**Assignment No:-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.

**Theory**:-Linear Regression: It's a statistical technique for forecasting analysis. Predictions are made using linear regression for continuous, real, or numerical variables like sales, earnings, age, and product price, among others.   
The term "linear regression" refers to a procedure that displays a linear relationship between one or more independent (y) variables and a dependent (y) variable.   
Given that linear regression displays a linear relationship, it can be used to determine how the value of the independent variable affects the value of the dependent variable.

Mathematically these slant lines follow the following equation,

Y = m\*X + b

Where X = dependent variable (target)

Y = independent variable

m = slope of the line (slope is defined as the ‘rise’ over the ‘run’)

However,[machine learningOpens a new window](https://www.unite.ai/what-is-machine-learning/) experts have a different notation to the above slope-line equation,

y(x) = p0 + p1 \* x

where,

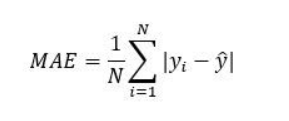
* y = output variable. Variable y represents the continuous value that the model tries to predict.
* x = input variable. In [machine learning](https://www.spiceworks.com/tech/artificial-intelligence/articles/what-is-machine-learning/), x is the feature, while it is termed the independent variable in statistics. Variable x represents the input information provided to the model at any given time.
* p0 = y-axis intercept (or the bias term).
* p1 = the regression coefficient or scale factor. In classical statistics, p1 is the equivalent of the slope of the best-fit straight line of the linear regression model.
* pi = weights (in general).

Thus, regression modeling is all about finding the values for the unknown parameters of the equation, i.e., values for p0 and p1 (weights).

The equation for multiple linear regression is similar to the equation for a simple linear equation, i.e., y(x) = p0 + p1x1 plus the additional weights and inputs for the different features which are represented by p(n)x(n). The formula for multiple linear regression would look like,

y(x) = p0 + p1x1 + p2x2 + … + p(n)x(n)

The Mean absolute error represents the average of the absolute difference between the actual and predicted values in the dataset. It measures the average of the residuals in the dataset.



Mean Squared Error represents the average of the squared difference between the original and predicted values in the data set. It measures the variance of the residuals.

A mathematical equation with numbers and symbols

Description automatically generated

Root Mean Squared Error is the square root of Mean Squared error. It measures the standard deviation of residuals.

A math equations with numbers and symbols

Description automatically generated with medium confidence

The coefficient of determination or R-squaredrepresents the proportion of the variance in the dependent variable which is explained by the linear regression model. It is a scale-free score i.e. irrespective of the values being small or large, the value of R square will be less than one.

A mathematical equation with numbers and symbols

Description automatically generated

Linear Regression Types:

* Basic Linear Regression:   
  A linear regression procedure is referred to as simple linear regression if it uses one independent variable to predict the value of a number of dependent variables.
* Multiple Linear Regression: This type of linear regression method is employed when multiple independent variables are combined to predict the value of a numerical dependent variable.

**Advantages:**

1. Interpretability: Easy interpretation of coefficients.
2. Simplicity: Simple to implement and understand.
3. Efficiency: Computationally efficient for large datasets.
4. Assumption Testing: Allows for testing of key assumptions.
5. Feature Importance: Provides insights into feature importance.

**Disadvantages:**

1. Assumes linearity: Regression models, particularly linear regression, assume a straight-line relationship between the independent and dependent variables. If the underlying relationship is more complex (curved, exponential, etc.), the model may not accurately capture the true association.
2. Sensitive to outliers: Outliers (data points significantly different from the majority) can disproportionately influence the regression line, leading to misleading results.
3. Doesn't establish causation: Even if a strong correlation is found between variables, regression models cannot determine causality. There might be a third, unseen variable influencing both the independent and dependent variables, creating a false association.

**Applications with example:**

1. Student grades determined by the number of hours studied (ideally):In this case, exam scores are dependent on the number of hours studied, but the number of hours studied is independent.
2. Estimating agricultural yields using rainfall data: The measure of precipitation is an independent variable, and yield is a dependent variable.
3. Estimating an individual's salary based on years of experience: Experience is now the independent variable, and salary is the dependent variable.

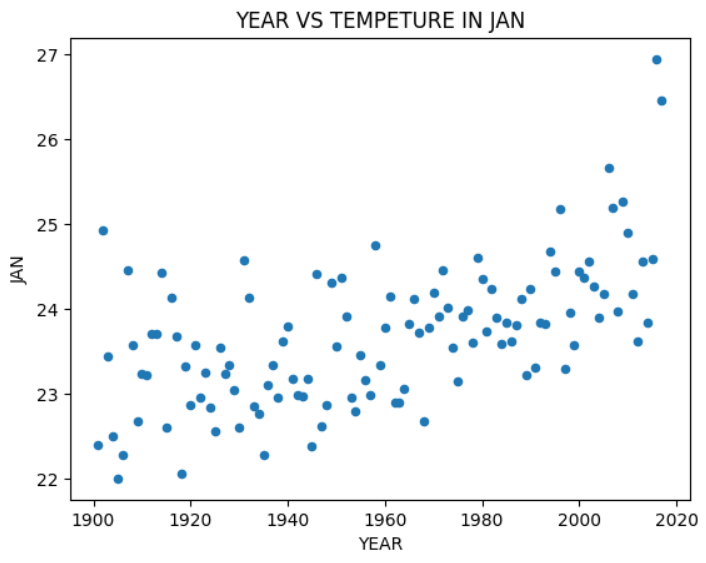
**Methodology:**

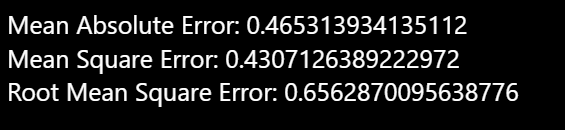
1. Loading the Data:
   * Download the temperature data from the provided link and load it into a DataFrame using a suitable library like pandas in Python or read.csv in R.
2. Data Preprocessing:
   * Check for any missing values or outliers in the dataset and handle them appropriately. Convert the categorical variables like month into numerical format if needed.
3. Splitting the Data:
   * Split the data into features (X) and target variable (y), where X represents the independent variables (e.g., month) and y represents the dependent variable (temperature).
4. Applying Linear Regression:
   * Use a suitable library function (e.g., sklearn.linear\_model.LinearRegression in Python or lm() function in R) to apply linear regression on the dataset.
5. Predicting Month-wise Temperature:
   * Fit the linear regression model on the training data and use it to predict the month-wise temperature for the test data.
6. Assessing Model Performance:
   * Calculate metrics such as Mean Squared Error (MSE), Mean Absolute Error (MAE), and R-Squared to assess the performance of the regression model.
7. Visualizing the Regression Model:
   * Plot the actual vs. predicted temperatures to visualize how well the regression model fits the data.

Advantages and Disadvantages & Limitation/Example:

1. Advantages:
   * Interpretability: Linear regression models are easy to interpret, making them suitable for explaining relationships between variables.
   * Simple and Fast: Linear regression is computationally efficient and easy to implement.
   * Versatility: Can be applied to both small and large datasets and can handle multiple input variables.
2. Disadvantages & Limitations/Example:
   * Assumption of Linearity: Linear regression assumes a linear relationship between the independent and dependent variables, which may not always hold true.
   * Sensitivity to Outliers: Linear regression models are sensitive to outliers in the data, which can skew the results.
   * Limited Complexity: Linear regression models may not capture complex relationships between variables compared to more advanced regression techniques.

**Diagram**

****

****

**Conclusion**  
  
In summary, the application of linear regression on the temperature dataset from India facilitated the prediction of month-wise temperatures. Evaluation of regression model performance through metrics such as MSE, MAE, and R-Square provided insights into the accuracy and effectiveness of the predictions. Additionally, visualization of the regression model enhanced the understanding of the relationship between independent and dependent variables, aiding in interpretation and decision-making processes.