MIT School of Computing
Department of Information Technology
Machine Learning Lab
Assignment no: -3

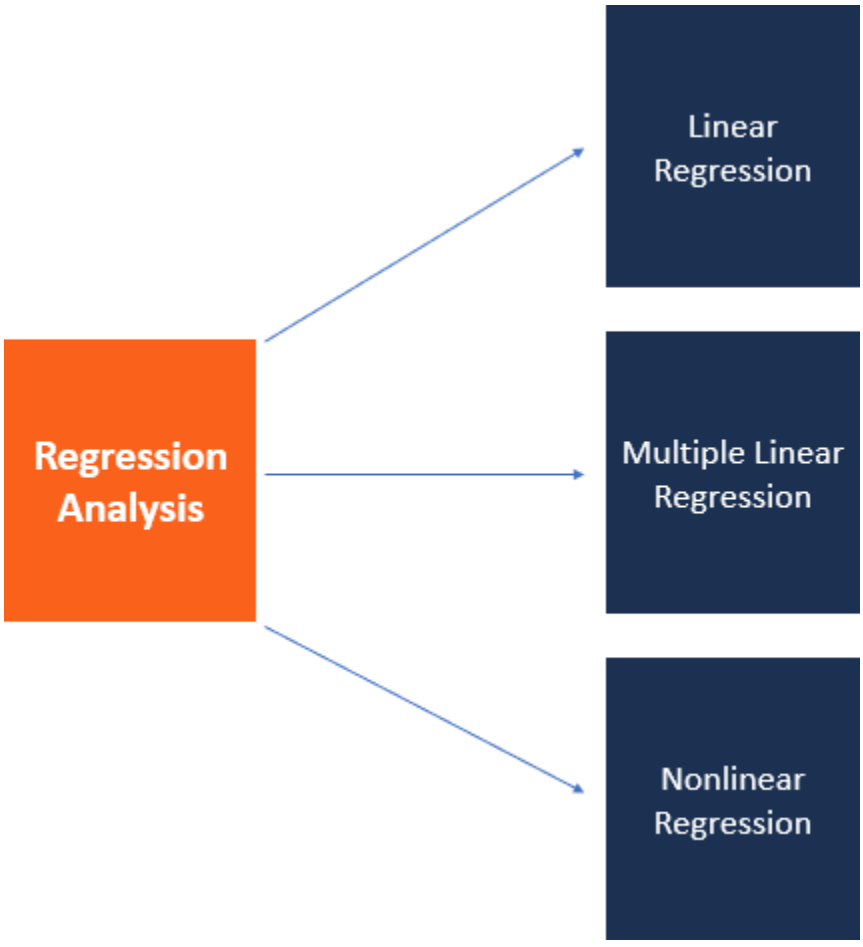
AIM: Perform a regression model and also calculate the mean error for any dataset.:

INDEX TERMS: Data Preprocessing, Regression Model, Mean error.

PROBLEM DEFINITION: Run a regression model on train data and compute the mean error in the test set.

Theory:

Regression Analysis: Regression analysis is a set of statistical methods used for the estimation of relationships between a dependent variable and one or more independent variables. It can be utilized to assess the strength of the relationship between variables and for modeling the future relationship between them.



Regression analysis includes several variations, such as linear, multiple linear, and nonlinear. The most common models are simple linear and multiple linear. Nonlinear regression analysis is commonly used

for more complicated data sets in which the dependent and independent variables show a nonlinear relationship.

Regression Analysis – Linear Model Assumptions

1. The dependent and independent variables show a linear relationship between the slope and the intercept.
2. The independent variable is not random.
3. The value of the residual (error) is zero.
4. The value of the residual (error) is constant across all observations.
5. The value of the residual (error) is not correlated across all observations.

The residual (error) values follow the normal distribution

Regression Analysis – Simple Linear Regression

Simple linear regression is a model that assesses the relationship between a dependent variable and an independent variable

Regression Analysis – Multiple Linear Regression

Multiple linear regression analysis is essentially similar to the simple linear model, with the exception that multiple independent variables are used in the model.

Mean error:

The mean error is an informal term that usually refers to the average of all the errors in a set. An "error"in this context is uncertainty in measurement or the difference between the measured value and true/correct value. The more formal term for error is measurement error, also called observational error.

R-Squared : a statistical measure between 0 and 1 which calculates how similar a regression line is to the data it's fitted to. If it's a 1, the model 100% predicts the data variance; if it's a 0, the model predicts none of the variance.

R-Squared = Explained variance of the model / Total variance of the target variable

Algorithm/steps:

Step1: Importing the required

librariesStep2: Loading the dataset

Step3: Let's check for any missing or NA values in the training and testing data set

Step4: Let's drop the record with the missing value in the training dataset. As it is only one record, removing it will not be much of a concern.

Step5: Let's define our dependent and independent variable for training and testing data

Step6: Let’s add a constant, to add a constant we will create a new variable.

Let’s define the model and fit it.

Step7: Let’s look at different parameters of the model summary and interpret it

Step8: Observe R2 score .

Write your comment in conclusion