Assignment 1

Aim:-

Predict the price of uber ride from a given pickup point to the agreed drop-off locations, perform the following tasks on it. Pre-process dataset, identify outliers. Check the correlation. Implement linear regression model and forest regression models.
Evaluate the models and commute their respective scores like R2, RMSE.

Requirements:-

Python installation, jupyter notebook, python libraries - pandas, numpy, matplotlib, seaborn

Theory:-

Data Pre-processing:

It refers to the techniques of preparing, cleaning and organizing the raw data to make it suitable for a building and training machine learning models.

Steps involved are:-

1) Getting the dataset.
2) Importing libraries.
3) Importing dataset.
4) Finding missing dataset
5) Encoding categorical data

6) Splitting dataset into training and testing

Data cleaning:

It is the process of adding missing data and correcting pre-processing or remaining incorned data

It involves -

- Ignoring the tuples.
- Regression
- Manually filling in missing data.
- Clustering

Data Transformation:

It is the process of turning the data into the prepared formats that you will need for -

- Aggregation
- Normalization
- Decentralization
- R features selection.

Data Reduction:

It is the process of reducing the data size without decreasing its significance. It also helps to actdown the data storage

Outliers:

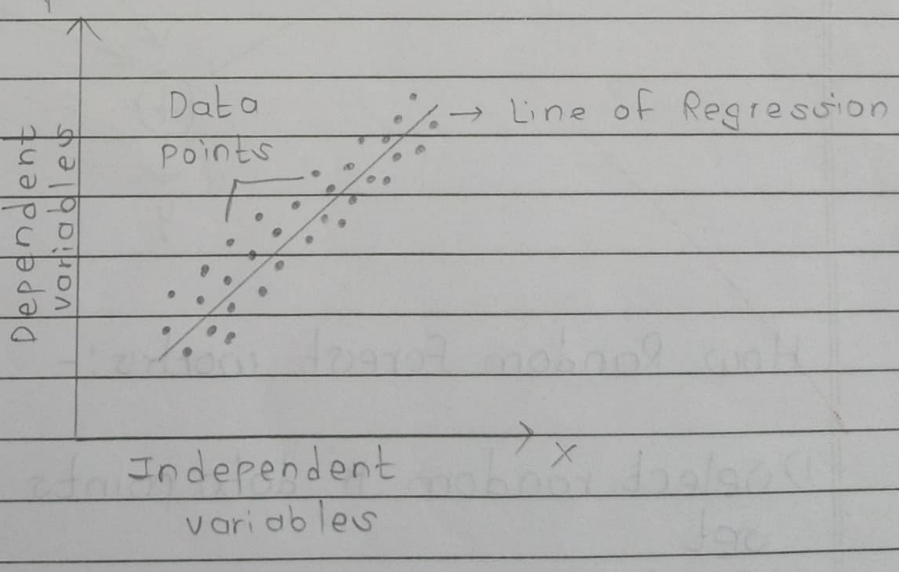
Outliers are the values that look different from the other values in the data Outliers removed may inflate the error metrix.

Linear Regression:

It is one of the easy test and most popular machine learning algorithm.
It is a statistics that is closed for predictive analysis.
It makes prediction for continuous or numerous variables such as sales, age, price, etc.

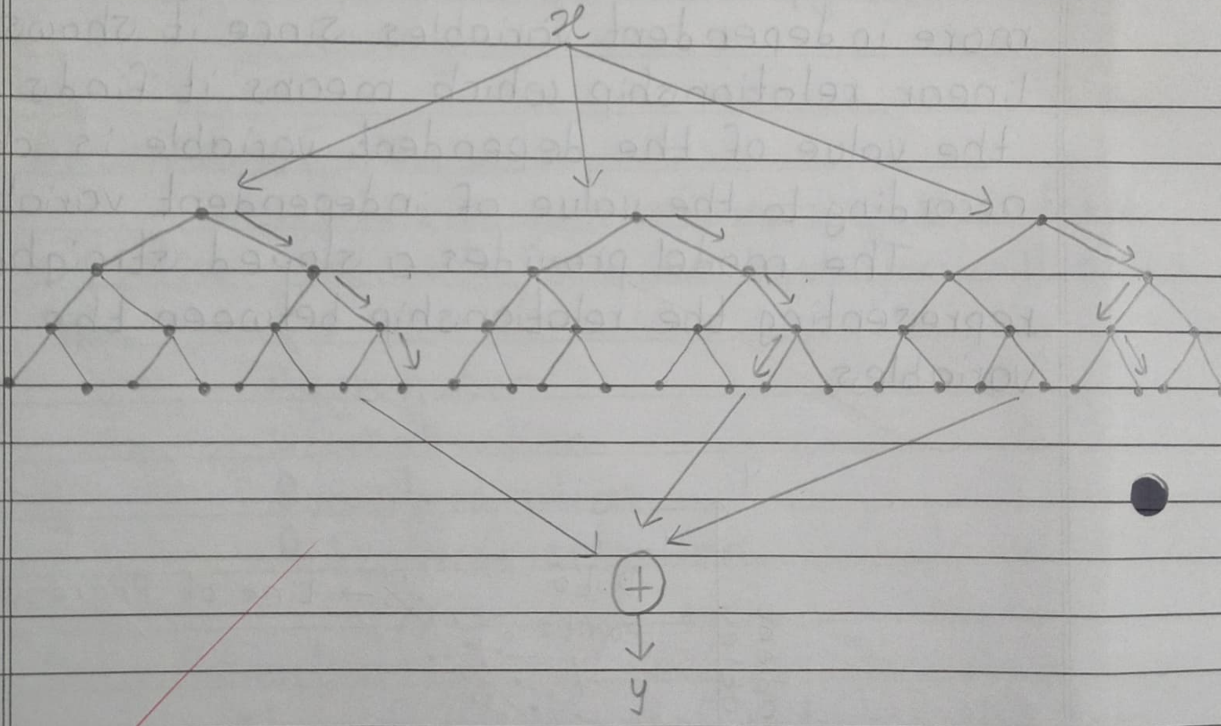
Linear regression algorithm shows a linear relationship between a dependent (y) and one or more independent variables since it shows linear relationship which means it finds how the value of the dependent variable is changing according to the value of independent variable.

The model provides a sloped straight line representing the relationship between the variables.



Random Forest Regression:

Every decision tree has high various variance but whereas we combine all of them together in parallel then resultant variance is low as each decision tree gets perfectly trained on that particular simple data and hence output does not depend on one decision tree but on multiple.



How Random Forest works:-

1) Select random 'K' data points from the training set.
2) Build the decision tree associated with selected data points.

3) Choose the number 'N' for decision trees that you want to build.
4) Repeat steps 1 and 2.
5) For new data points, find the prediction of each decision tree and assign the new data points to the category that wants to win the majority votes.

Conclusion:-

Hence, we have implemented linear regression model on the dataset.