**ASSIGNMENT: - 06**

**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 place’s 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.

**S/W, Library and Package:**

1. Software: Python
2. Library: scikit-learn (sklearn) - for linear regression model and metrics calculation
3. Package: pandas - for data manipulation and preprocessing

**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.

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.

Applications of Simple Linear Regression:

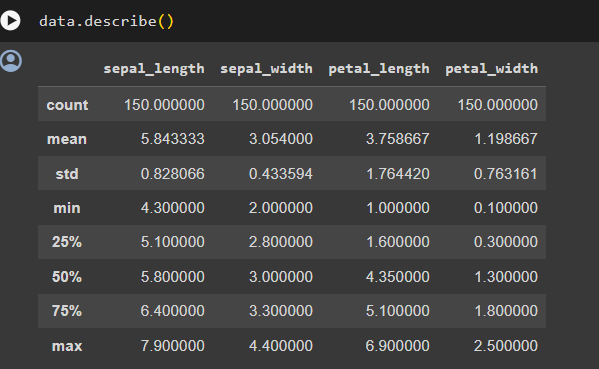
1. Student grades determined by the number of hours studied (ideally):In this case, exam scores are dependent on the number of hours studied, but the number of hours studied is independent.
2. Estimating agricultural yields using rainfall data: The measure of precipitation is an independent variable, 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.

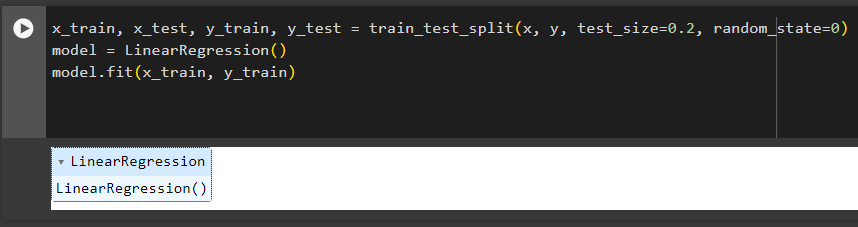
Limitations of Simple Linear Regression:

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.

**Working/ Algorithm:**

1. Data Loading and Initial Exploration:
   * Import the required libraries: Pandas, NumPy, Matplotlib.
   * Load the Iris dataset from a CSV file.
   * Display the first few rows of the dataset to understand its structure.
   * Check basic information about the dataset (columns, data types).
   * Perform a descriptive statistical analysis to get an overview of the data.
2. Data Cleaning and Preprocessing:
   * Remove any duplicate rows from the dataset.
   * Check for missing or null values in the dataset.
   * Handle missing values using an imputation strategy (SimpleImputer with mean strategy).
   * Separate the dataset into independent variables (**x**: 'sepal\_length', 'sepal width', 'petal\_length') and dependent variable (**y**: 'class').
   * Split the data into training and testing sets using train\_test\_split.
3. Model Training, Prediction, and Evaluation:
   * Create a Linear Regression model object.
   * Train the model using the training data (x\_train, y\_train).
   * Make predictions (y\_pred) using the trained model on the test data (x\_test).
   * Evaluate the model's performance by printing its coefficients and intercept.
   * Visualize the model's predictions by plotting a scatter plot of the test data and overlaying the regression line.





**Conclusion:**

Linear regression is a versatile and widely used statistical technique with several advantages, such as interpretability, simplicity, and efficiency. It finds applications in predictive modeling, risk assessment, marketing analysis, and resource planning. However, it has limitations like the assumption of linearity, sensitivity to outliers, and potential overfitting/underfitting issues. Overall, linear regression provides valuable insights and predictive capabilities but should be used judiciously considering its assumptions and limitations.