confusion matrix is a performance measurement tool for classification models. It consists of four components:

- True Positives (TP): Correctly predicted positive cases.
- True Negatives (TN): Correctly predicted negative cases.
- False Positives (FP): Incorrectly predicted positive cases (Type I Error).
- False Negatives (FN): Incorrectly predicted negative cases (Type II Error).

Evaluation Metrics:

- Accuracy: Measures the overall correctness of the model.
- Precision: Measures how many predicted positive cases were actually positive.
- Recall: Measures how many actual positive cases were correctly predicted.
- F1-Score: Harmonic mean of precision and recall, balancing both metrics.

Methodology:

1. **Data Preprocessing:**

- Load the dataset using Pandas.
- Handle missing values (imputation or removal).
- Encode categorical variables (e.g., gender) using one-hot encoding.
- Normalize numerical features using MinMaxScaler or StandardScaler.
- Split the dataset into training and testing sets (e.g., 75% training, 25% testing).

2. **Choosing the ML Algorithm:**

Since the problem is a binary classification task, suitable algorithms include:

- Logistic Regression
- Decision Tree Classifier
- Random Forest Classifier
- Support Vector Machine (SVM)
- K-Nearest Neighbors (KNN)
- Neural Networks (optional for advanced modeling)

3. **Model Training & Prediction:**

- Train the selected ML model on the training dataset.
- Predict customer responses on the test dataset.

4. **Confusion Matrix & Performance Metrics Calculation:**

- Compute the confusion matrix (True Positives, True Negatives, False Positives, False Negatives).
- Derive the following metrics from the confusion matrix:

a) Accuracy
b) Precision
c) Recall (Sensitivity)
d) F1-Score

Conclusion:

- The regression model effectively predicts monthly temperatures with a reasonable error margin.
- The evaluation metrics provide insight into model accuracy.