School of Computer Science Engineering and Technology

Course- BTech	Type- Specialization Core
Course Code- CSET228	Course Name- Data Mining and Predictive Modelling
Year- 2024-25	Semester- Even
Date- 10/02/2025	Batch-

Lab # No. (6.1)

Regression and error analysis

CO Mapping

Exp No.	Name	CO1	CO2	CO3
Q1, Q2, Q3	Regression and error Analysis		√	

Objectives

1. Students will be able to gain a deeper understanding Regression Analysis.

**Q1.** Explore regression analysis.   (40 Minutes)

a. Create a regression model using 'citympg' as the dependent variable and 'horsepower' as the independent variable. Report all model statistics and comment on the model diagnostics.
b. Create the same regression model keeping 'highwaympg' as the dependent variable and 'horsepower' as the independent variable. Report all model statistics and comment on the model diagnostics.
c. Include a basic scatterplot of the data for both above as well.
d. Create another regression model using 'citympg' and 'highwaympg' as independent variables separately and 'price' as the dependent variable. Examine in your both regression models whether these data are positively or negatively correlated or not.
e. Create a regression model using two other variables from the 'CarPrice_Assignment.csv' dataset (i.e., not 'highwaympg', 'citympg' or 'horsepower' as independent variable). You can keep 'price' as dependent variable. Take any other two separately as independent variable and create two separate regression model if it can be ensured the existence of a reasonable underlying connection between the independent and dependent variables.
f. Report all model statistics and comment on the model diagnostics.   **(20 Min)**

   https://archive.ics.uci.edu/ml/datasets/Automobile

Q 2. Use auto-mpg dataset, attribute information of dataset is given in auto-mpg.txt file.

a) Clean dataset i.e., adjust the missing values. Select cylinder, horsepower, mpg and acceleration columns and create a new dataset.
b) Plot a diagram between (mpg, acceleration) and (cylinder and acceleration). Create a train and test dataset by dividing the data in 80% and 20% respectively.
c) From the train dataset, plot (horsepower, acceleration), use LinearRegression() from linear_model in sklearn library to train the model. For training take "Horsepower" as train_x and "Acceleration" as train_y and finally evaluate the model using testing dataset. .   **(20 Min)**

School of Computer Science Engineering and Technology
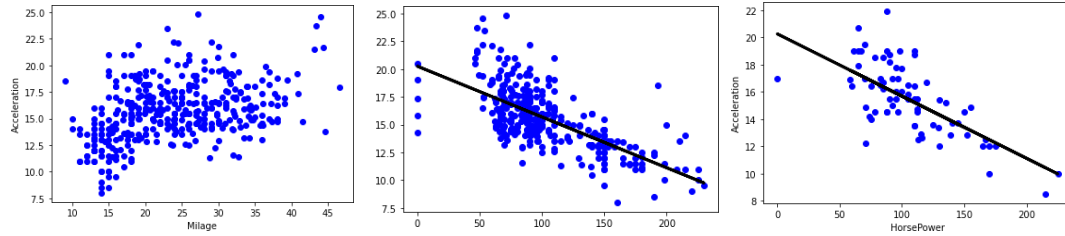
Q 3. Create two columns using random function. First as actual values and second as predicted values.

a) On these values find accuracy, f1-score and precision respectively.

b) Create a confusion matrix and display the confusion matrix. Display the error graph between actual and predicted values.

c) Using confusion matrix find True positive, False positive, True Negative and False Negative. Using these values plot a ROC curve. **(20 Min)**