Assignment No. 4

= Title of assignment :- create a linear regression model using python/R to predict home prices using boston housing dataset.
Boston housing dataset contains information about various houses.

= Objective of the assignment :- Students should be able to data analysis using linear regression using python for any open source dataset.

= prerequisites :-
1) Basic of python programming
2) concept of regression

= Concept of theory :-
1) linear regression : Univariate & Multivariate.
2) Least square method for linear regression.
3) Measuring performance of linear regression.
4) Example of linear regression
5) Training dataset & testing data set.

= Linear regression :-
It is machine learning algorithm based on supervised learning. It targets prediction value on the basis of independent variables.

$$Y = mX + b + e$$

= Multivariate regression :- It concern the study of two or more predictor variables usually a transformatn of original features into polynomial features from a given degree.

$$Y = a + bX + cx_2$$

2) Least square method for linear regression :-

- Linear regression involves establishing linear relationships between dependent & independent variables. such a relationship ~~between~~ is portrayed in the form of an equatⁿ also known as linear model..

3) Measuring performance of linear Regression Mean square error:-

The mean squared error represents the error of the estimator or predective model created based on the given set of observations in the sample. Two or more regression model created using a given sample data can be compared based on their MSE.

$$MSE = \frac{1}{n} \Sigma (y - \hat{y})^2$$

An MSE of zero (0) represents the fact that the predictor is a perfect predictor. RMSE.

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{1}{n} (\hat{y}_i - y_i)^2}$$

RMSE - least square regression method

Edureka R-squared:-

$$SST = \sum_{i=1} (y_i - \bar{y})^2$$

$$SSR = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

$$R^2 = \frac{SSR}{SST} = \frac{\Sigma(\hat{y}_i - \bar{y})^2}{\Sigma(y_i - \bar{y})^2}$$

A value of R-squared closer to 1 would mean that regression model covers most part of variance.

4) Example of linear regression:-

student	score in x standard	score in xII std.
1	95	85
2	85	95
3	80	70
4	70	65
5	60	70

$$\hat{y} = \beta_0 + \beta_1 x$$

$$\beta_1 = 476/730 = 0.644$$

$$\beta_0 = 77 - (0.644 \times 78) = 26.768$$

$$\hat{y} = 26.76 + 0.644x$$

= Integration of regression line :-

Increase in value of x by 0.644 units

Integrat^n 2; if $x=0$ value of independent variable, it is expected that value of y is 26.768.

if student score is 65 in std x.

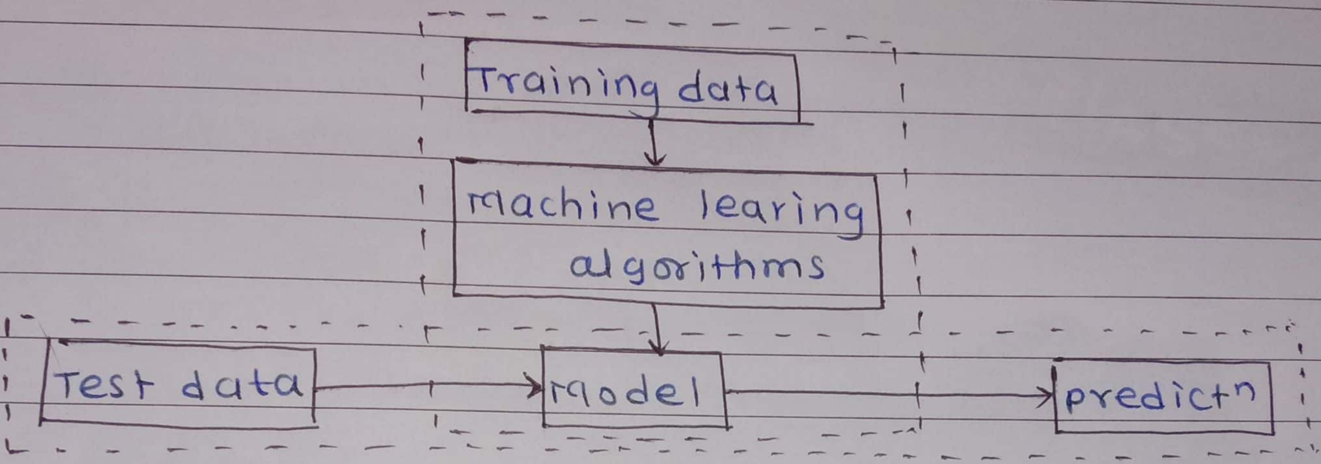
for $x = 80$

$$\hat{y} = 26.76 + 0.644 \times 65 = 68.38$$

= training data set & testing data set.

- ML algorithm has two phases.

1) Training & testing.



a) Training phase :-

- Training dataset is provided as input to this phase.
- Training dataset is dataset having attributes & class labels & used for training ML.

Testing phase:

- Testing dateset is provided as input to this phase.
- Test dataset is a dataset ~~for~~ which class label is unknown. it is tested using model.
- A test dataset used ~~for~~ assessment of the finall chosen model.

Generalizatn :-

- Is predictn of future based on past system.
- It needs to generalize beyond training data to some future data.