Assignment No. 4

Name - Kuldharan Sumit Dattatraya
Class - TE
DIV - 4
Subject - DSBDAL

Problem statement -

Create a Linear Regression Model using Python/R to predict home prices using Boston Housing Dataset (https://www.kaggle.com/c/boston-housing). The Boston Housing dataset contains information about various houses in Boston through different parameters. There are 506 samples & 14 feature variables in this dataset.

The objective is to predict the value of prices of the house using the given features.

Theory -

1) Explain Regression.

→ Regression is defined as a method of estimating the value of one variable when that of the other is known & the variables are correlated.

Regression analysis is used to predict or estimate one variable in terms of the other variables.

It is useful in statistical estimation of demand curves, Supply curves, Production function, Cost function, etc.

Types of Regression -

1. Simple Regression & Multiple Regression
2. Linear Regression & Nonlinear Regression.

2) Explain Linear Regression.

→ Linear Regression is the supervised machine learning model in which the model finds the best fit linear line between the independent & dependent variable i.e. it finds the linear relationship between the dependent & independent variable.

Linear Regression is of two types -

① Simple Linear Regression

② Multiple Linear Regression.

① Simple Linear Regression -

In this regression only one independent variable is present & the model has to find the linear relationship of it with the dependent variable.

② Multiple Linear Regression -

In this regression there are more than one independent variables for the model to find the relationship.

Equation for simple linear Regression

$$y = b_0 + b_1 x$$

Equation for multiple linear regression

$$y = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3 \cdots + b_n x_n$$

3) What is Scikit-learn library.

→ scikit-learn is probably the most useful library for machine learning in Python. The sklearn library contains a lot of efficient tools for machine learning & statistical modeling including classification, regression, clustering and dimensionality reduction.

4) How to use heatmap function from the Seaborn library to plot the correlation matrix.

→ ① Import all required modules first.
② Import the file where your data is stored.
③ Plot a heatmap
④ Display it using matplotlib.

Syntax -

```
heatmap(data, Vmin, Vmax, center, cmap, .....)
```

Example -

```
import matplotlib.pyplot as mp
import pandas as pd
import seaborn as sb
data = pd.read_csv("data.csv")
print (data.corr())
dataplot = sb.heatmap(data.corr(), cmap= "YIGnBU"
                      annot = True)
mp.show()
```

5) What is scatterplot and how to use it.

→ In scatter plot, the values of two variables are plotted along two axes and the resulting pattern can reveal correlation present between the variables if any.

A scatterplot is also useful for assessing the strength of the relationship and to find if there are any ~~conflicts~~ outliers in the data.

Code

```
import numpy
import matplotlib.pyplot as plt
x = numpy.random.normal (5.0, 1.0, 1000)
y = numpy.random.normal (10.0, 2.0, 1000)
plt.scatter(x, y)
plt.show()
```

6) What is mean_absolute_error?

→ Mean absolute error calculates the average difference between the calculated values and actual values. It is also known as scale-dependent accuracy as it calculates error in observations taken on the same scale.