

Parle Tilak Vidyalaya Associations

# SATHAYE COLLEGE (Autonomous)

Vile-Parle (East), Mumbai – 400 057.

Practical Journal

Big Data Analytics

Submitted by

Rakesh Kumar Budhiram

Roll No.: 15

M.Sc. [I.T.]-Information Technology Part I

2022 – 2023



Parle Tilak Vidyalaya Association’s

SATHAYE COLLEGE (Autonomous)

Vile-Parle (East), Mumbai – 400 057.

# CERTIFICATE

This is to certify that \_ Rakesh Kumar Budhiram\_\_\_\_\_\_

Seat No \_\_\_\_15\_\_\_\_\_\_ has successfully completed all the practical in the subject of \_Big Data Analytics\_ for M.Sc.I.T. Part-I SEM – II as prescribed by University of Mumbai for the year 2022-2023.

Coordinator Professor in Charge External Examiner

M.Sc. [I.T.]

Date: Date Date

INDEX

|  |  |  |  |
| --- | --- | --- | --- |
| Sr.No | Topic | Date | Sign |
| 1 | Implementing k-means classification Technique. | 08-Feb-  2022 |  |
| 2 | Implementing Linear Regression | 15-Feb-  2022 |  |
| 3 | Implementing Logistic Regression | 28-Feb-  2022 |  |
| 4 | Implement an application that stores big data in MongoDB and manipulate it using python. | 21-Mar-  2022 |  |
| 5 | Implement SVM classification Technique. | 06-Apr-  2022 |  |
| 6 | Implement Decision Tree classification Technique | 30-Apr-  2022 |  |

## PRACTICAL NO.1

Aim: Implementing k-means classification Technique.

Description :- The algorithm will categorize the items into k groups of similarity. To calculate that similarity, we will use the euclidean distance as measurement. The algorithm works as follows: First, we initialize k points, called means, randomly. We categorize each item to its closest mean and we update the mean’s coordinates, which are the averages of the items categorized in that mean so far. We repeat the process for a given number of iterations and at the end, we have our clusters.

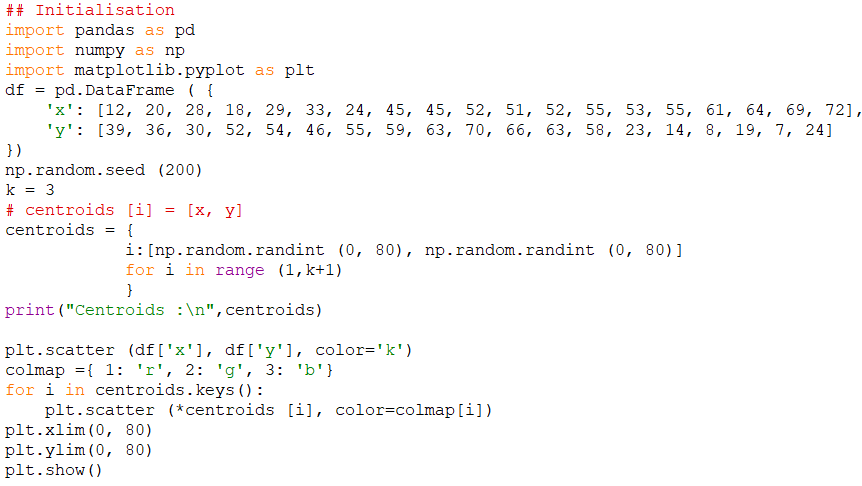
Methods :-

1. numpy.random.randint(low, high=None, size=None) : Return random integers from low (inclusive) to high (exclusive).
2. matplotlib.pyplot.figure(figsize=(x,y)) :Create a new figure, or activate an existing figure.
3. matplotlib.pyplot.scatter(x, y, color =’k’) : With Pyplot, you can use the scatter() function to draw a scatter plot.
4. matplotlib.pyplot.xlim(\*args, \*\*kwargs) : The xlim() function in pyplot module of matplotlib library is used to get or set the xlimits of the current axes.
5. matplotlib.pyplot.ylim(\*args, \*\*kwargs) : The ylim() function in pyplot module of matplotlib library is used to get or set the ylimits of the current axes.
6. matplotlib.pyplot.show() : This method is used to display the graph.
7. df.head() : This method is used to obtain size of the dataset.

Here we are taking 3 number of cluster I.e Red, Green, Yellow.

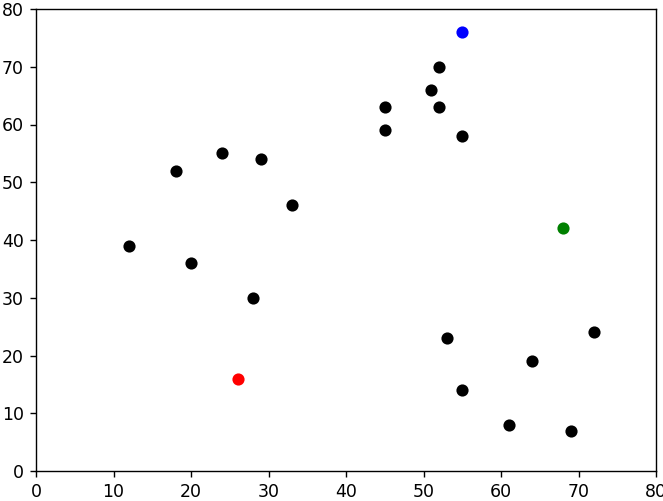
Program Code :

**1. Initialize Stage**



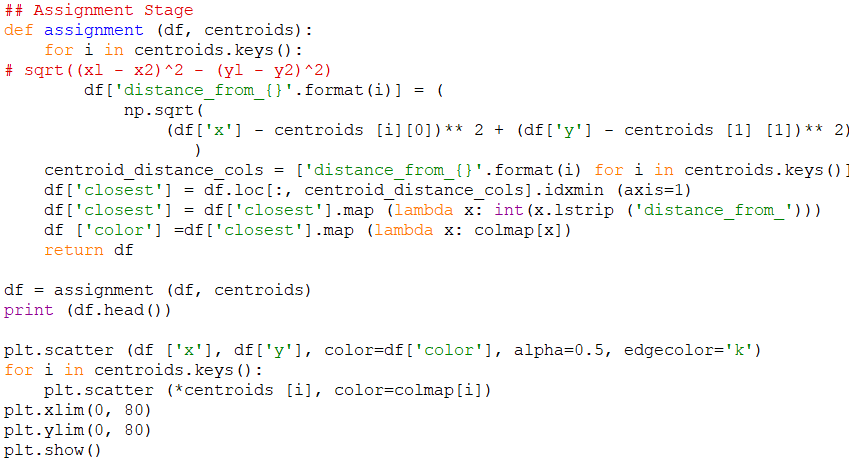
Output :



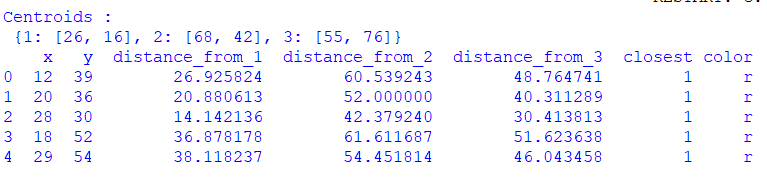


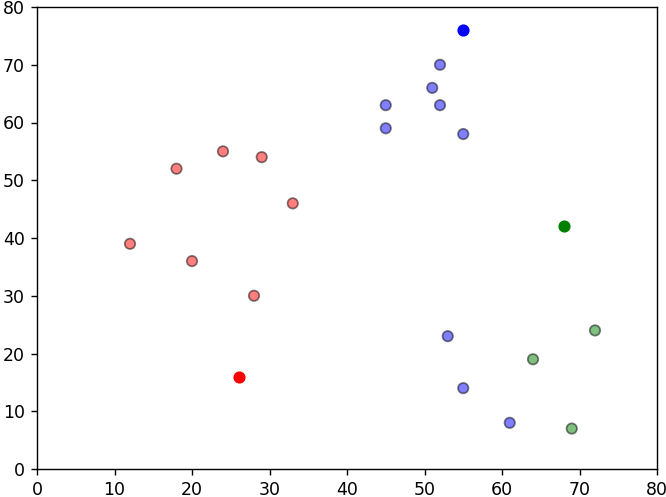
**Stage 2 : Assignment stage.**

**Code :**

****

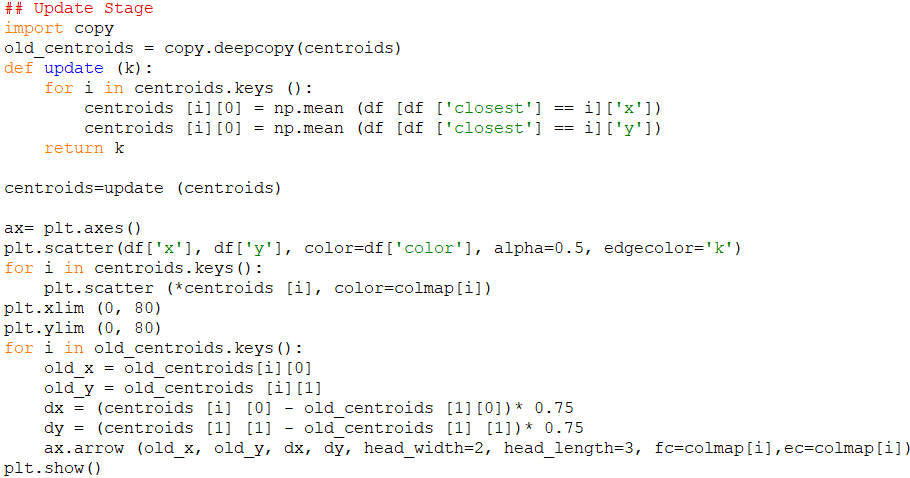
Output :



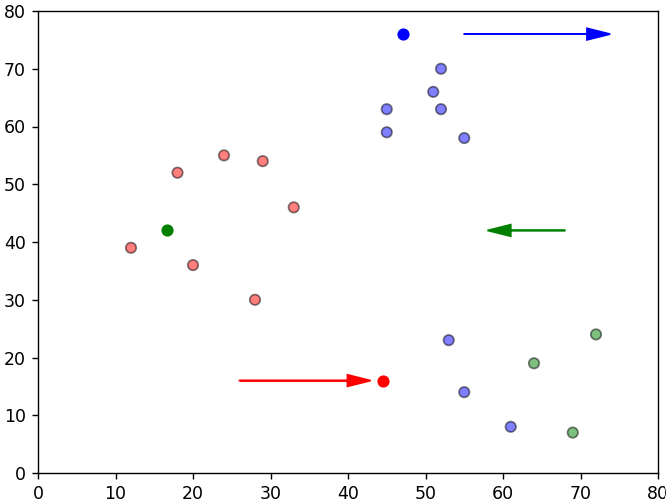


**Stage 3: Update Stage.**

**Code :**

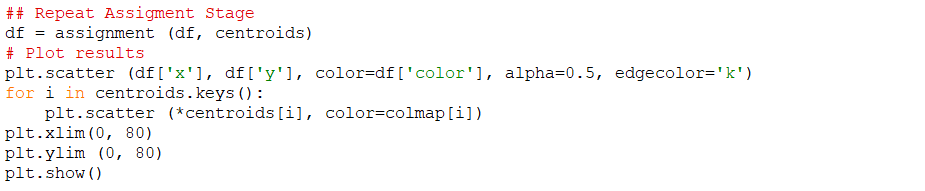
****

Output :

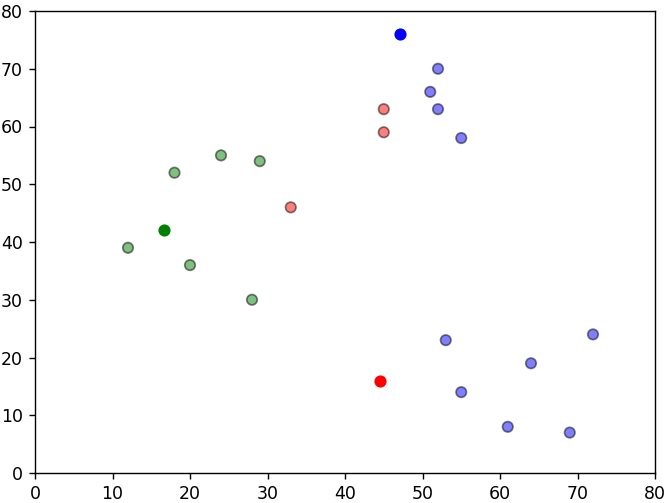


**Stage 4: Repeat Assignment stage**

**Code:**

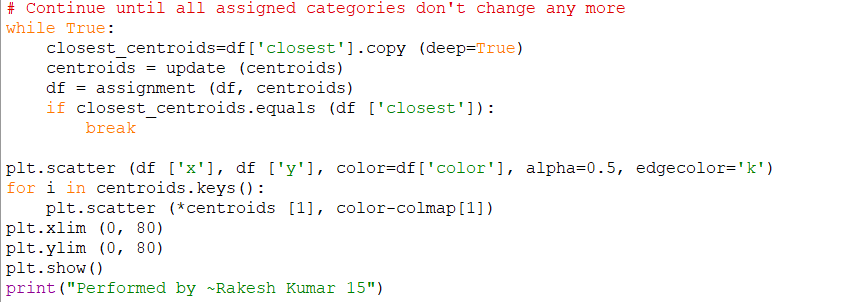
****

Output :

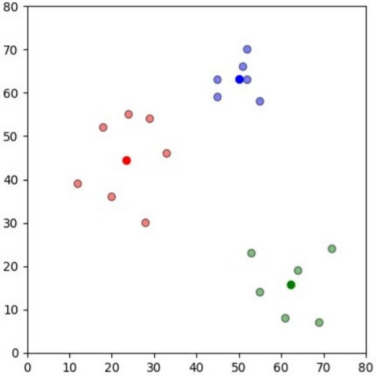


**Stage 5 :** Continue until all assigned Categories don’t change anymore. \

Code :



Output :



Output Conclusion : So we have 3 clusters with 3 means at the centre of these clusters.

**PRACTICAL NO.2**

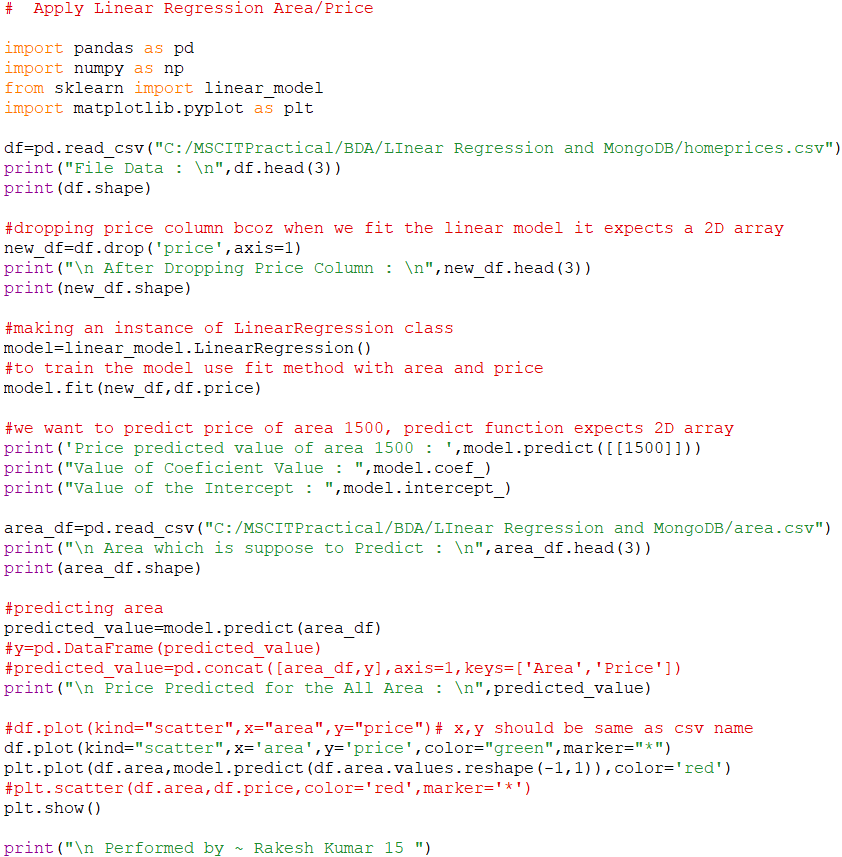
Aim: Implementing Linear Regression

Description: - Linear regression is one of the easiest and most popular Machine Learning algorithms. It is a statistical method that is used for predictive analysis. Linear regression makes predictions for continuous/real or numeric variables such as sales, salary, age, product price, etc. Linear regression algorithm shows a linear relationship between a dependent (y) and one or more independent (y) variables, hence called as linear regression. Since linear regression shows the linear relationship, which means it finds how the value of the dependent variable is changing according to the value of the independent variable.

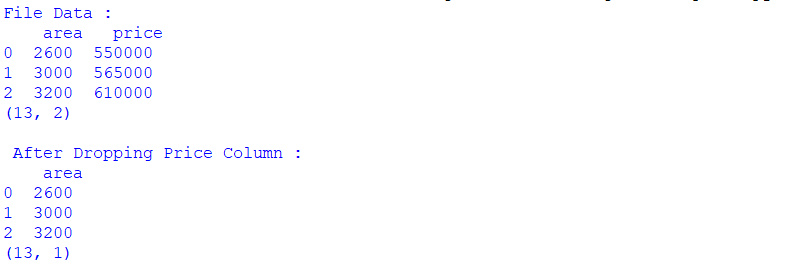
Methods :-

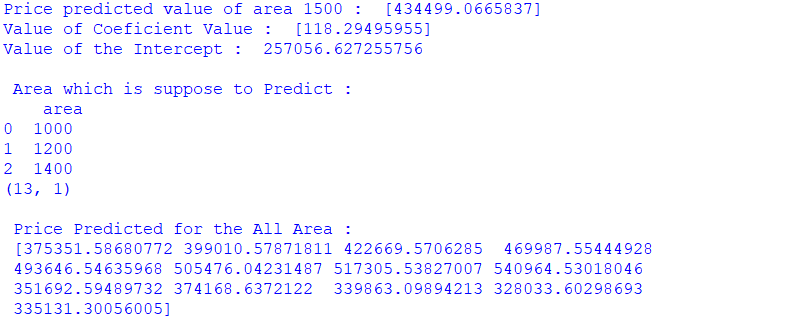
1. pd.read\_csv(inputfilename) : This method is used to read the csv files.
2. dataframe.iloc[:,[colno\_1,colon\_3]] : This method is used to fetch specific row of specific columns.
3. train\_test\_split(x,y,test\_size=0.25,random\_state=0) : This method is used to split dataframe into training and testing dataset.
4. StandardScaler() :This method is used for feature scaling.
5. SVC(kernel='linear', random\_state=0) : This method is used for linear support vector classifier.
6. metrics.accuracy\_score(y\_test,y\_pred) : This method is used to check the accuracy score.
7. model.coef\_ : model.coef\_ is used to obtain coefficient value.
8. model.intercept\_ : model.intercept\_ is used to obtain intercept value.
9. model.score(waist,weight) : This method is used to check the accuracy of the model.
10. model.predict(waist\_new) :This method is used to predict the value based on trained dataset.
11. data.corr() : This method is used to obtain correlation.
12. lm.fit(waist, weight) : fit() is used to train model.

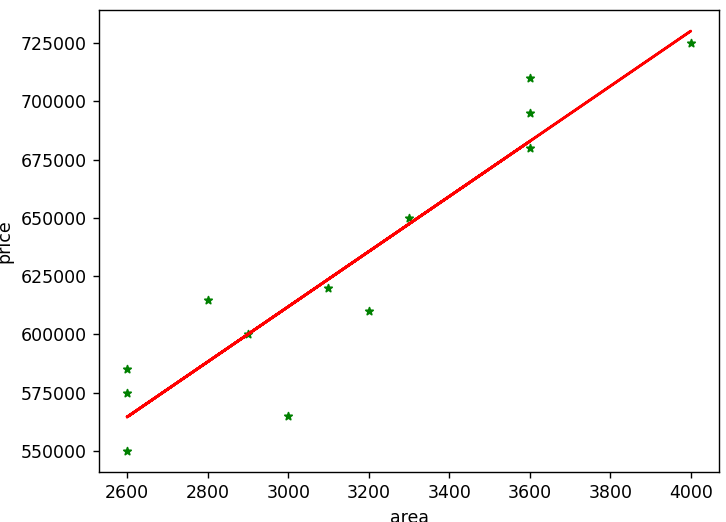
Code :



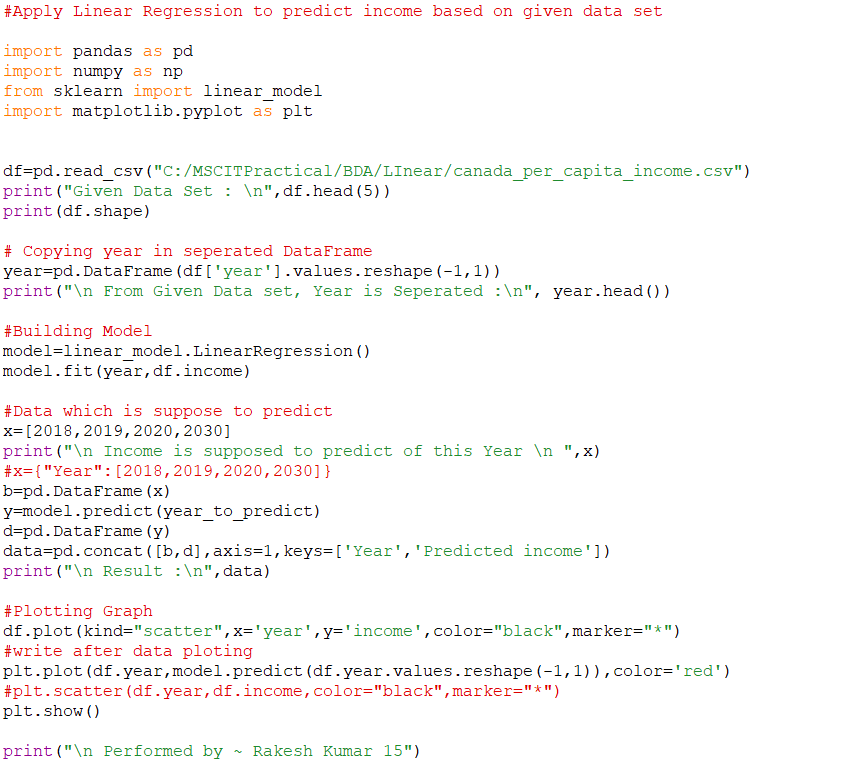
Output :



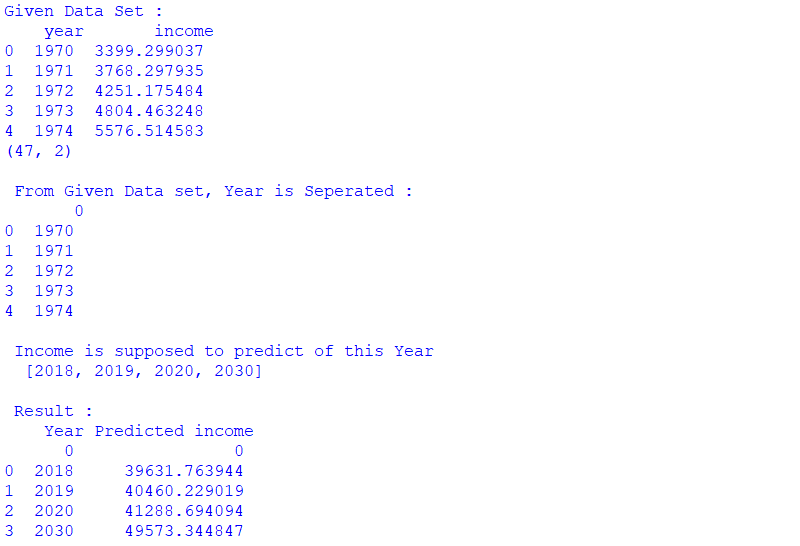


****

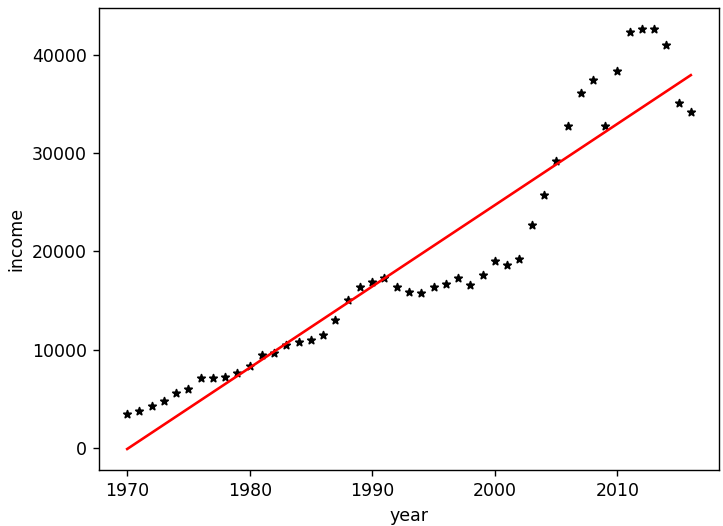
Code :



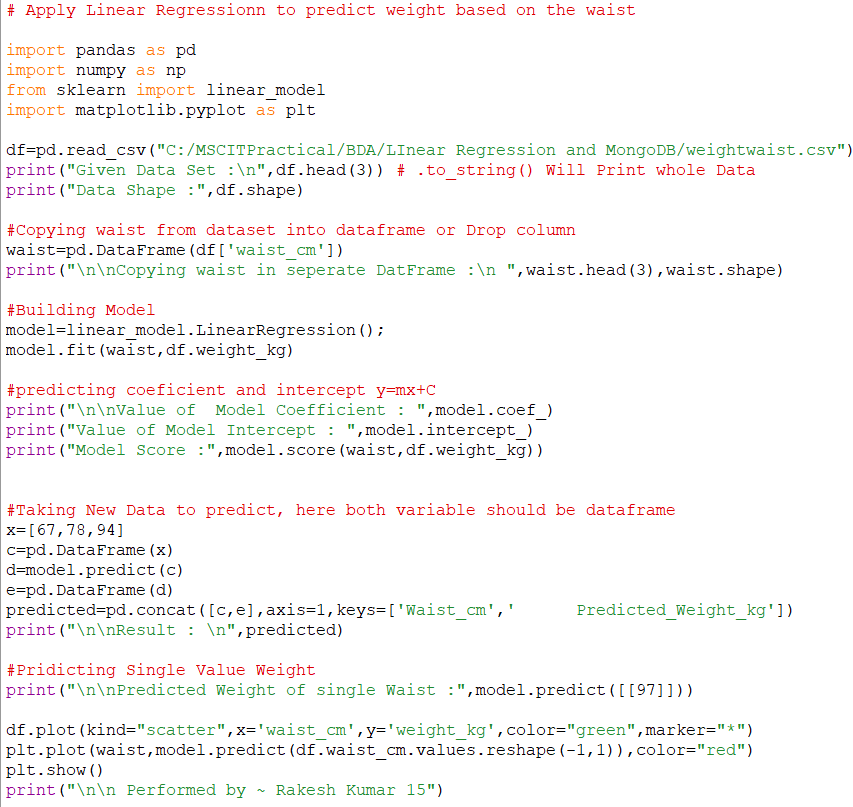
Output :



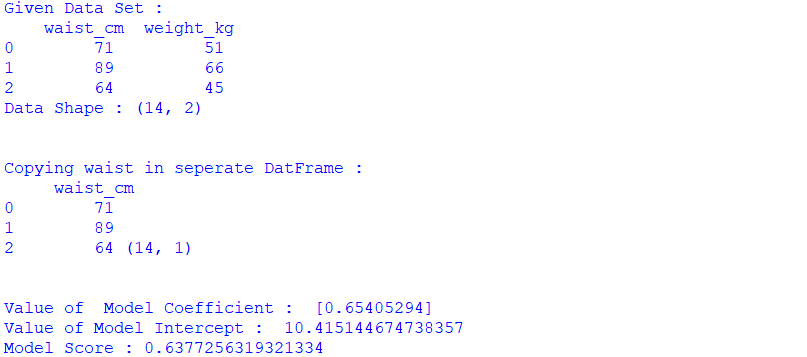


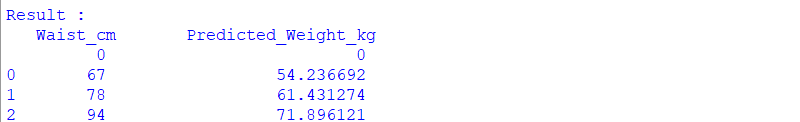


Code :



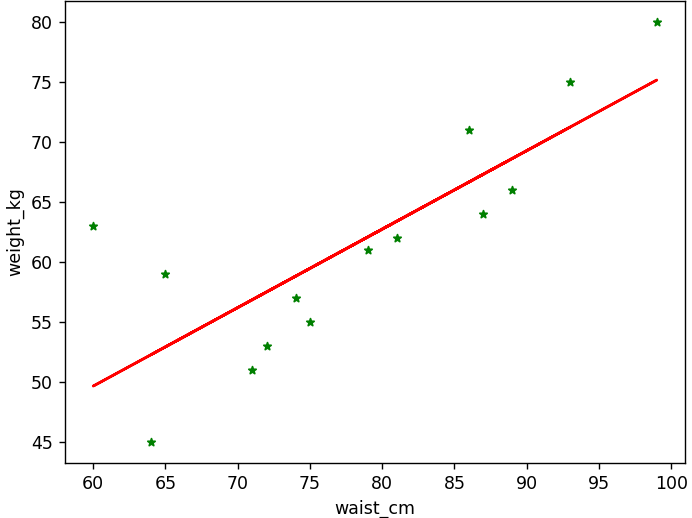
Output :











## PRACTICAL NO.3

Aim : Implementing Logistic Regression.

Description :-

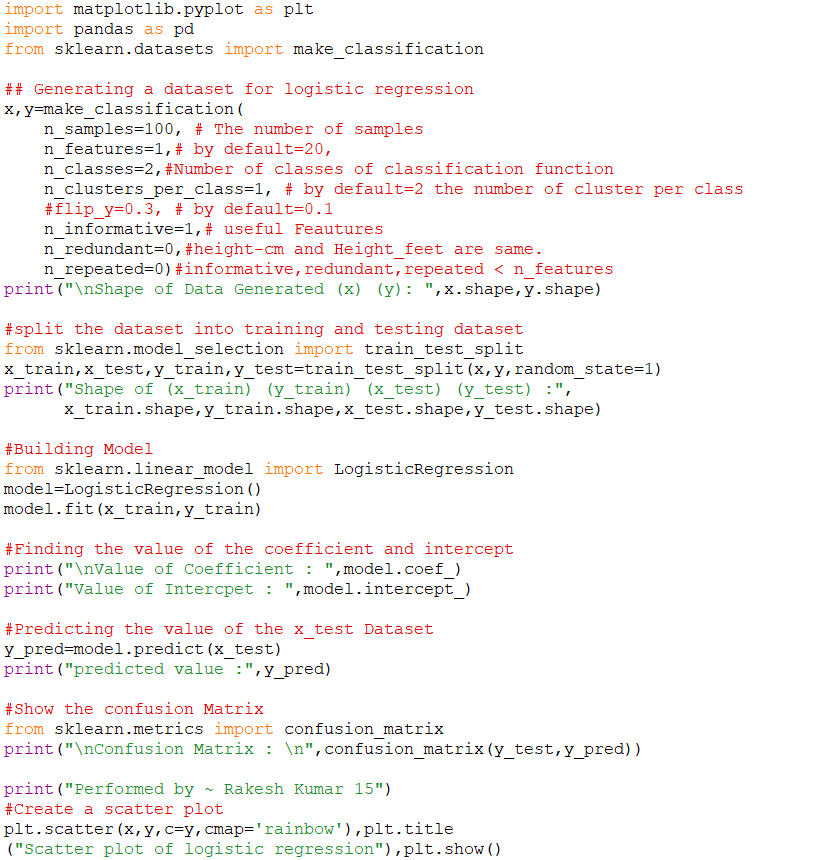
Logistic regression is one of the most popular Machine Learning algorithms, which comes under the Supervised Learning technique. It is used for predicting the categorical dependent variable using a given set of independent variables. Logistic regression predicts the output of a categorical dependent variable. Therefore the outcome must be a categorical or discrete value. It can be either Yes or No, 0 or 1, true or False, etc. but instead of giving the exact value as 0 and 1, it gives the probabilistic values which lie between 0 and 1. Logistic Regression is much similar to the Linear Regression except that how they are used. Linear Regression is used for solving Regression problems, whereas Logistic regression is used for solving the classification problems.

In Logistic regression, instead of fitting a regression line, we fit an "S" shaped logistic function, which predicts two maximum values (0 or 1). The curve from the logistic function indicates the likelihood of something such as whether the cells are cancerous or not, a mouse is obese or not based on its weight, etc. Logistic Regression is a significant machine learning algorithm because it has the ability to provide probabilities and classify new data using continuous and discrete datasets.

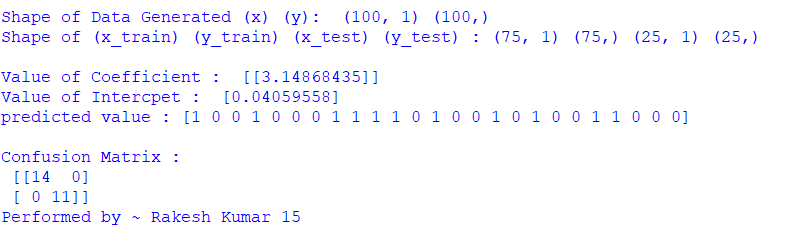
Method:

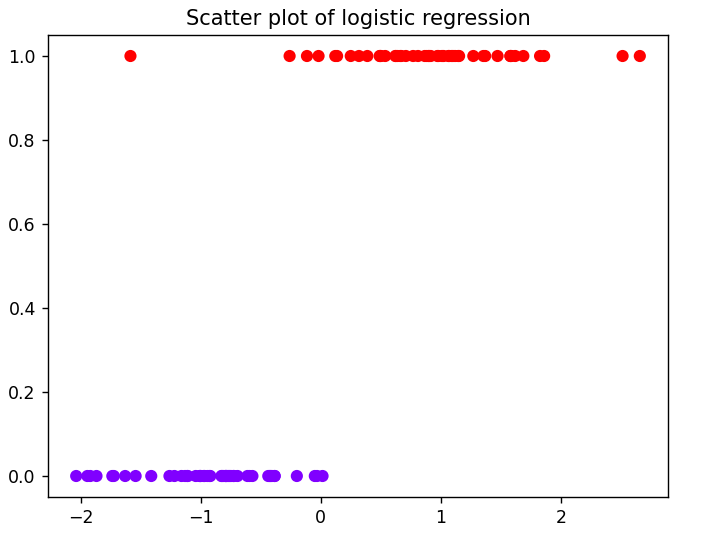
LogisticRegression() : This method is used to implement logistic regression.

Program Code :



Output:





**PRACTICAL NO.4**

Aim: Implement an application that stores big data in MongoDB and manipulate it using python.

Description :

MongoDB, the most popular NoSQL database, is an open-source document-oriented database. The term ‘NoSQL’ means ‘non-relational’. It means that MongoDB isn’t based on the table-like relational database structure but provides an altogether different mechanism for storage and retrieval of data.

SQL databases store data in tabular format. This data is stored in a predefined data model which is not very much flexible for today’s real-world highly growing applications. Modern applications are more networked, social and interactive than ever. Applications are storing more and more data are accessing it at higher rates.

Relational Database Management System (RDBMS) is not the correct choice when it comes to handling big data by the virtue of their design since they are not horizontally scalable. If the database runs on a single server, then it will reach a scaling limit. NOSQL databases are more scalable and provide superior performance. MongoDB is such a NoSQL database that scales by adding more and more servers and increases productivity with its flexible document model.

Methods :

1. MongoClient('localhost:27017') : This method is used to get at which port monodb is running.
2. client.get\_database('database\_name') : This method is used to access the database.
3. db.records\_name : This method is used to access the collection of database.
4. records.count\_documents({}) : This method is used to count the number of records in the collection.
5. list(records.find()) : This method is used to print all the records in collection.
6. records.update\_one({"$set":{"key","value"}}) : This method is used to update one record in collection.
7. records.insert\_one({"eno":6,"name":"Raj","location":"India"}) : This method is used to insert one record in collection.
8. records.delete\_one({"name":"Raj"}) : This method is used to delete one record from collection.

Steps : For mongoDB operation

Here We need Python -version 3.11.2 and MongoDB Version 5.0.15 2008R2Plus

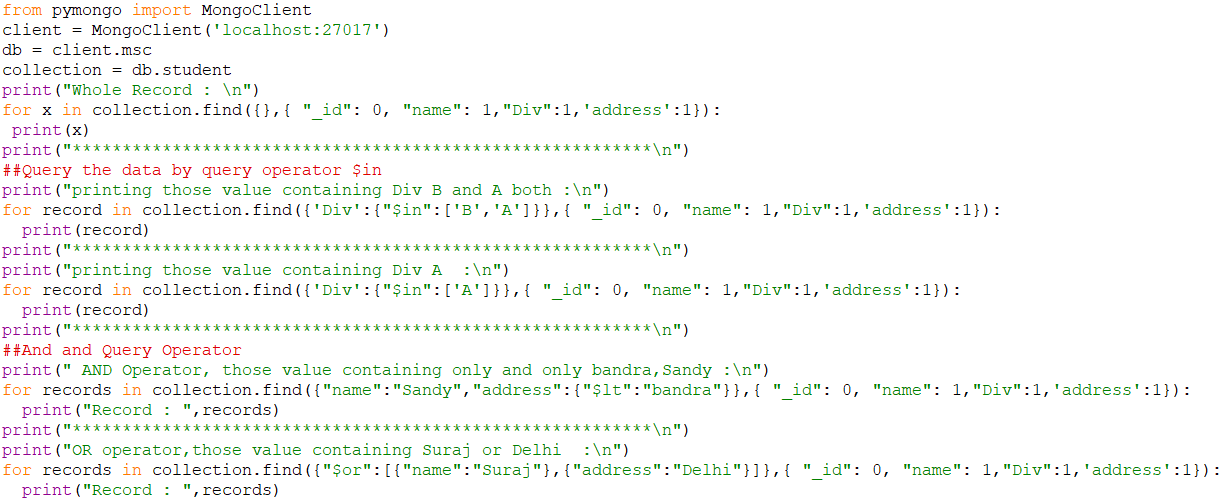




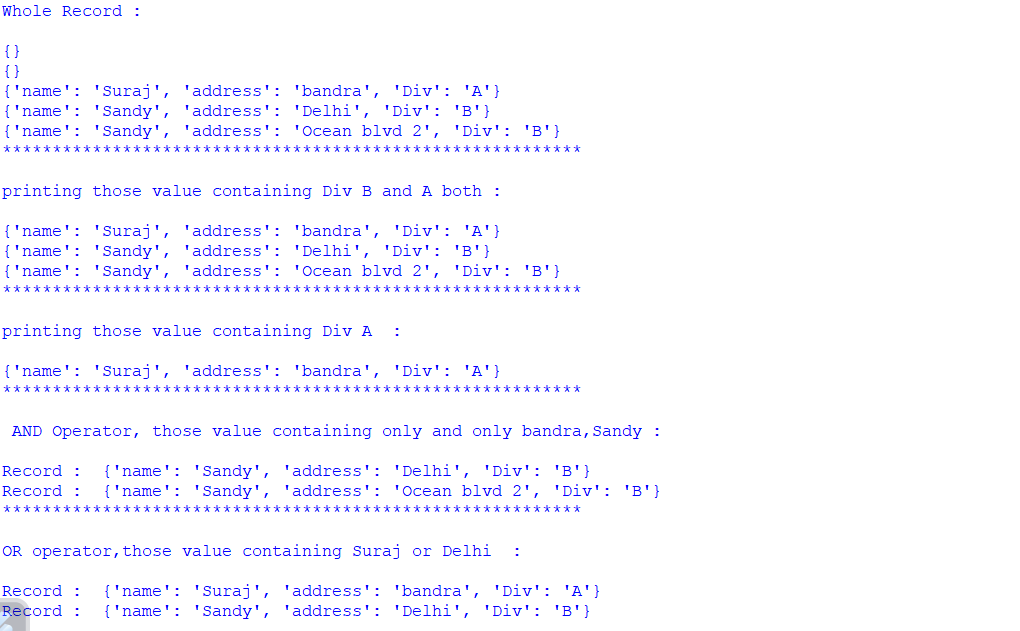




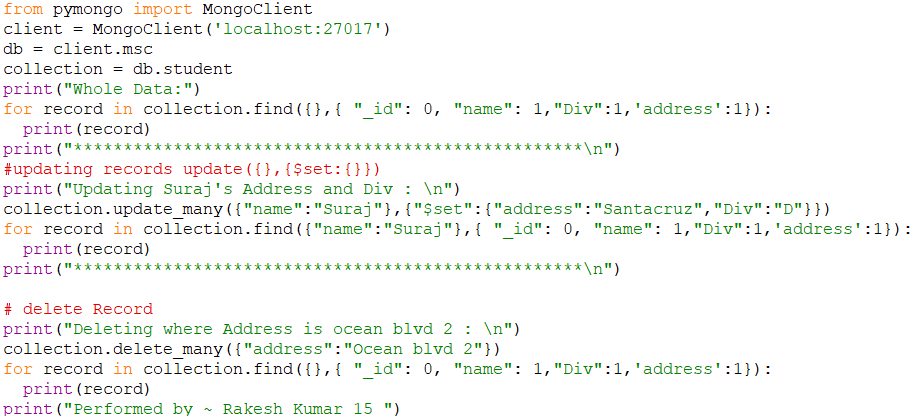
Code : For the Retrieve data



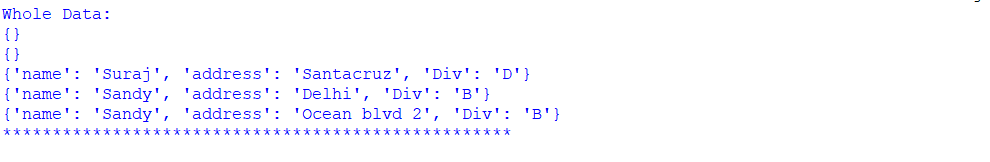
Output :

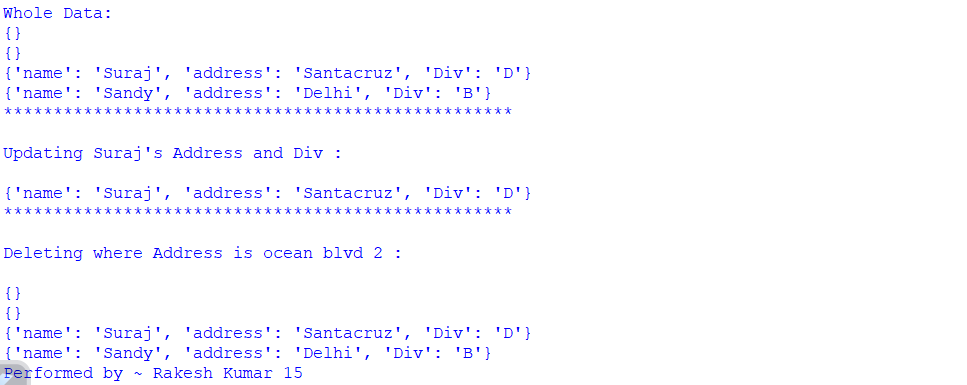


Code: for deleting and updating records



Output :

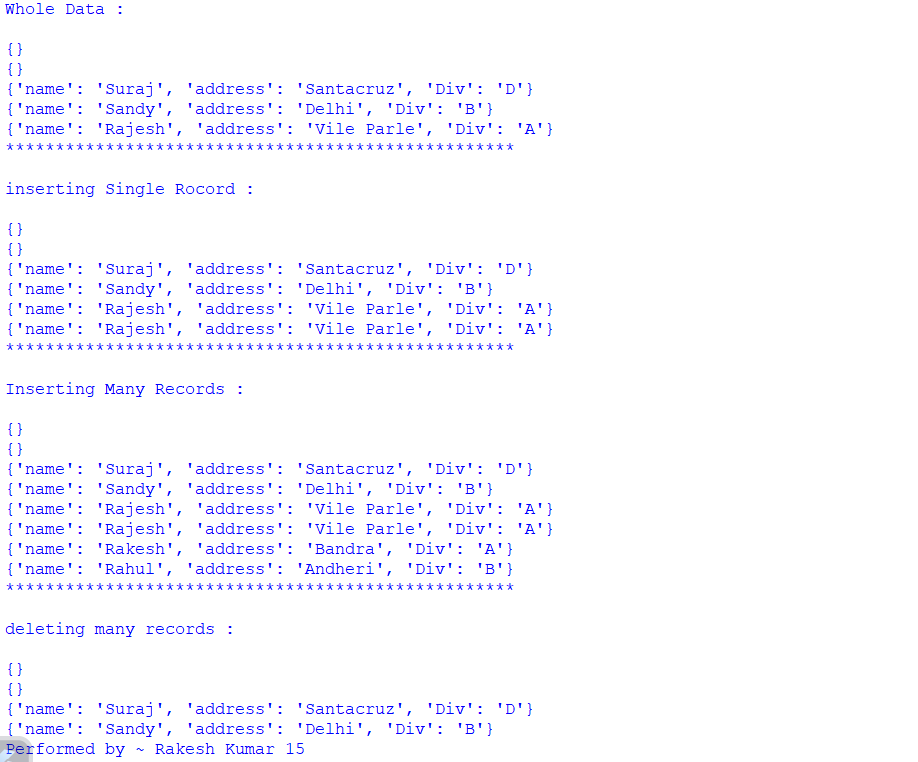




Code : For inserting single/Multiple Records also Deleting Multiple Records



Output :



## PRACTICAL NO.5

Aim: Implement SVM classification Technique.

Description :- SVM is a famous supervised machine learning algorithm used for classification as well as regression algorithms. However, mostly it is preferred for classification algorithms. It basically separates different target classes in a hyperplane in n-dimensional or multidimensional space. The main motive of the SVM is to create the best decision boundary that can separate two or more classes(with maximum margin) so that we can correctly put new data points in the correct class. Because It chooses extreme vectors or support vectors to create the hyperplane, that’s why it is named so.

Methods :-

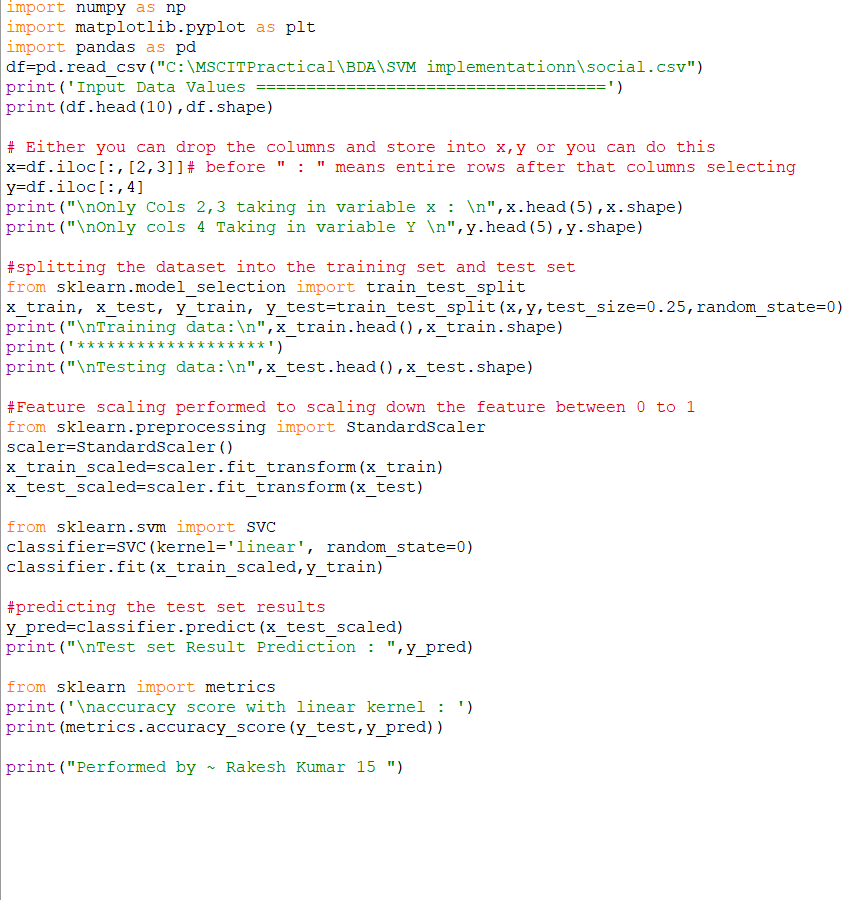
1. StandardScaler() : It is used for feature scaling. Use to transform the not scale data to scale data. X is in less range y is in high range then might be x get less importance by model itself so avoid this, where It calculates the mean and standard deviation for the independent and dependent variable. Each datapoint – mean of columns and divide by the standard deviation. By doing so mean of the respected columns become 0 and deviation become 1. When you plot your data you will see that all your data points centered towards the mean value. So all the features start getting the equal importance by the model
2. SVC(kernel='linear', random\_state=0) : This method is used for implementing SVM.
3. metrics.accuracy\_score(y\_test,y\_pred) : This method is used to check the accuracy score of the model.
4. df.iloc() : The iloc function in Python returns a view of the selected rows and columns from a Pandas DataFrame.

Fit.transform() : where fit with parameter calculates the mean and standard deviation and transform() does subtract form each data point – mean divide by deviation

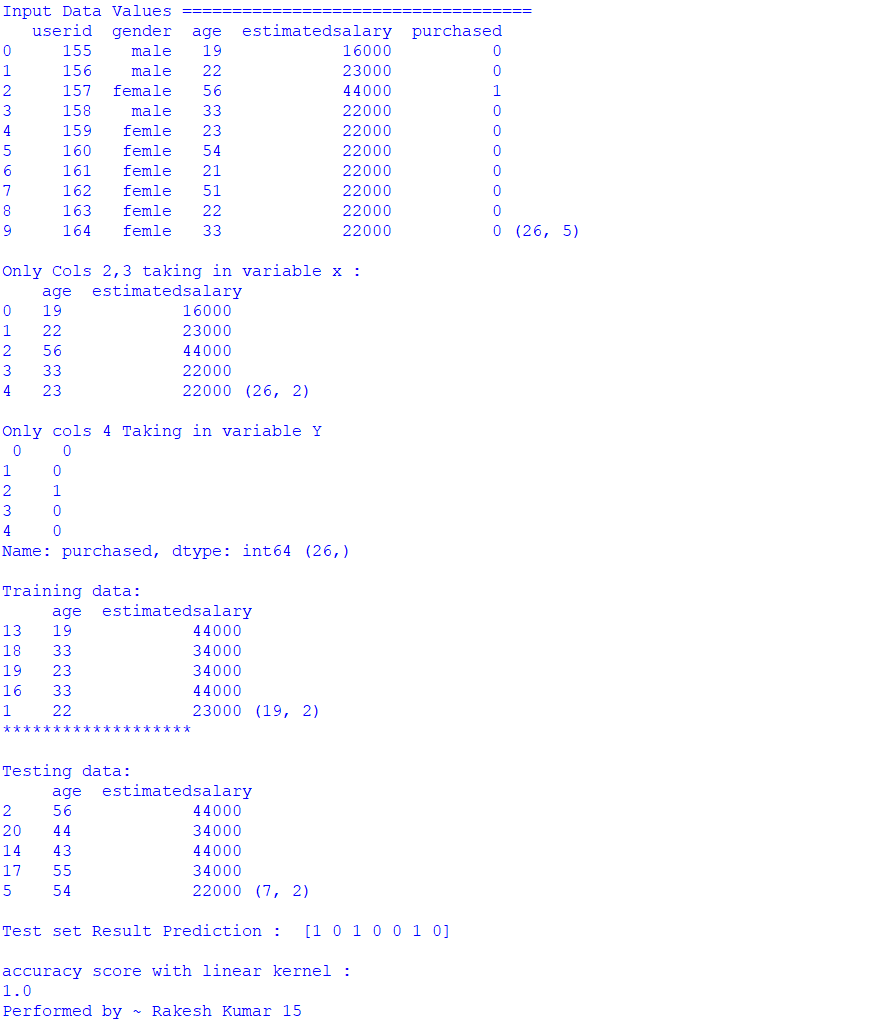
1. : Dependent variable is not in the form of binary then ---df[“ purchased ”] = df[“purchased”] . map ({“any name”:0 , “anyanme” : 1 })
2. If data points in purchased are more than 2 types then

df=df [df [“purchased”] ! = ” 3rd type ”]

Program code:



Output:



## PRACTICAL NO.6

## 

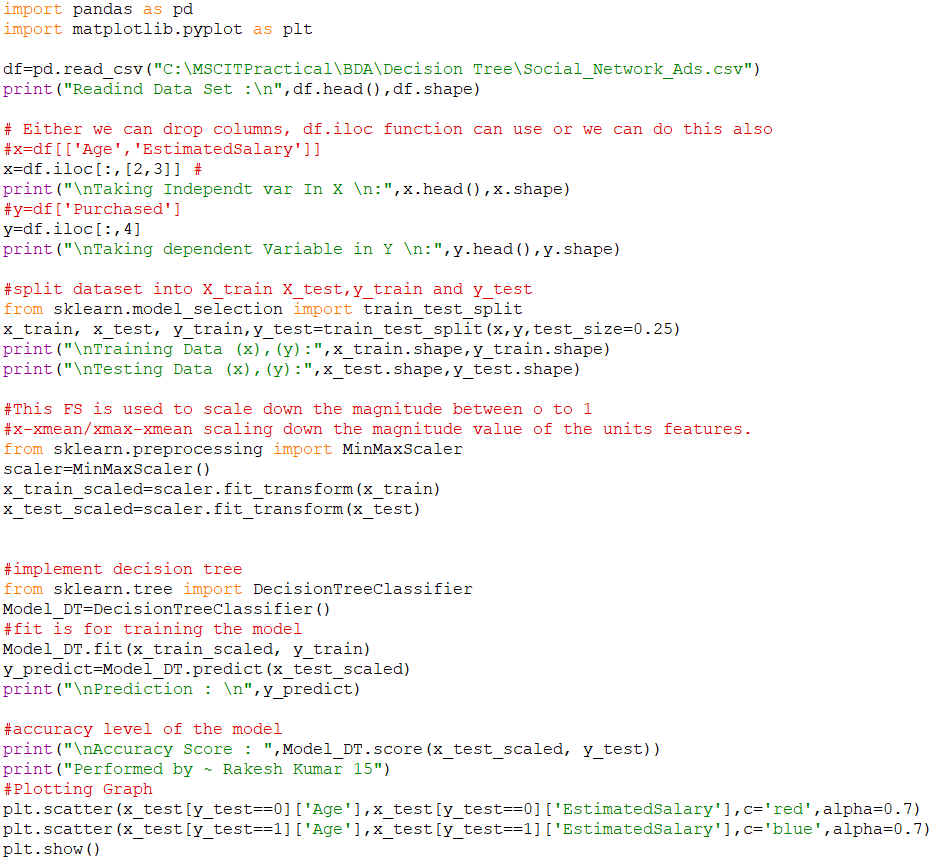
Aim: Implement Decision Tree classification Technique.

Description :- Decision Tree is a supervised learning method used in data mining for classification and regression methods. It is a tree that helps us in decision-making purposes. The decision tree creates classification or regression models as a tree structure. It separates a data set into smaller subsets, and at the same time, the decision tree is steadily developed. The final tree is a tree with the decision nodes and leaf nodes. A decision node has at least two branches. The leaf nodes show a classification or decision. We can't accomplish more split on leaf nodesThe uppermost decision node in a tree that relates to the best predictor called the root node. Decision trees can deal with both categorical and numerical data.

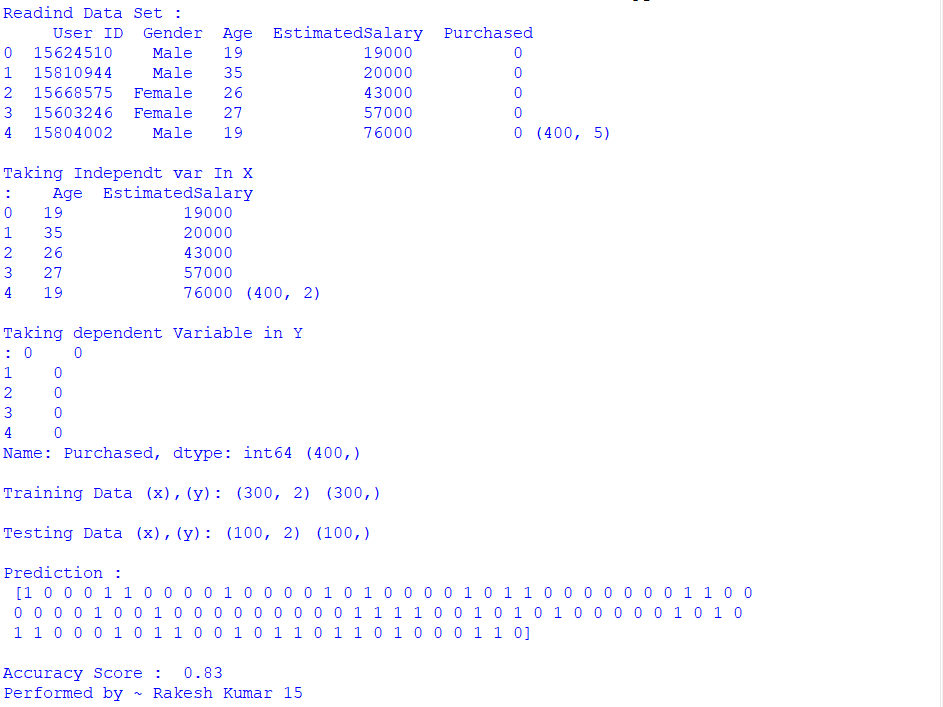
Methods :

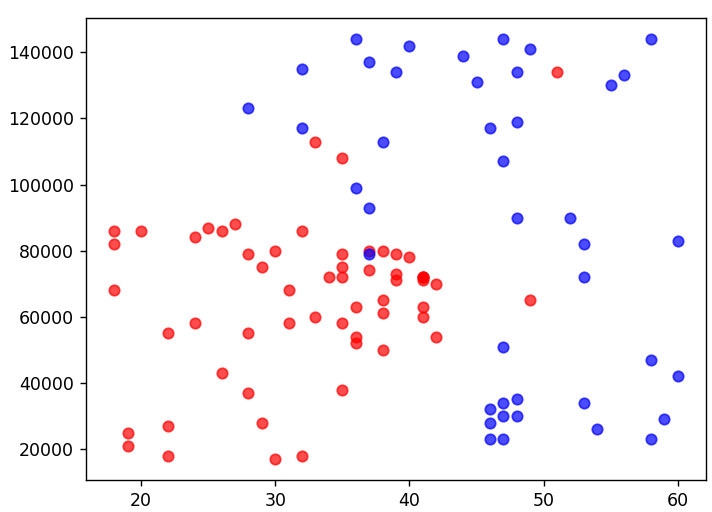
1. MinMaxScaler() : This method is used for feature scaling.
2. DecisionTreeClassifier() : This method is used to implement decision tree

Program code:



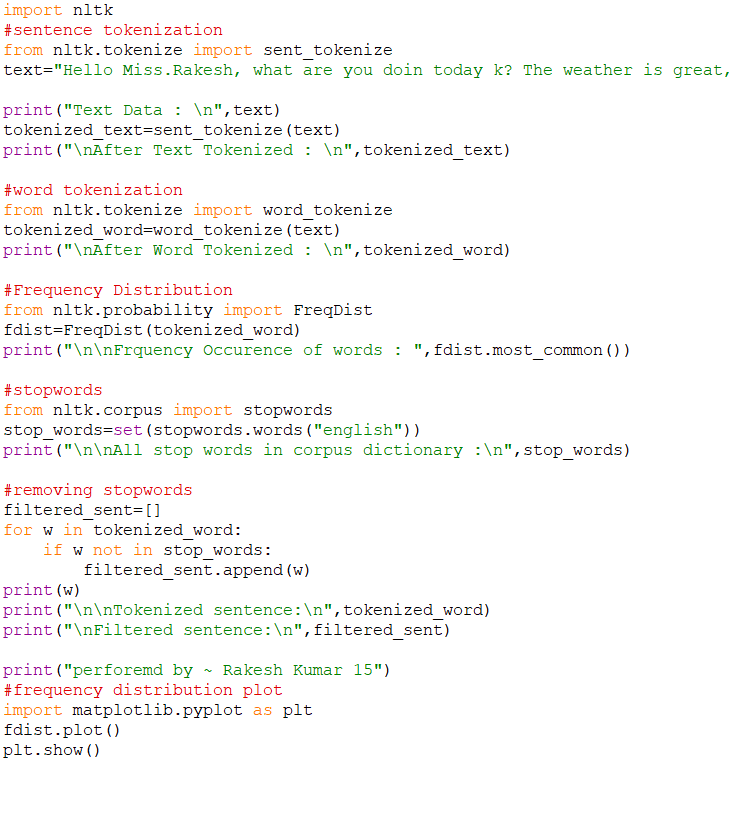
Output:



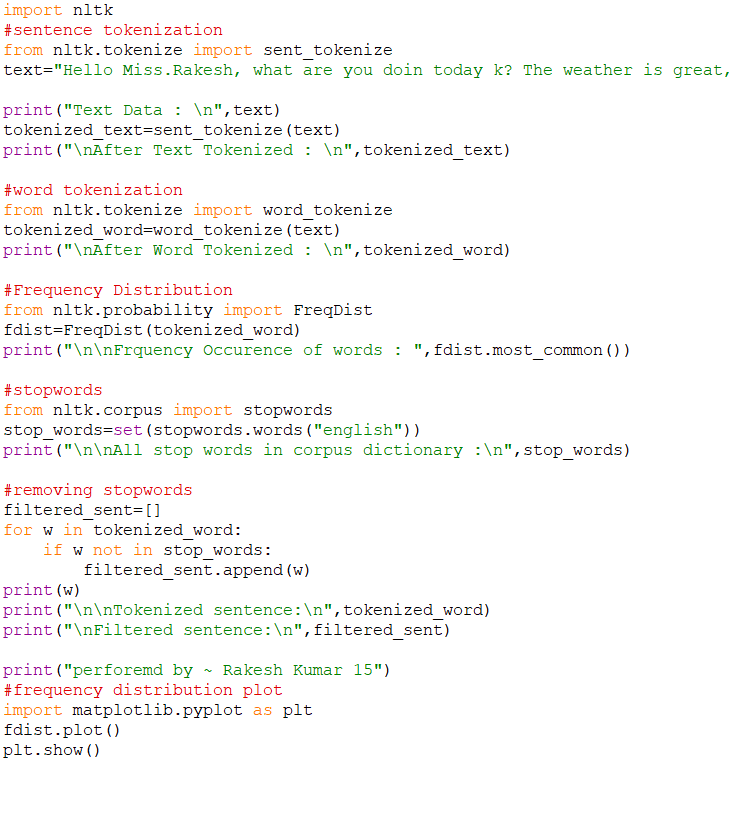


Practical No : 7

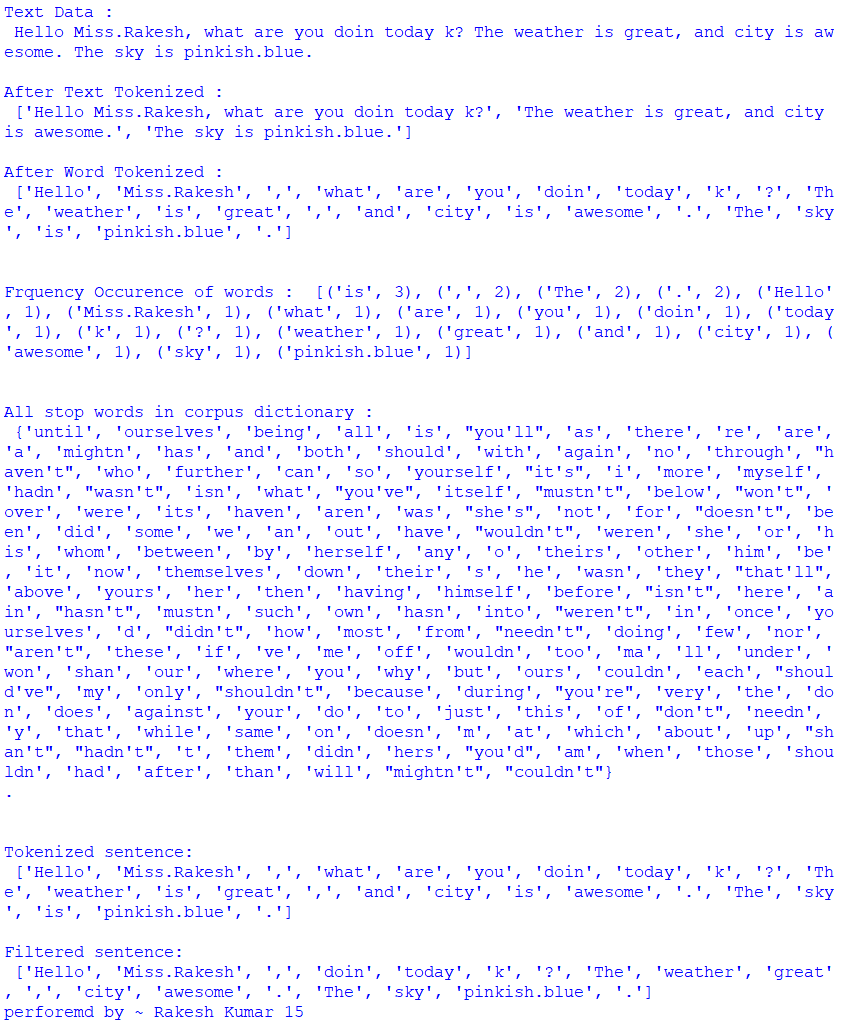
Program Code :

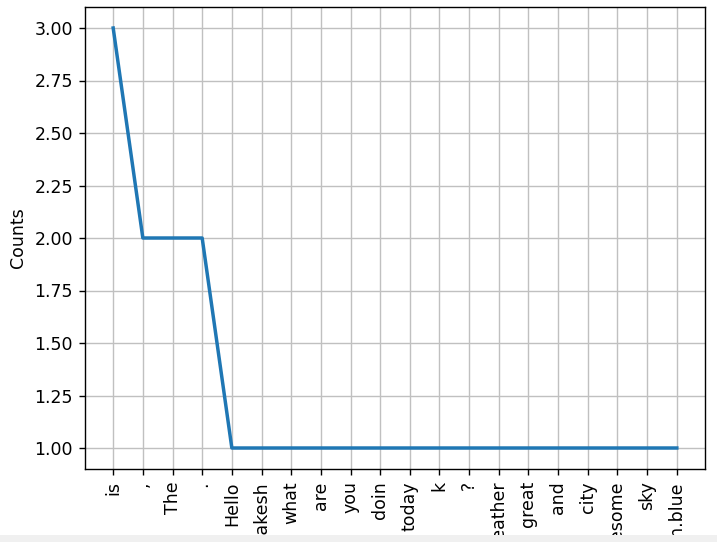






Output :





**Practical No. 8**

Sentimental Analysis : Sentiment analysis is a [natural language processing](https://en.wikipedia.org/wiki/Natural_language_processing) technique that identifies the polarity of a given text. It refers to analyse an opinion or feelings about something using data like text or images, regarding almost anything. Sentiment analysis helps companies in their decision making process. For example, if public sentiments towards a product is not so good, a company may try to modify the product or stop the production altogether in order to avoid the losses. There are different flavors of sentiment analysis, but one of the most widely used techniques labels data into positive, negative and neutral.

unstack() is used to reshape the given Pandas DataFrame by transposing specified row level to column level. By default, it transposes the innermost row level into a column level. This is one of the technique for reshaping the DataFrame. When we want to analyze or reshape the data, Pandas provides in-built functions.

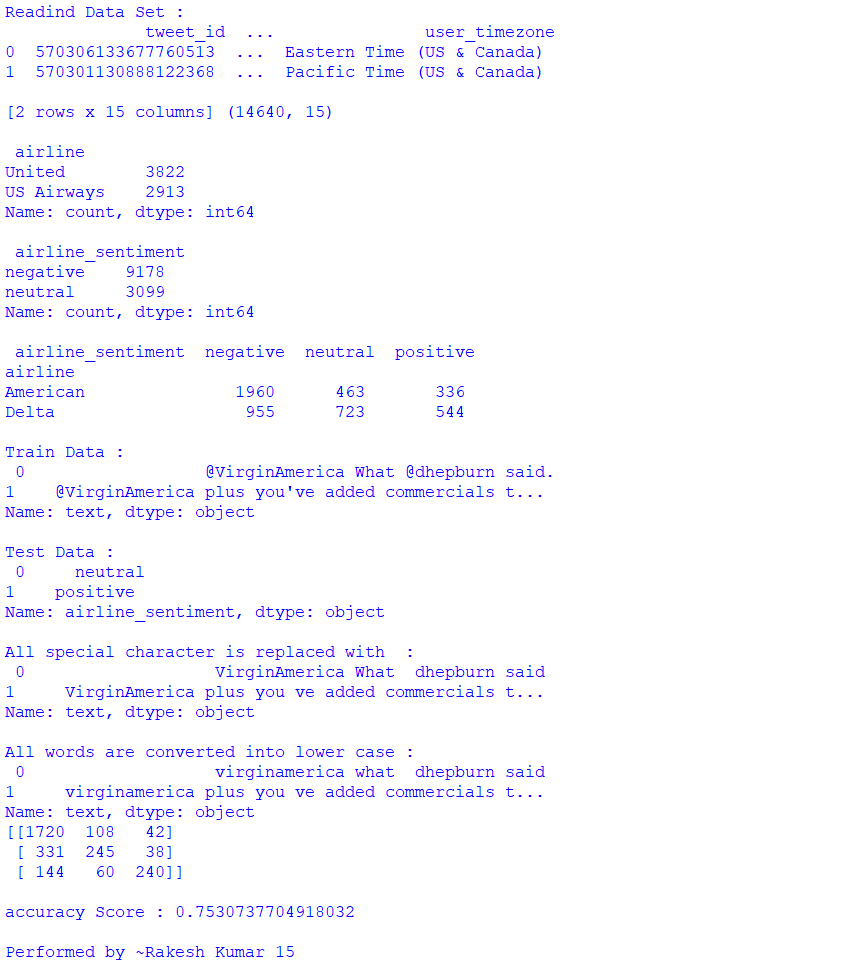
The count() method counts the number of not empty values for each row, or column if you specify the axis parameter as axis='columns' , and returns a Series object with the result for each row (or column).

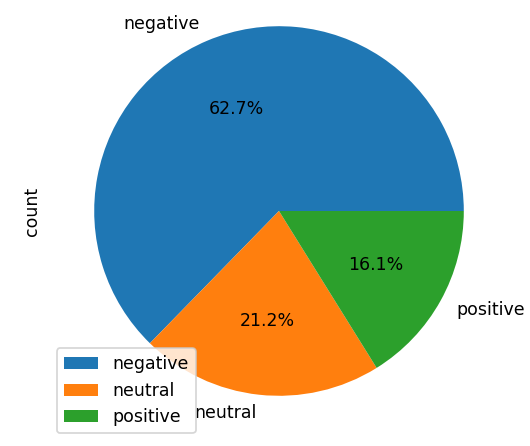
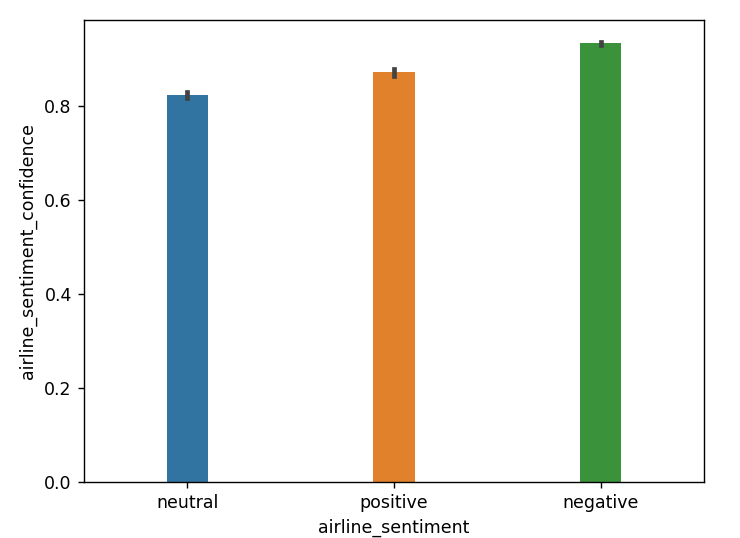
Groupby(): name of the column on the basis of you can create the group like school , sector means creating a group inside your data

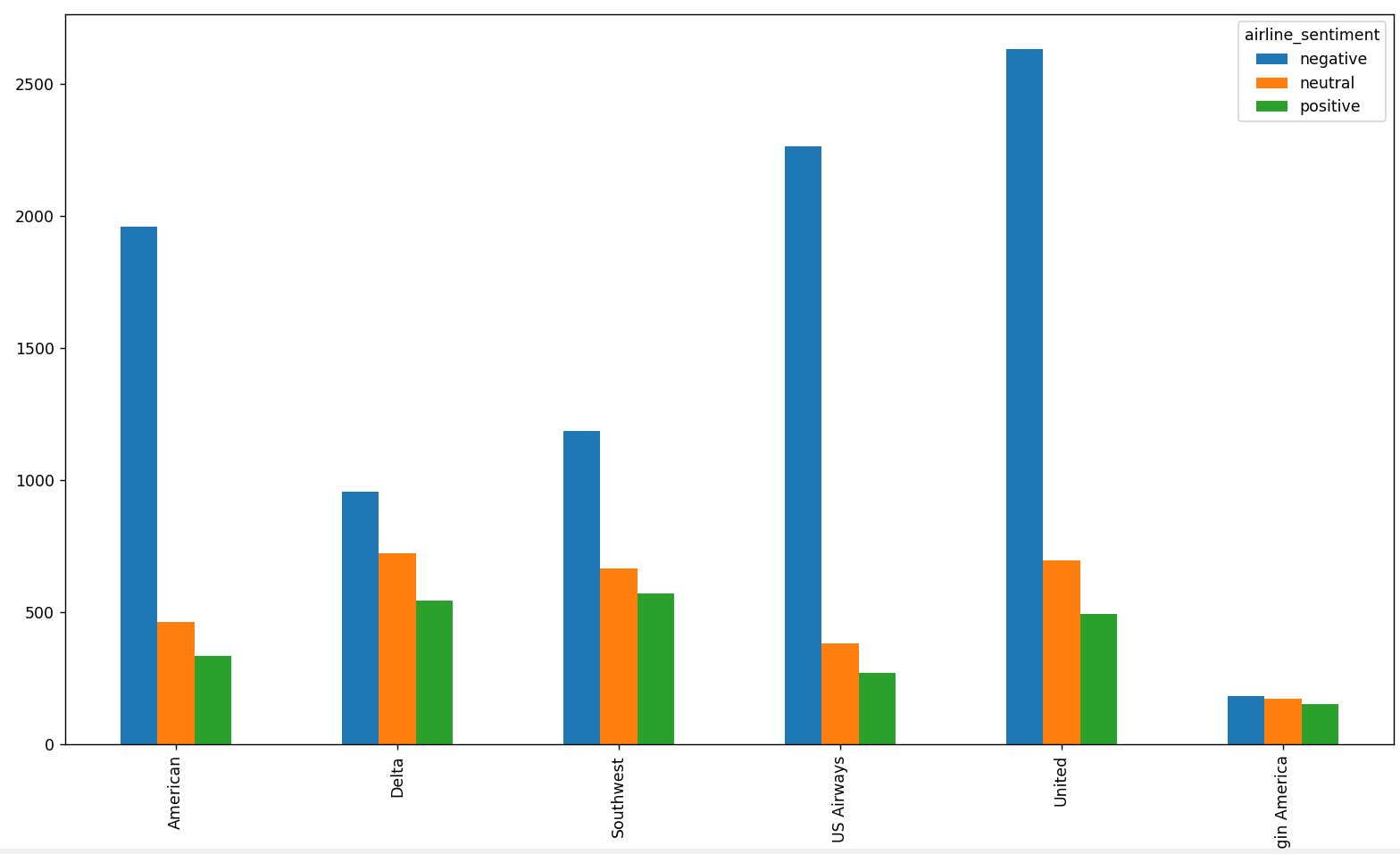
Program Code :



Output :



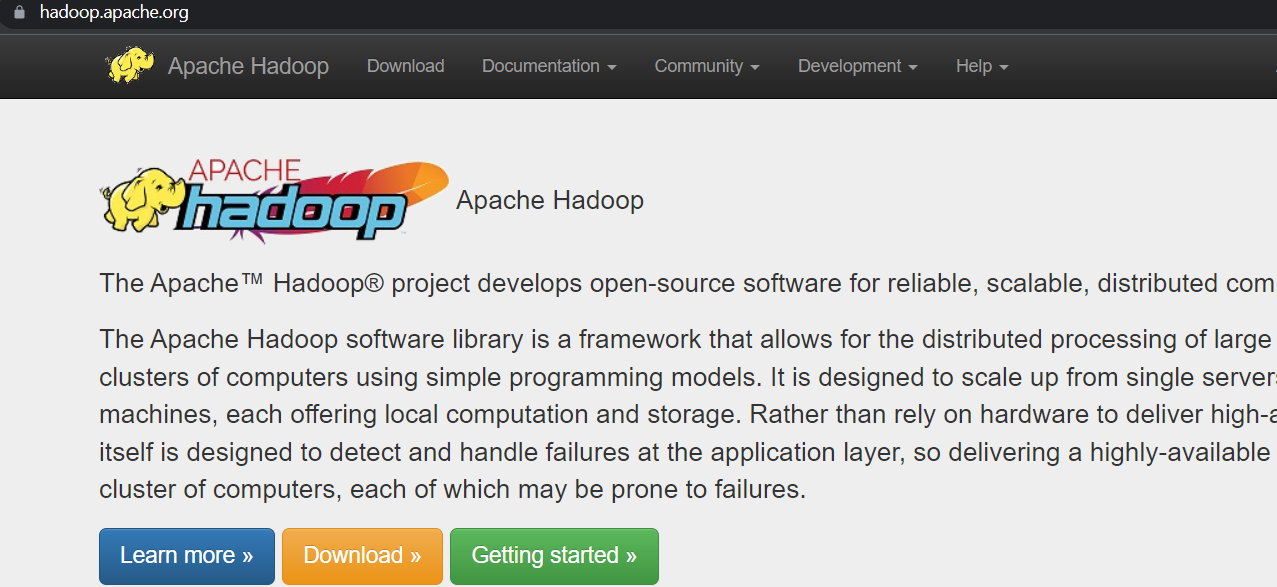
 

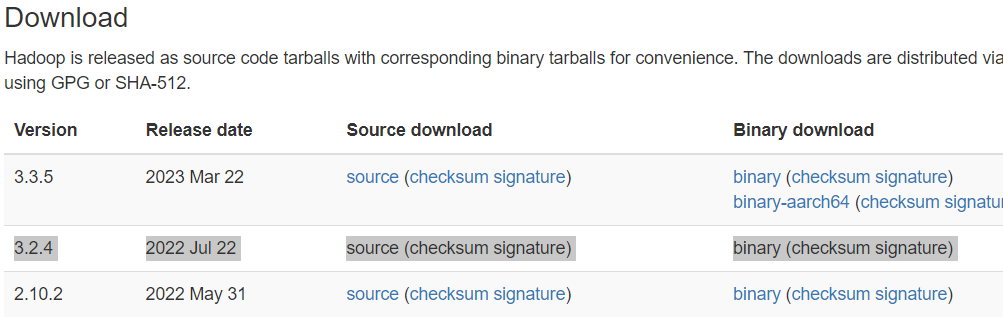


**Practical No. 9**

**Steps : For Downloading Hadoop.**

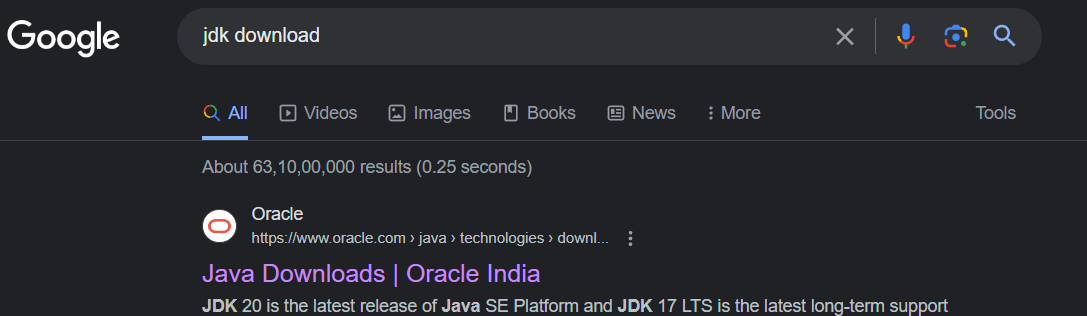
****

****

****

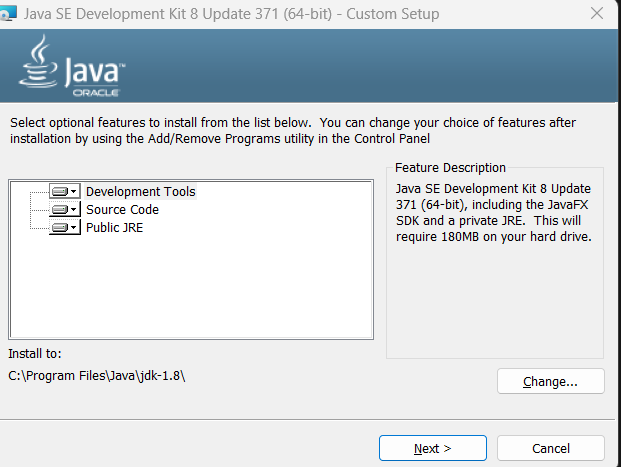
Extract with WinRAR.

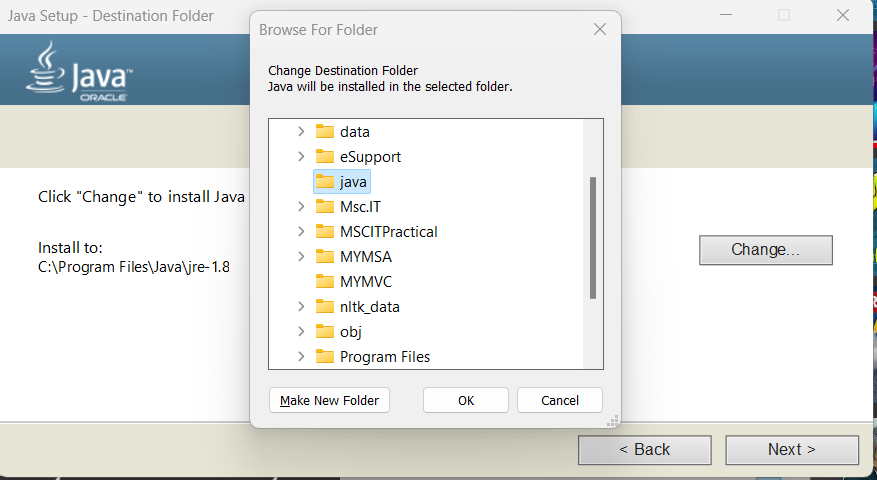
**Step 2:for downloading JDK.**

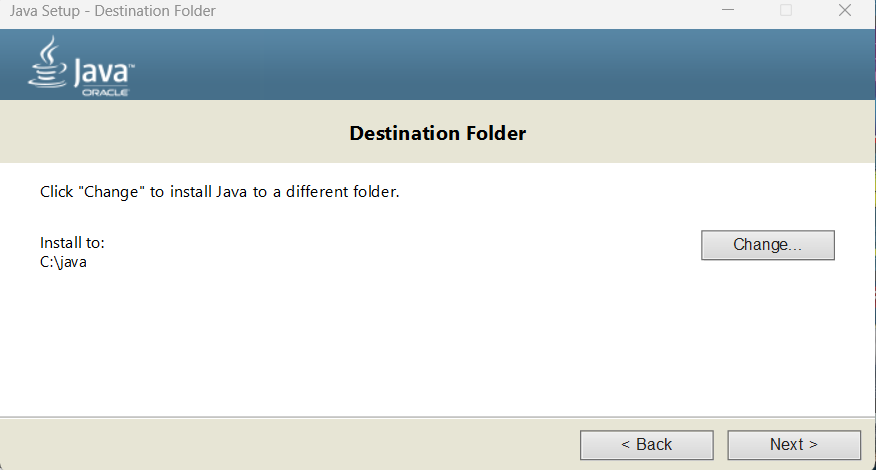
****

****

Lets Start installing JDK.







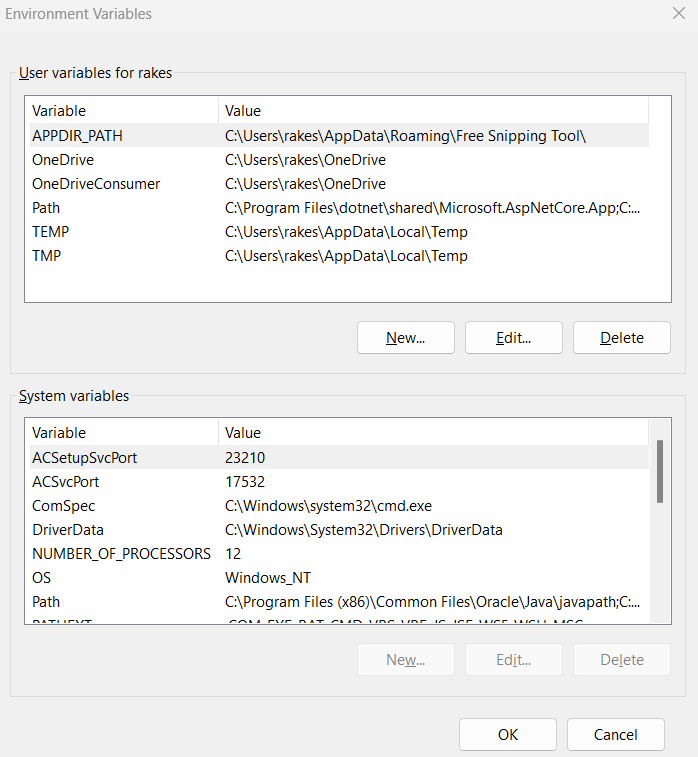
**Step 3:** We have to move jdk file from C:\Program Files\Java to newly created java folder C:\java.

Delete “java folder from C:\Program Files\Java” in C drive.

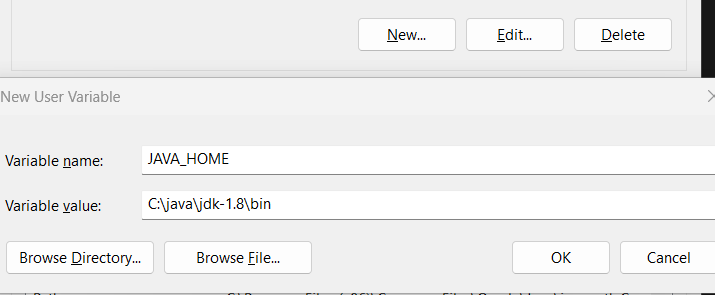
C:\java because sometimes It throws error while setting environment variable.

**Step 4:** Now we have to set environment variable and path for JDK

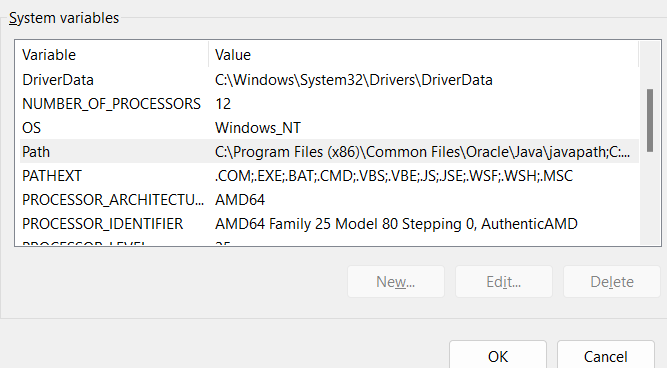
Click window and search for



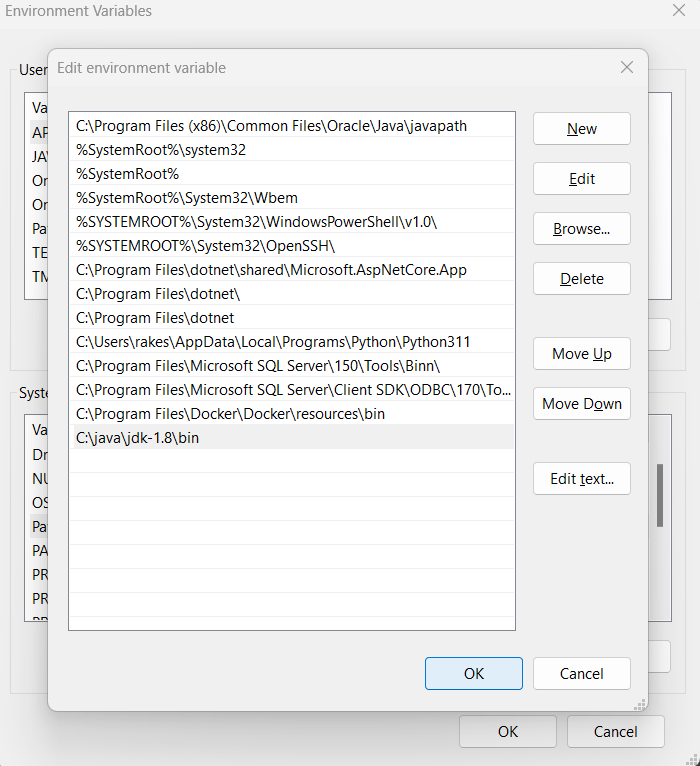
Click on New and Give Variable name and path.



Click on path

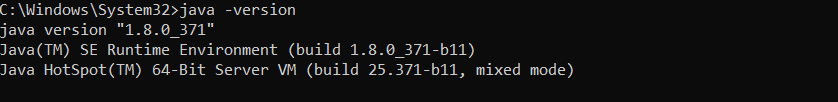


Click on edit and then click on new and paste the jdk/bin path

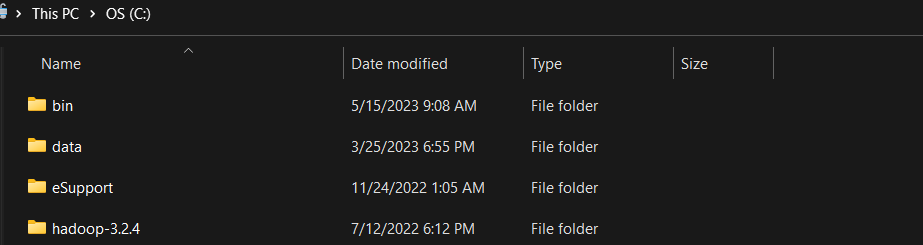


Now you can check installation of jdk in cmd.

Open CMD run as administrator.



**Step 5:** Now extract the Hadoop tar file with WinRar and move the file in C drive.



Now we need to perform some configuration in ETC folder files.

To edit this, click the desire file and right click open all simultaneously.

1. core-site

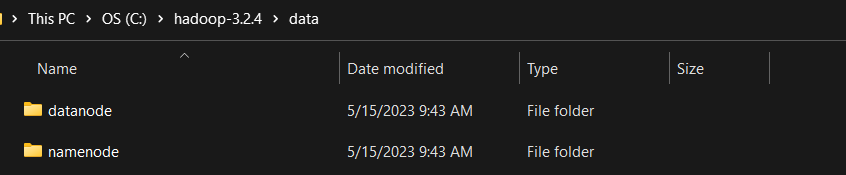
2. hdfs-site

3. mapred-site

4. yarn-site

5. hadoop-env here set JAVA\_HOME=C:\java\jdk-1.8

**Lets first add the folder data and two more folder datanode and namenode in hadoop/data folder.**

****

a) File C:/Hadoop-3.2.1/etc/hadoop/core-site.xml, paste below xml paragraph and

save this file.

<configuration>

<property>

<name>fs.defaultFS</name>

<value>hdfs://localhost:9000</value>

</property>

</configuration>

b) C:/Hadoop-3.2.1/etc/hadoop/mapred-site.xml, paste below xml paragraph and save

this file.

<configuration>

<property>

<name>mapreduce.framework.name</name>

<value>yarn</value>

</property>

</configuration>

**c) Create folder "data" under "C:\Hadoop-3.2.1"**

**1) Create folder "datanode" under "C:\Hadoop-3.2.1\data"**

**2) Create folder "namenode" under "C:\Hadoop-3.2.1\data" data**

d) Edit file C:\Hadoop-3.2.1/etc/hadoop/hdfs-site.xml, paste below xml paragraph

and save this file.

<configuration>

<property>

<name>dfs.replication</name>

<value>1</value>

</property>

<property>

<name>dfs.namenode.name.dir</name>

<value>**C:\hadoop-3.2.1\data\namenode**</value>

</property>

<property>

<name>dfs.datanode.data.dir</name>

<value>**C:\hadoop-3.2.1\data\datanode**</value>

</property>

</configuration>

e) Edit file C:/Hadoop-3.2.1/etc/hadoop/yarn-site.xml, paste below xml paragraph

and save this file.

<configuration>

<property>

<name>yarn.nodemanager.aux-services</name>

<value>mapreduce\_shuffle</value>

</property>

<property>

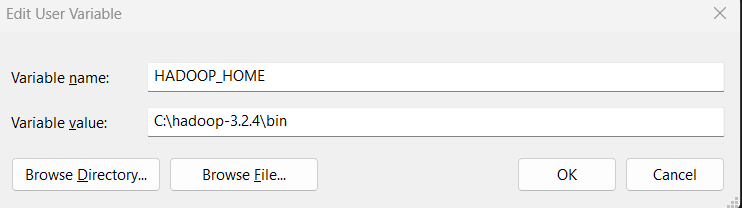
<name>yarn.nodemanager.auxservices.mapreduce.shuffle.class</name>

<value>org.apache.hadoop.mapred.ShuffleHandler</value>

</property>

</configuration>

**Step 5:** Now we have to set the path for hadoop in environment variable.



Path



**Step 7:**

If you see hadoop/bin folder, you’ll see some important files are missing.

Download the bin configuration file and extract it.

Now delete the bin folder of hadoop and paste new bin folder.