SAVEETHA SCHOOL OF ENGINEERING
SAVEETHA INSTITUTE OF MEDICAL AND TECHNICAL SCIENCES
COMPUTER SCIENCE AND ENGINEERING

LAB EXPERIMENTS INDEX

SUBJECT CODE : CSA1611

SUBJECT NAME : DATA WAREHOUSING AND DATA MINING USING PREDICTIVE ANALYSIS

1. CENTRAL TENDENCY AND DATA DISPERSION MEASURES USING R-TOOL

Download the dataset from the UCI repository (or) any other appropriate website
Implement the following operations using R.
a)Dispersion measures (mean,median,mode,midrange), .
b)Data Dispersion techniques-Five number Summary .
c)List out the Range and outliers values.

2. PLOTTING GRAPHS USING R-TOOL

Download the dataset from the UCI repository (or) any other appropriate website
Draw Boxplot, Histogram, Scatterplot for the dataset.

3. PERFORM CORRELATION ANALYSIS AND NORMALIZATION USING R-TOOL

Download the dataset from the UCI repository (or) any other appropriate website
Implement the following operations using R.
a)Perform the correlation analysis for the numerical attribute using Pearson coefficient
b)Perform Correlation Analysis for categorical attribute using chi-square Test
c) Apply Normalization technique using min_ max, z score and decimal scaling for the given data frames of particular dataset.
.

4. REGRESSION ANALYSIS USING R TOOL

Download the dataset from the UCI repository (or) any other appropriate website
Implement the following operations using R.
a)Linear regression .
b)Multipl regression .

5. DATA PREPROCESSING AND ANALYSIS FOR DATASET USING WEKA

Download the dataset from the UCI Repository .
Using Weka,for each attribute find the following information,
a. Attribute Type
b. Percentage of missing values
c. Find min, max, mean, standard deviation on the given dataset
d. Are there any records that have a value but no other record has
e. Write a note on class distribution for each of the attributes

6. DATA SEGMENTATION BY K- MEANS CLUSTER AND EXPECTATION MAXIMISATION ALGORITHM USING WEKA .

a)Apply K-means algorithm to your dataset downloaded from UCI repository/use .arff file from Weka . Perform the following operations by setting the number of clusters and seed value of the algorithm for the generation of initial cluster centres.

A. Choose a set of attributes for clustering give a motivation.
B. Experiment with atleast 2 different numbers of clusters but with a same seed value.
(by default: Seed=10).Analysis the resulting clusters obtained and give a report.
C. Try with different seed value .Explain what the seed value controls.
D. Do you think the clusters are good clusters?. What does each cluster represent?.

b)Analyse the dataset using expectation maximisation algorithm by setting the minimum standard deviation for normal density calculation and compare the results with simple k-means algorithm using Weka.

7 DATA SEGMENTATION BY COBWEB – HIERARCHIAL CLUSTERING ALGORITHM USING WEKA TOOL.

Apply Hierarchial Clustering algorithma and Cobweb Hierarchial Clustering Algorithm to your dataset downloaded from UCI repository/use .arff file from Weka .Analyse the results obtained and give Comparison report.

8. FREQUENT PATTERN MINING USING ASSOCIATION RULE THROUGH WEKA

Run the Apriori algorithm and explore the association rules by changing the following parameters. Use supermarket.arff dataset to perform the following opetations.

1: Upper bound min support
2: Lower bound min support
3: Metric type
4: Output item sets

9. FREQUENT PATTERN MINING USING FP GROWTH THROUGH WEKA TOOL

Run the FP Growth algorithm and explore the association rules by changing the following parameters: Use supermarket.arff dataset to perform the following operations.

1: Upper bound min support
2: Lower bound min support
3: Metric type
4: Output item sets

10. PREDICTION OF CATEGORICAL DATA USING DECISION TREE ALGORTIHM THROUGH WEKA

Prediction of categorical data using decision tree algorithm through weka.Use credit.arff dataset for performing the following operations.

a)Build a model using Decision Tree .Train a Decision Tree using the complete dataset as the training data. Report the model obtained after training.

b) Suppose you use your above model trained on the complete dataset, and classify credit good/bad for each of the examples in the dataset. What percentage of examples can you classify correctly ? Why do you think you cannot get 100 % training accuracy ?

c) Is testing on the training set as you did above a good idea ? Why orWhy not ?

11. PREDICTION OF CATEGORICAL DATA USING NAÏVE BAYES ALGORTIHM THROUGH WEKA

Perform the Naïve Bayes classification algorithm using weka. Compare the results obtained with decision tree,

12. PREDICTION OF CATEGORICAL DATA USING SMO ALGORTIHM THROUGH WEKA

Perform the SMO algorithm using weka. Compare the results obtained with decision tree.

13. EVALUATING ACCURACY OF THE CLASSIFIERS

Compute the confusion matrix in weka for the Japanese Credit Screening Data Set (download using Weka) using the following algorithm.

1. Logistic Regression
2. Naive Bayes
3. Decision Tree
4. k-Nearest Neighbors
5. Support Vector Machines

Compare the results obtained

14. NUMERICAL PREDICTION ANALYSIS USING LINEAR REGRESSION THROUGH WEKA

Using regression analysis create a model to calculate the price of house (Dataset provided).

a) Create .arff file

b) Create a model based on other comparable houses in the neighborhood and

c) Predict the Price of a House with a model created by using Linear Regression ..

15. Use the dataset downloaded from the UCI repository and perform the following using KNIME

a)Extract the data through csv file loader.

b)Use row filter, rule-based filter, pivot and group by to transform the data into an understandable knowledge.

c)Report the result (visualization)

d)Apply simple K-means Algorithm and visualize the clusters .

EX.NO: 01
Date:

CENTRAL TENDENCY AND DATA DISPERSION MEASURES USING R-TOOL

PROBLEM STATEMENT:

Download the dataset from the UCI repository (or) any other appropriate website and perform (or) implement the central tendency measures.(mean, median, mode and midrange) and Data dispersion technique including summary.

DESCRIPTION:

This data comes from the 2010 census profile of general population and housing characteristics. Zip codes and limited to those that fall at least partially within LA city boundaries. The dataset will be updated after the next census in 2020.

CENTRAL TENDENCY:

i. **Mean** : The mean is the average of the numbers: a calculated "central" value of a set of numbers.

ii. **Median :** The median is a statistical term that is one way of finding the 'average' of a set of data points.

iii. **Mode :** The mode of a set of data values is the value that appears most often.

iv. **Summary :** A summary table stores data that has been aggregated in a way that answers a common (or resource-intensive) business query.

```
> View(census)
> mean(census$Total.Males)
[1] 16391.56
> median(census$Total.Males)
[1] 15283
> mode(census$Total.Males)
[1] "numeric"
> summary(census$Total.Males)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
      0    9764   15283   16392   22220   52794
> IQR=(census$Total.Males)
>
```

MEASURES OF DISPERSION:

i. **Inter Quartile Range :** The interquartile range (IQR) is a measure of variability, based on dividing a data set into quartiles. Quartiles divide a rank-ordered data set into four equal parts.

ii. **Quartiles :** A quartile is a statistical term describing a division of observations into four defined intervals based upon the values of the data and how they compare to the entire set of observations.

iii. **Mid Range :** The arithmetic mean of the largest and the smallest values in a sample or other group.

iv. **Outlier :** An outlier is an observation that lies an abnormal distance from other values in a random sample from a population.

a) Lower Fence : Q1 - 1.5*IQ
b) Upper Fence : Q3 + 1.5*IQ
c) Outlier Values

```
> IQR(census$Total.Males)
[1] 12456
> quantile(census$Total.Males,0.25)
   25%
9763.5
> quantile(census$Total.Males,0.75)
    75%
22219.5
> range(census$Total.Males)
[1]     0 52794
> mean(range(census$Total.Males))
[1] 26397
> Lf<-quantile(census$Total.Males,0.25)-1.5*(IQR(census$Total.Males))
> print(Lf)
    25%
-8920.5
```

```
> Lf<-quantile(census$Total.Males,0.25)-1.5*(IQR(census$Total.Males))
> print(Lf)
    25%
-8920.5
> Uf<-quantile(census$Total.Males,0.25)+1.5*(IQR(census$Total.Males))
> print(Uf)
    25%
28447.5
> outlier_values<-boxplot.stats(census$Total.Males)$out
> print(outlier_values)
 [1] 52794 43128 50658 45113 46321 52364 45229 52358 45786 42283 42564
>
```

RESULT:

Thus the central tendency and measures of dispersion have been executed successfully. The outlier values are from more than upper fence there are no lower fence values.

EX.NO: 02
Date :

PLOTTING GRAPHS USING R-TOOL

PROBLEM STATEMENT:

Plot the boxplot, histogram and scatterplot for the dataset which was taken in the previous exercise.

DESCRIPTION:

Consider a dataset travel.times.csv, where it contains the attributes are Date, StartTime, DayOfWeek , Goingto , Distance, MaxSpeed , AvgSpeed, AvgMovingSpeed, FuelEconomy, TotalTime, MovingTime, Take407All, Comments for plotting the graphs.

IMPLEMENTATION:

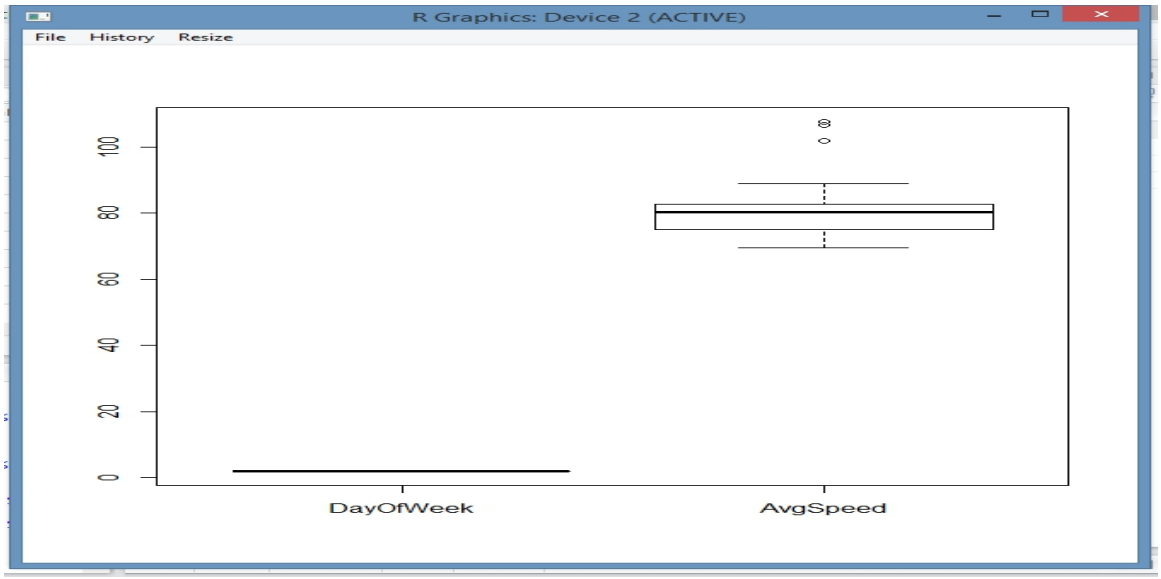
i. BoxPlot
ii. Histogram
iii. ScatterPlot

i. BOXPLOT :

A box plot is a graphical rendition of statistical data based on the minimum, first quartile, median, third quartile, and maximum. The term "box plot" comes from the fact that the graph looks like a rectangle with lines extending from the top and bottom. Because of the extending lines, this type of graph is sometimes called a box-and-whisker plot.Boxplot analysis made among the DayOfWeek and AvgSpeed.

- travel1<-travel.times[which(travel.times$DayOfWeek==”Friday”),names(travel.times)%in% c(“DayOfWeek”,”AvgSpeed”)]
- travel2<-travel.times[which(travel.times$DayOfWeek==”Monday”),names(travel.times)%in% c(“DayOfWeek”,”AvgSpeed”)]
- boxplot(travel1,travel2)
-

OUTPUT:

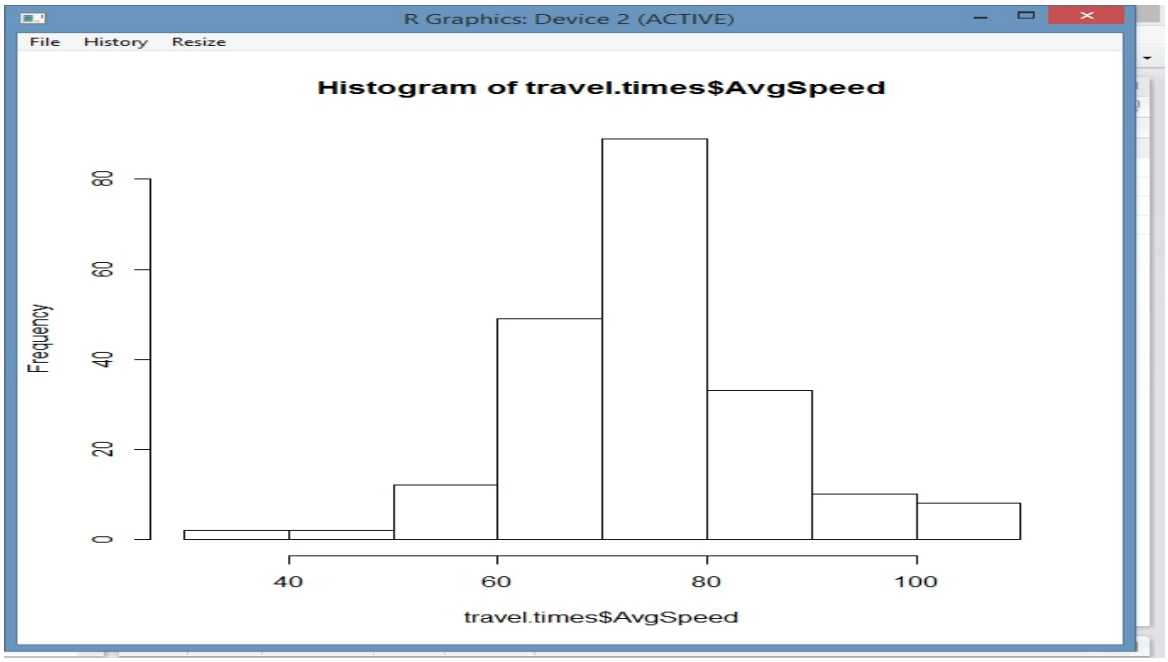


ii. HISTOGRAM:

A histogram is a display of statistical information that uses rectangles to show the frequency of data items in successive numerical intervals of equal size. In the most common form of histogram, the independent variable is plotted along the horizontal axis and the dependent variable is plotted along the vertical axis. The data appears as coloured or shaded rectangles of variable area.

- Hist(travel.times$AvgSpeed)

OUTPUT:

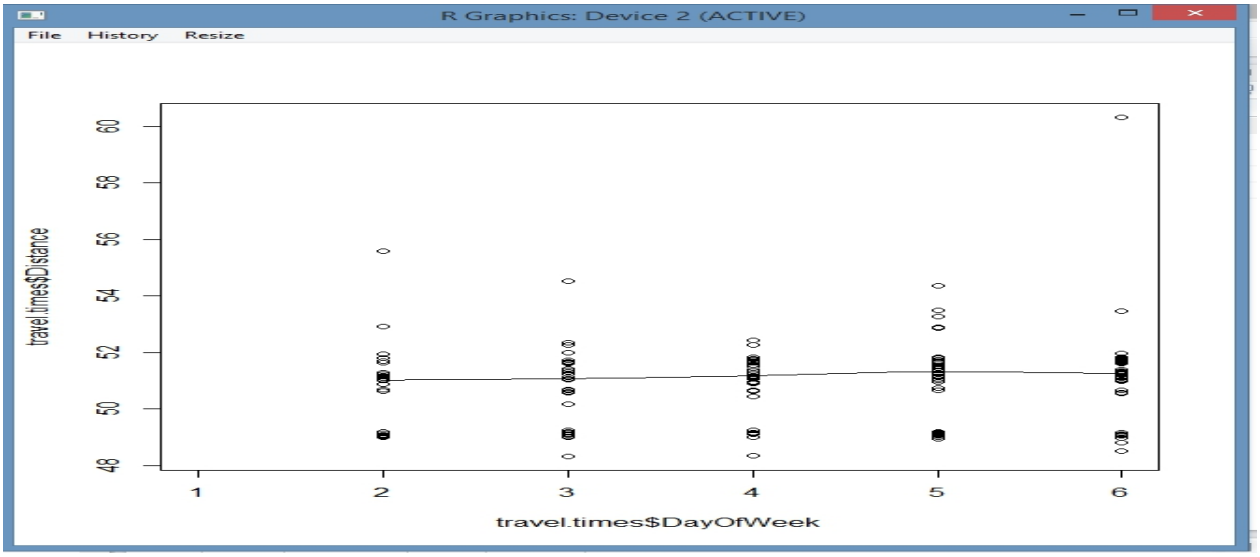


iii. SCATTERPLOT:

Scatter plots are important in statistics because they can show the extent of correlation, if any, between the values of observed quantities or phenomena (called variables). If no correlation exists between the variables, the points appear randomly scattered on the coordinate plane. Scatterplot is made between DayOfWeek and distance of the dataset travel.times.csv

- Scatter.smooth(travel.times$DayOfWeek,travel.times$Distance)

OUTPUT:



RESULT:

Thus, the plotting of graphs like boxplot, histogram and scatterplot for the given dataset has been successfully completed.

EX.NO : 03
Date :

PERFORM CORRELATION ANALYSIS AND NORMALIZATION USING R-TOOL

PROBLEM STATEMENT :

Perform the correlation analysis for the numerical attribute using pearson coefficient and for categorical attribute using chi-square and also, perform the normalization technique using min, max, z score and decimal scaling for the given data frames of particular dataset.

DESCRIPTION :

A dataset of name diabetes.csv is given for the correlation analysis, to calculate or to correlate between Age and Insulin and the same dataset for the performance of normalization technique.

- **CORRELATION ANALYSIS:**

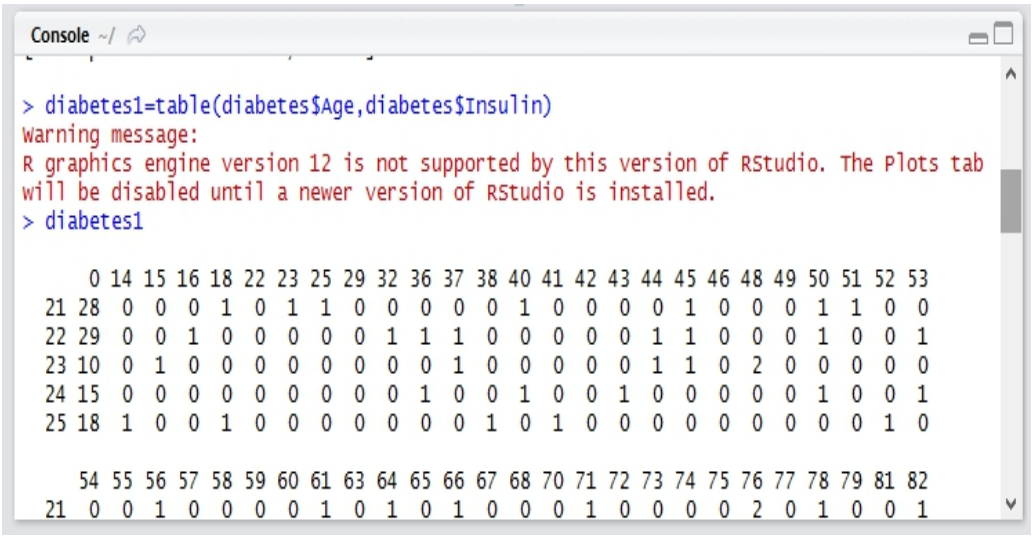
  STEPS INVOLVED:

  i. Create a new table with required dataframes.
  ii. After that apply the formula or query for the chi-square test.

  QUERIES:

  - diabetes1<-table(diabetes$Age,diabetes$Insulin)
  - diabetes1
  - chi sq.test(diabetes1)

  OUTPUT:

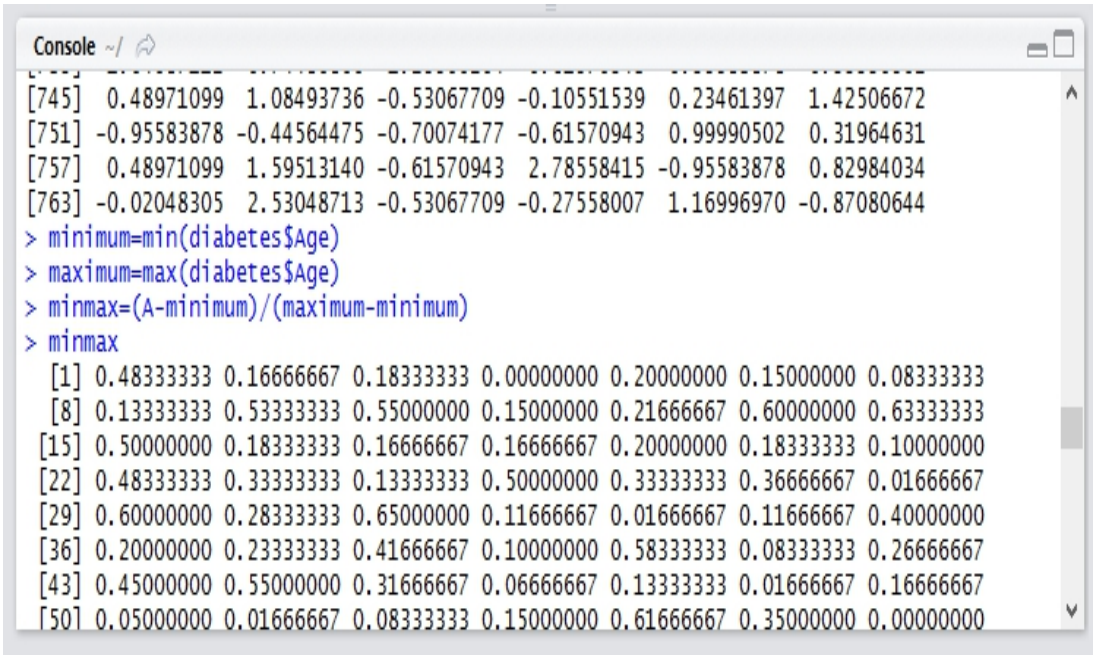


- **NORMALIZATION :**

i. MIN MAX NORMALIZATION:

- A<- c(diabetes$Age)
- Mean<-mean(A)
- Minimum<-min(diabetes$Age)
- Maximum<-max(diabetes$Age)
- MinMax<- (A-Minimum)/(Maximum-Minimum)
- MinMax

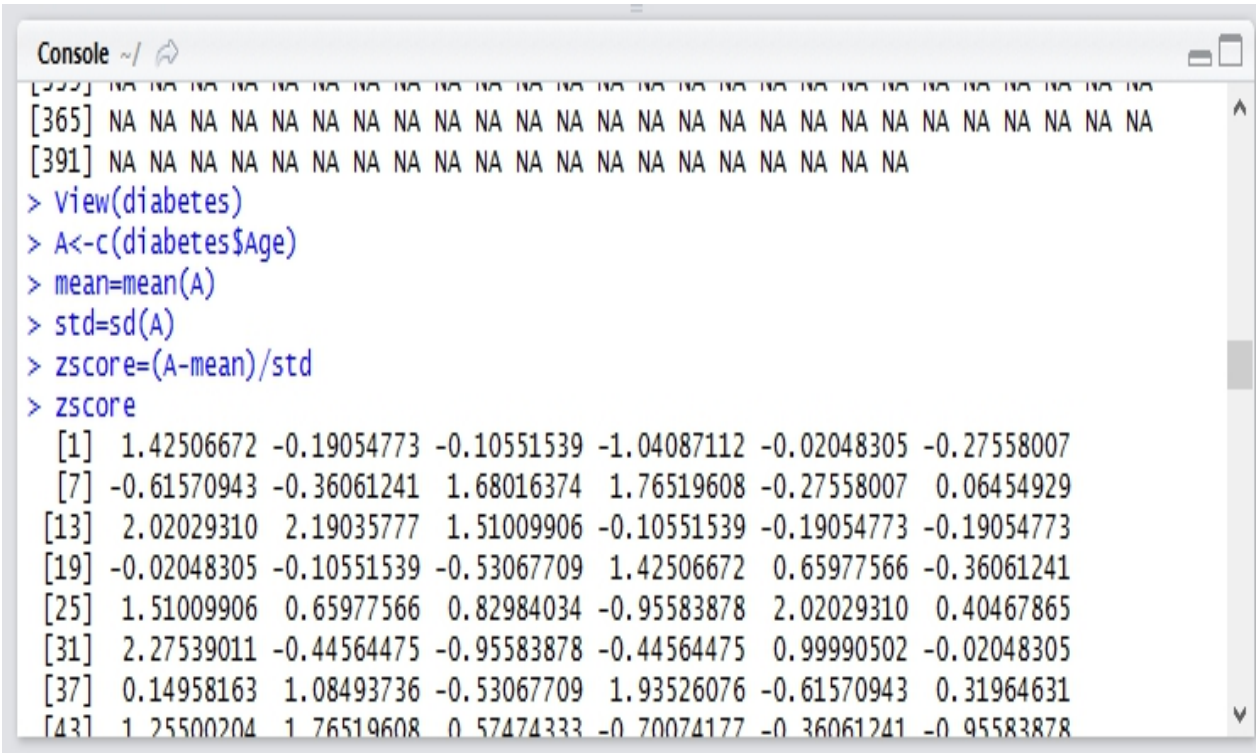
OUTPUT:



ii. Z SCORE NORMALIZATION :

- A<- c(diabetes$Age)
- Mean<- mean(A)
- Std<- sd(A)
- Zscore <- (A-Mean)/Std
- Zscore

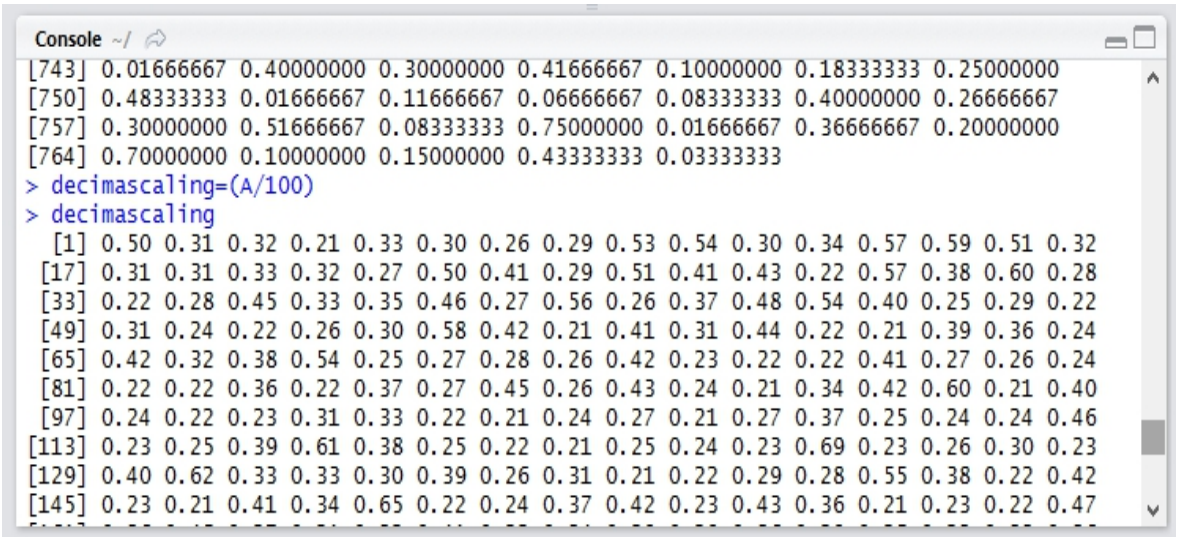
OUTPUT:



iii. DECIMAL SCALING NORMALIZATION :

- Decimalscaling =(A/100)
- Decimalscaling

OUTPUT:



RESULT:

Thus, the correlation analysis and normalization for the given dataset has been successfully executed and observed.

EX.No: 04
Date :

REGRESSION ANALYSIS USING R TOOL

PROBLEM STATEMENT :

Perform the linear regression and multiple regression for the given dataset.

DESCRIPTION :

Consider a dataset of diabetes.csv with the attributes pregnancies, Glucose, BloodPressure, SkinThickness, BMI, Diabetes, Age, Outcome for the analysis. There will be linear regression analysis between Age and BloodPressure . Where, for the multiple regression, the analysis is between Age, BloodPressure, Glucose from the dataset.

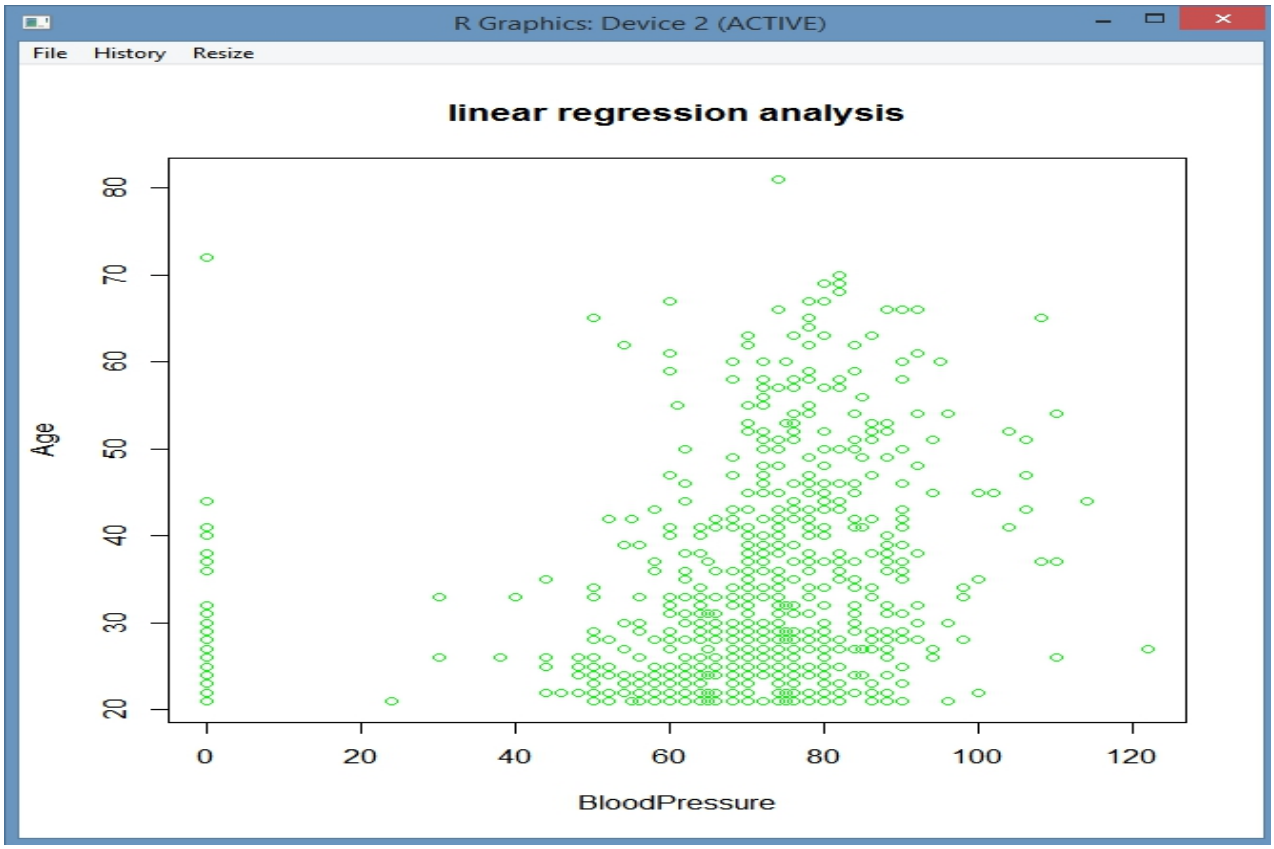
- **LINEAR REGRESSION :**

Linear regression is a kind of statistical analysis that attempts to show a relationship between two variables. Linear regression looks at various data points and plots a trend line. Linear regression can create a predictive model on apparently random data, showing trends in data, such as in cancer diagnoses or in stock prices.

QUERIES :

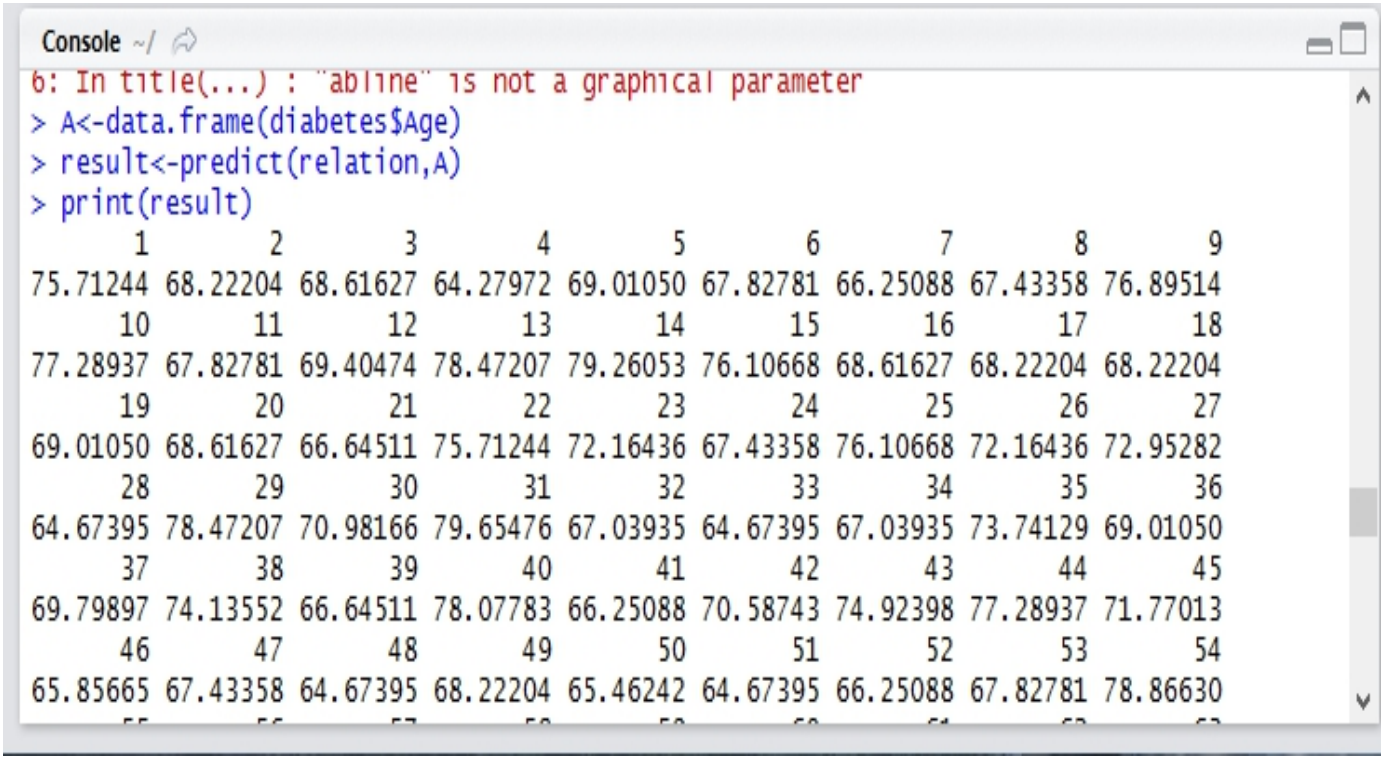
- Relation <- lm(diabetes$BloodPressure~diabetes$Age)
- Png<- (file=”linear regression.png”)
- Plot(diabetes$Age, diabetes$BloodPressure, col=”green”, main= “ Linear Regression Analysis” , abline= (lm(diabetes$BloodPressure~ diabetes$Age)), xlab = “BloodPressure”, ylanb= “Age”)

OUTPUT:



- A<- data.frame(diabetes$Age)
- Result<- predict(relation, A)
- Print(Result)

OUTPUT :



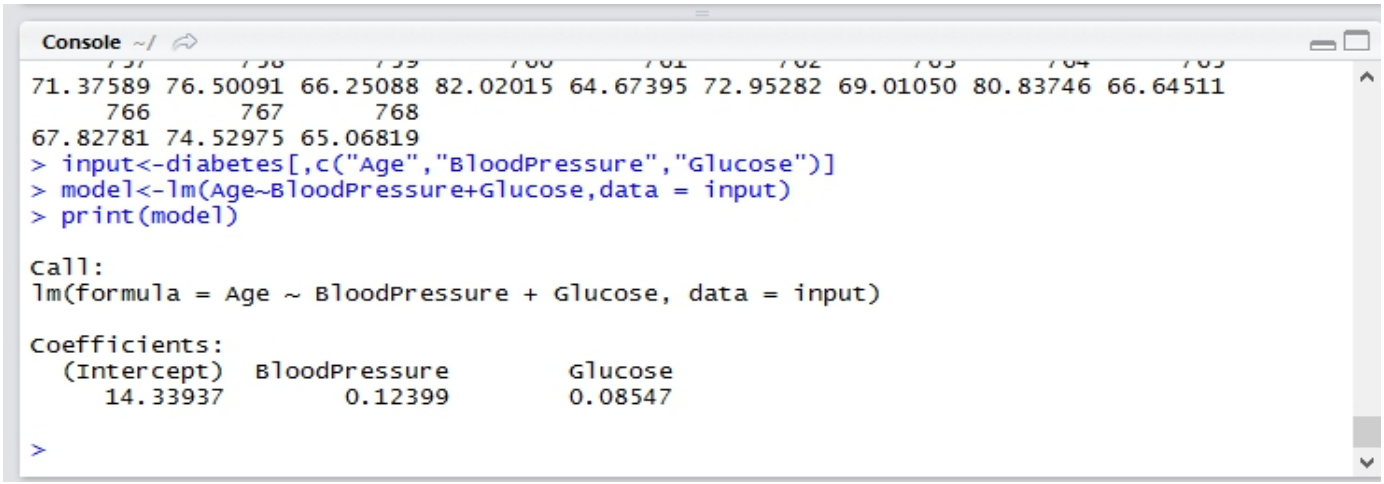
❖ **MULTIPLE REGRESSION :**

Multiple regression is a statistical tool used to derive the value of a criterion from several other independent, or predictor, variables. It is the simultaneous combination of multiple factors to assess how and to what extent they affect a certain outcome.

- **QUERIES :**

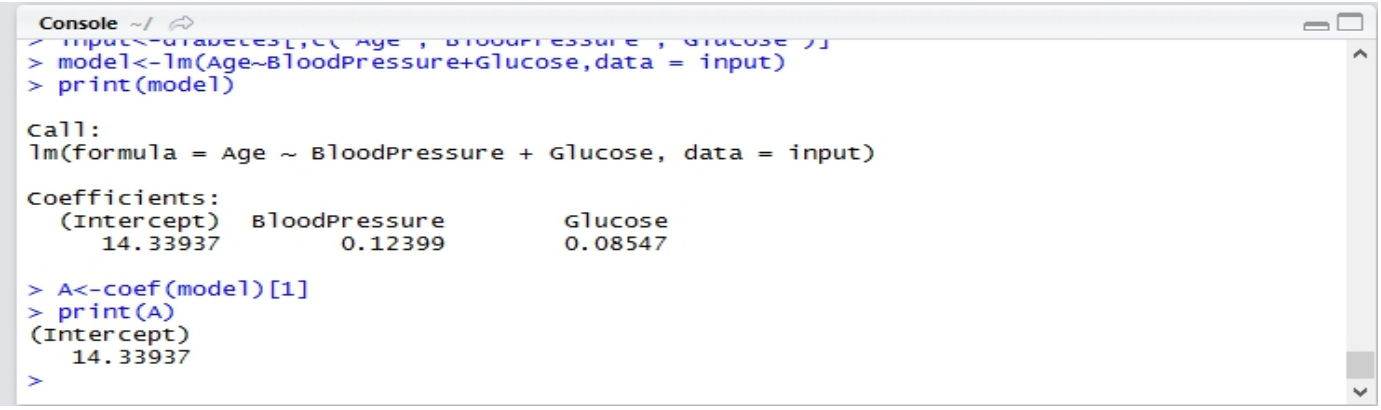
  - Input <- diabetes[,c(“Age”, “BloodPressure”, “Glucose”)]
  - Model <- lm(Age~ BloodPressure+Glucose,data=input)
  - Print(model)

OUTPUT:



- A<- coef(model)[1]
- Print(A)

OUTPUT:



- xBloodPressure<- coef(model)[2]
- yGlucose<- coef(model)[3]
- print(xBloodPressure)
- print(yGlucose)

OUTPUT:

- y = A+ xBloodPressure + yGlucose
- print(y)

OUTPUT:

RESULT :

Thus, the linear regression and the multiple regression analysis for the given dataset has been successfully completed.

EX.No: 05
Date :

DATA PREPROCESSING AND ANALYSIS FOR DATASET USING WEKA

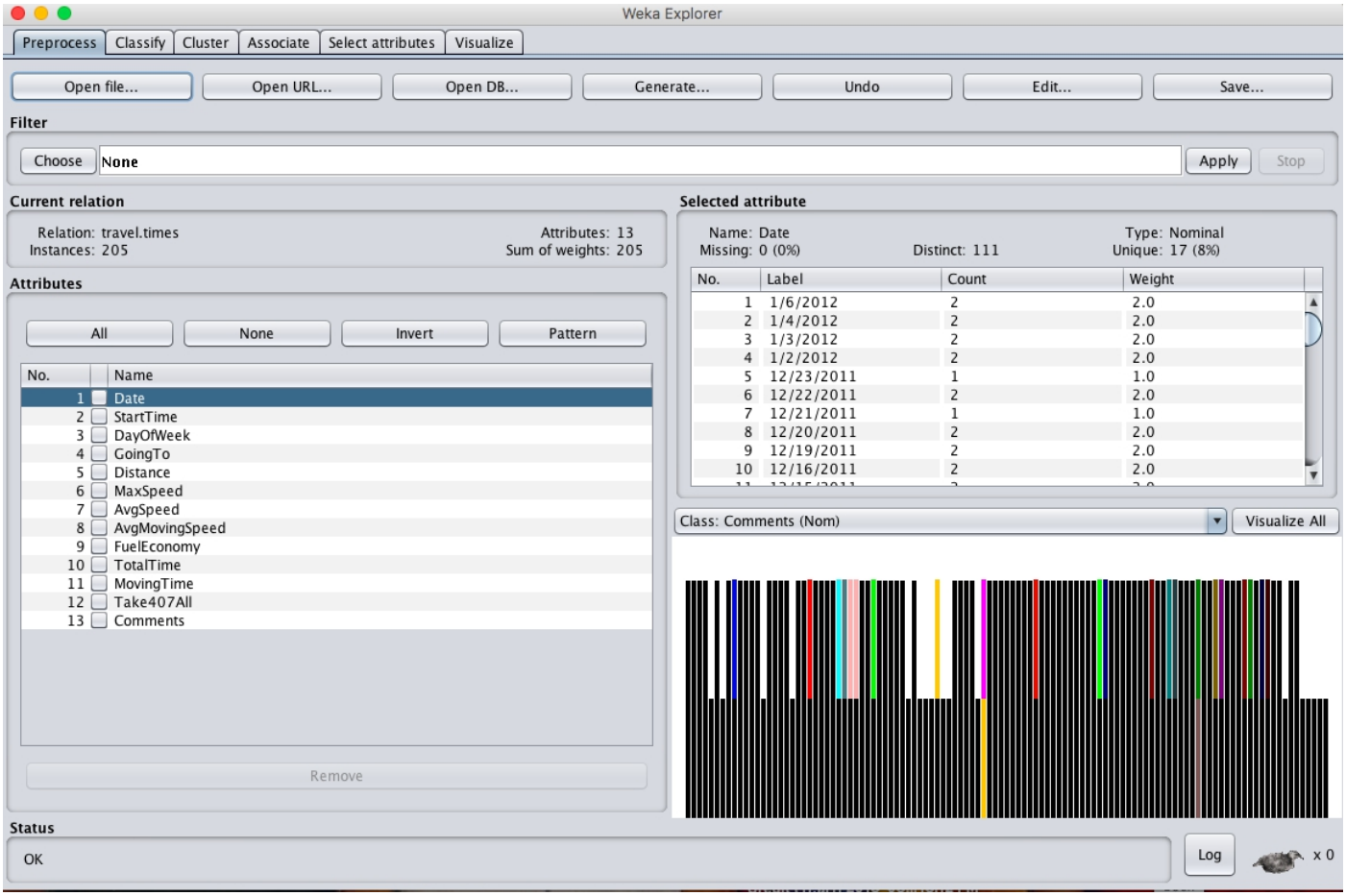
PROBLEM STATEMENT :

For each attribute in the dataset find the following information:

A. Attribute type
B. Percentage of missing values
C. Find the minimum, maximum, mean, Standard Deviation with numerical attributes.
D. Are there any records that have a values that no other record has ?
E. Write a note on class distribution for each of the attributes.
F. Apply attribute selection measures under filter supervised selection attribute.

DESCRIPTION :

Consider a dataset of traveltimes.csv file where it contains the columns of (or) attributes as Date, StartTime, DayOfWeek, GoingTo, Distance, MaxSpeed, AvgSpeed, AvgMovingSpeed, FuelEconomy, TotalTime, MovingTime, Take407All comments.



OBSERVATION :

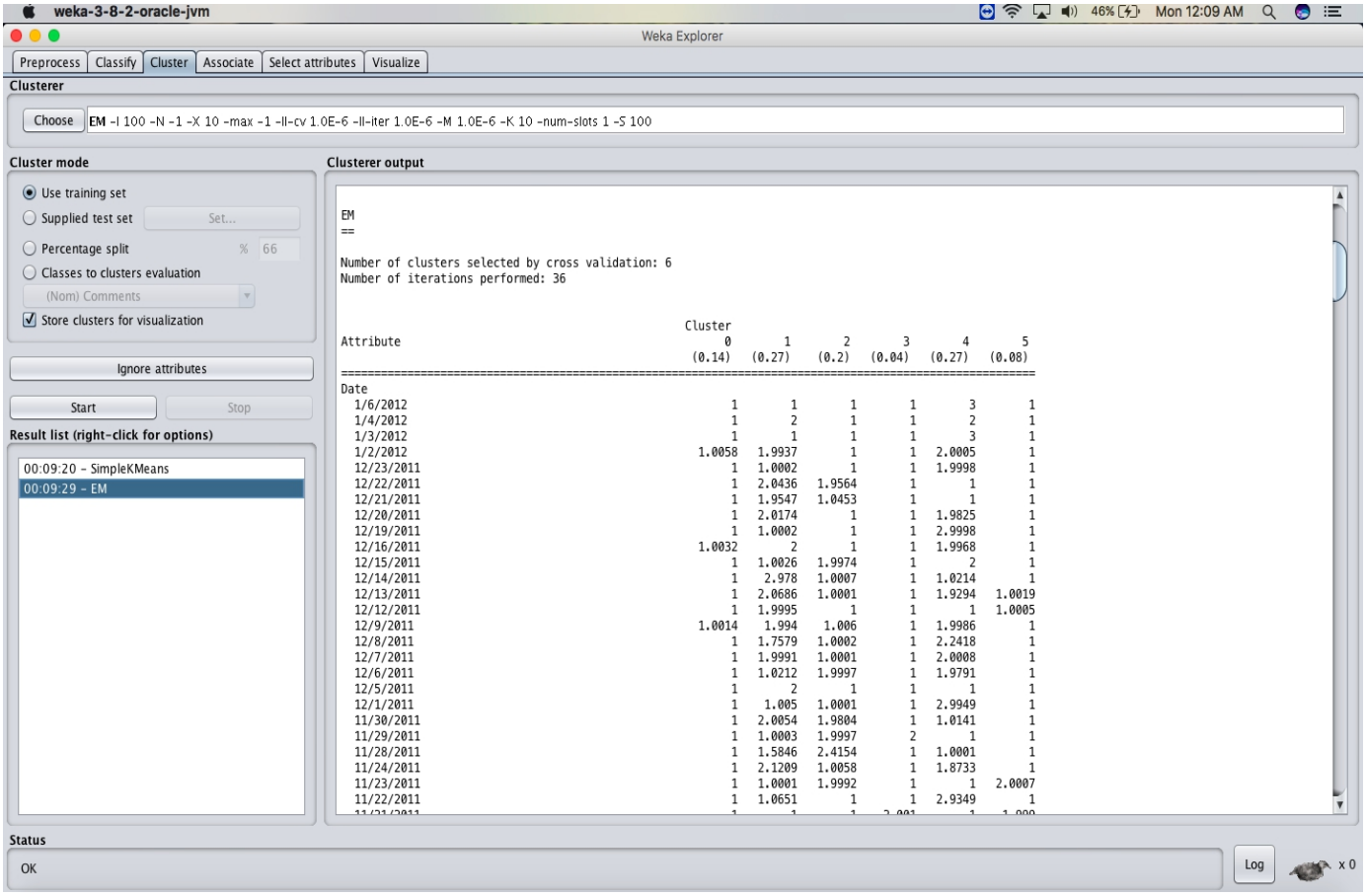
A. ATTRIBUTE TYPE :

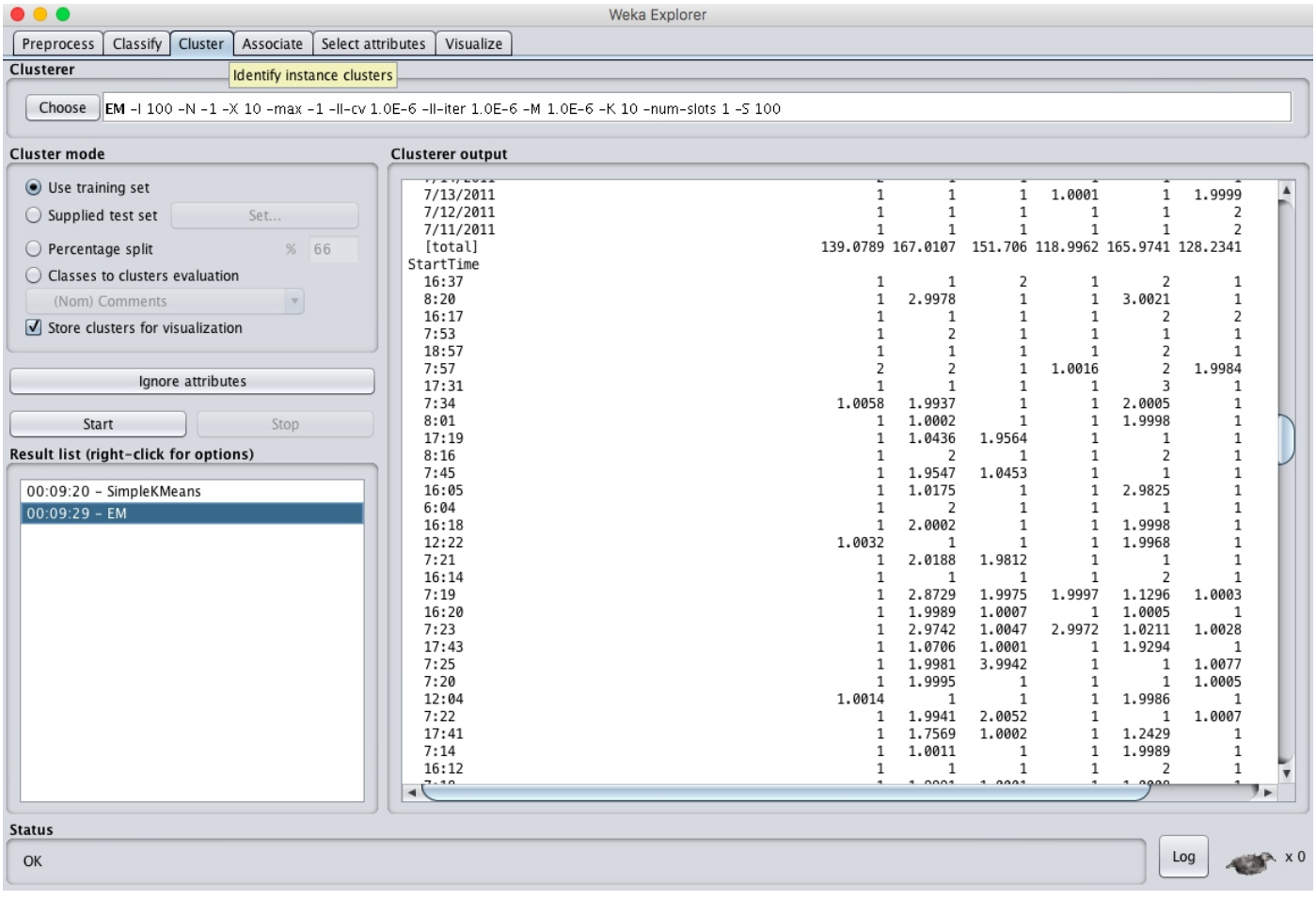
S.NO	ATTRIBUTE	TYPE
1.	Date	Nominal
2.	Start Time	Nominal
3.	Day Of Week	Nominal
4.	Going To	Nominal
5.	Distance	Numeric
6.	Max Speed	Numeric
7.	Avg Speed	Numeric
8.	Avg Moving Speed	Numeric
9.	Fuel Economy	Nominal
10.	Total Time	Numeric
11.	Moving Time	Numeric
12.	Comments	Nominal
13.	Take 407 All	Nominal

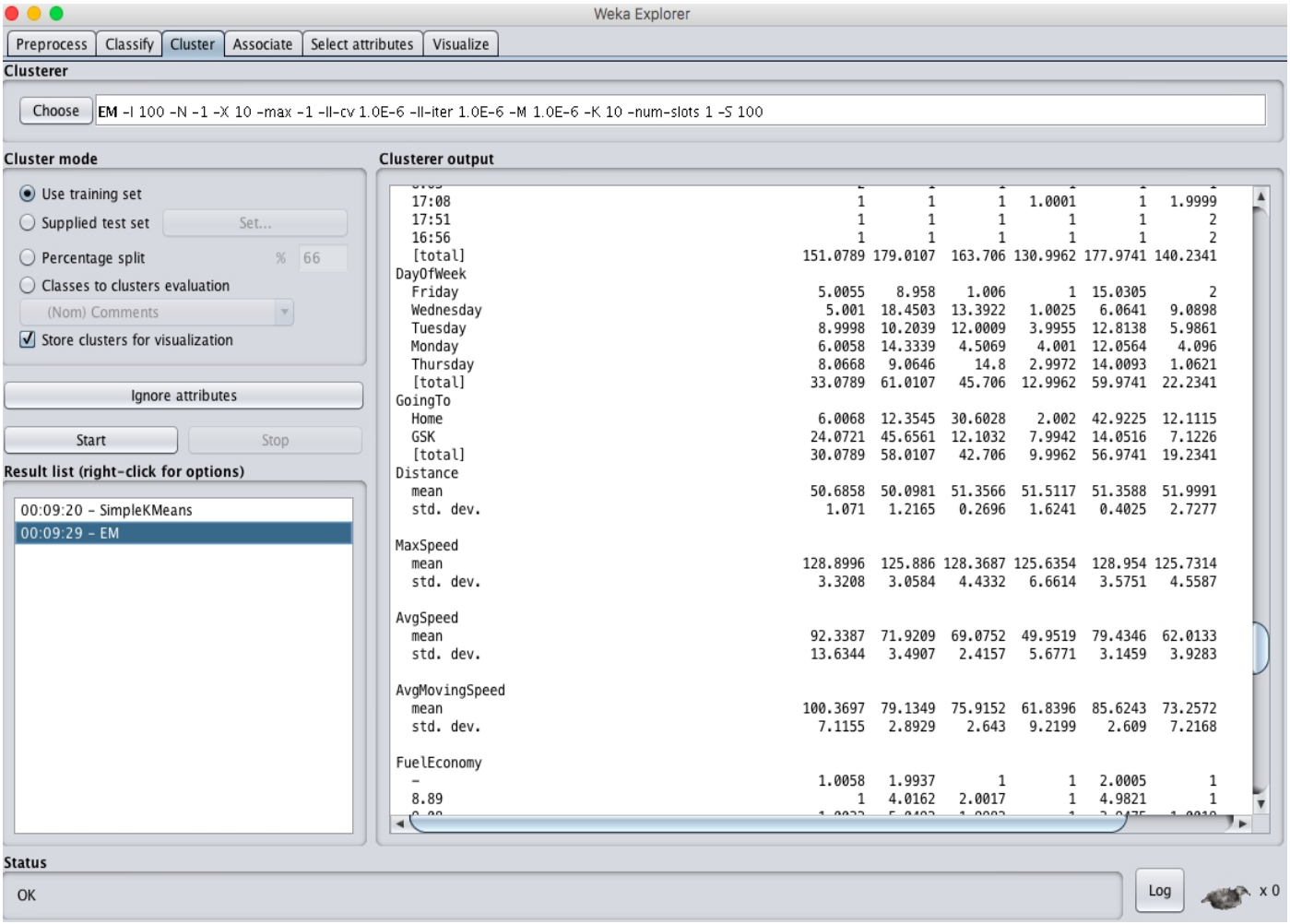
B. PERCENTAGE OF MISSING VALUES :

S.NO	ATTRIBUTE	Percentage Of Missing Values
1.	Date	0 %
2.	Start Time	0 %
3.	Day Of Week	0 %
4.	Going To	0 %
5.	Distance	0 %
6.	Max Speed	0 %
7.	Avg Speed	0 %
8.	Avg Moving Speed	0 %
9.	Fuel Economy	8 %
10.	Total Time	0 %
11.	Moving Time	0 %
12.	Comments	88 %
13.	Take 407 All	0 %

C. MIN, MAX, MEAN, STANDARD DEVIATION :







D. Are there any records that have a value that no other records has?

Avg Speed is unique from all of the other attributes records. Its uniqueness percentage is all about 60 % out of everything. So this would be the attribute where no other record has this records value.

E. Write a note on class distribution for each of the attributes.

Mainly, all the attributes are distributed to 2 types. They are :

1. Nominal
2. Numeric

- Nominal class attributes are : Date, Start Time, Day Of Week, Going To, Fuel Economy, Total 407 All, Comments.
- Numeric class attributes are : Distance, MaxSpeed, AvgSpeed, Avg Moving Speed, Total Time, Moving Time.

F. Apply attribute selection measures under filter supervised selection.

When we apply attribute selection under the filter of supervised attribute selection. Initially, it had 13 attributes but after the filter of the attributes count is been reduced to 11, where AvgMaxSpeed and Total 407 All are removed.

RESULT :

Thus, the dataprocessing and analysis for a dataset using weka tool has been successfully completed.

EX.No: 06
Date :

DATA SEGMENTATION BY K- MEANS CLUSTER USING WEKA AND R-TOOL

PROBLEM STATEMENT :

Apply K-means algorithm to your dataset experiment with the algorithm as follows: By setting the number of elements and seed of the random algorithm for generating initial cluster centres . Compare the results that has occurred between K-means and R-Tool .

DESCRIPTION :

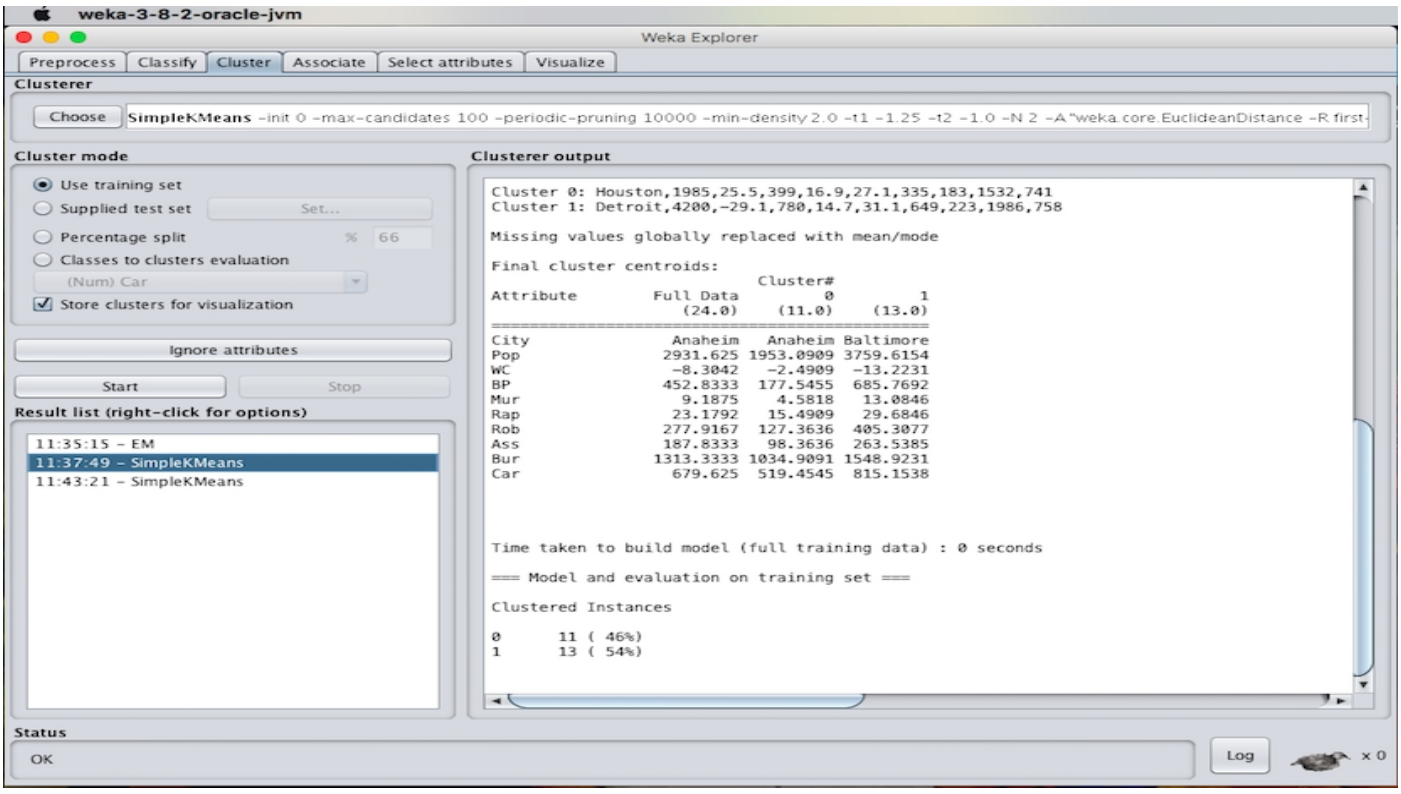
Consider a dataset of citycrimes.csv file of which it contains the attributes are City, Pop, WC, BP, Mur, Rap, Rob, Ass, Bus and car for the performance of the dataset by applying the K-means algorithm in weka and as well using R- tool.

- **USING WEKA TOOL :**

A. Choose a set of attributes for clustering and give a motivation.

STEPS INVOLVED :

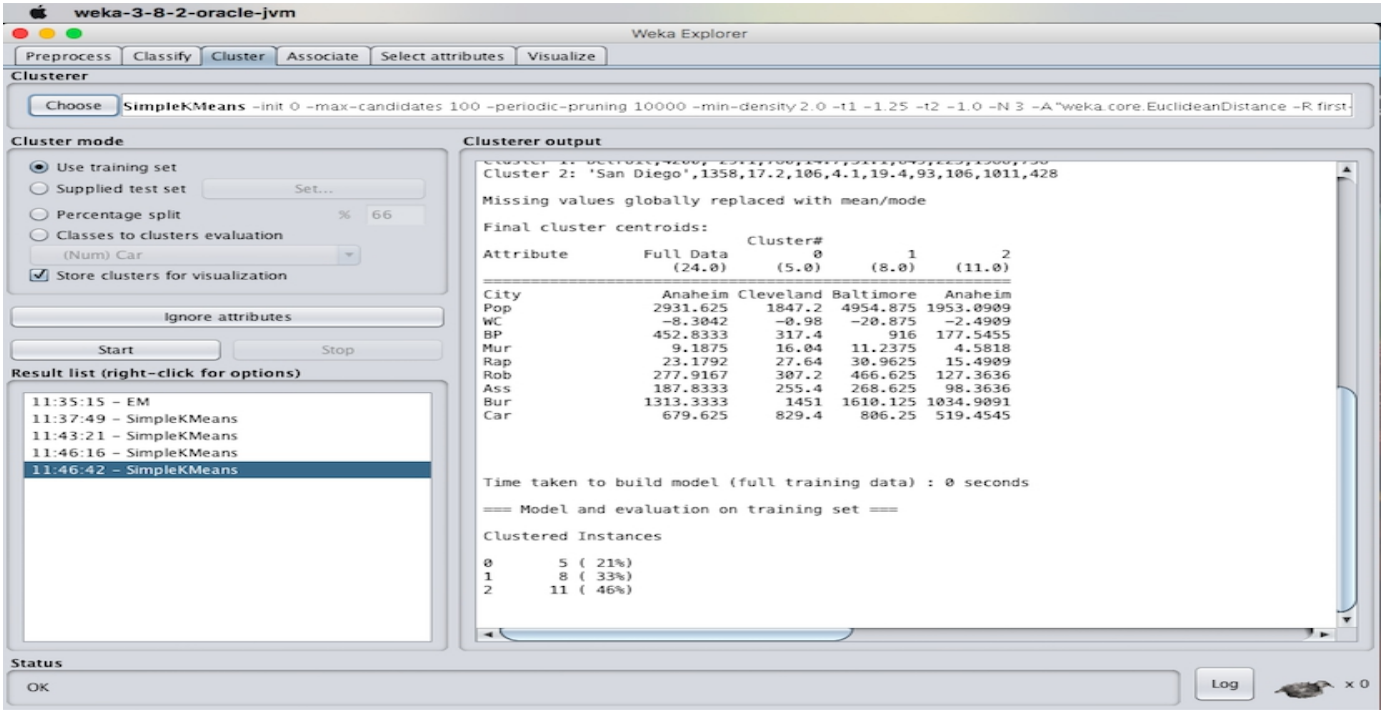
- Choose a set of attributes for clustering and for giving a motivation.
- Choose the dataset and import the dataset into Weka tool.
- Cluster the dataset and choose simple K-means algorithm and give the motivation.

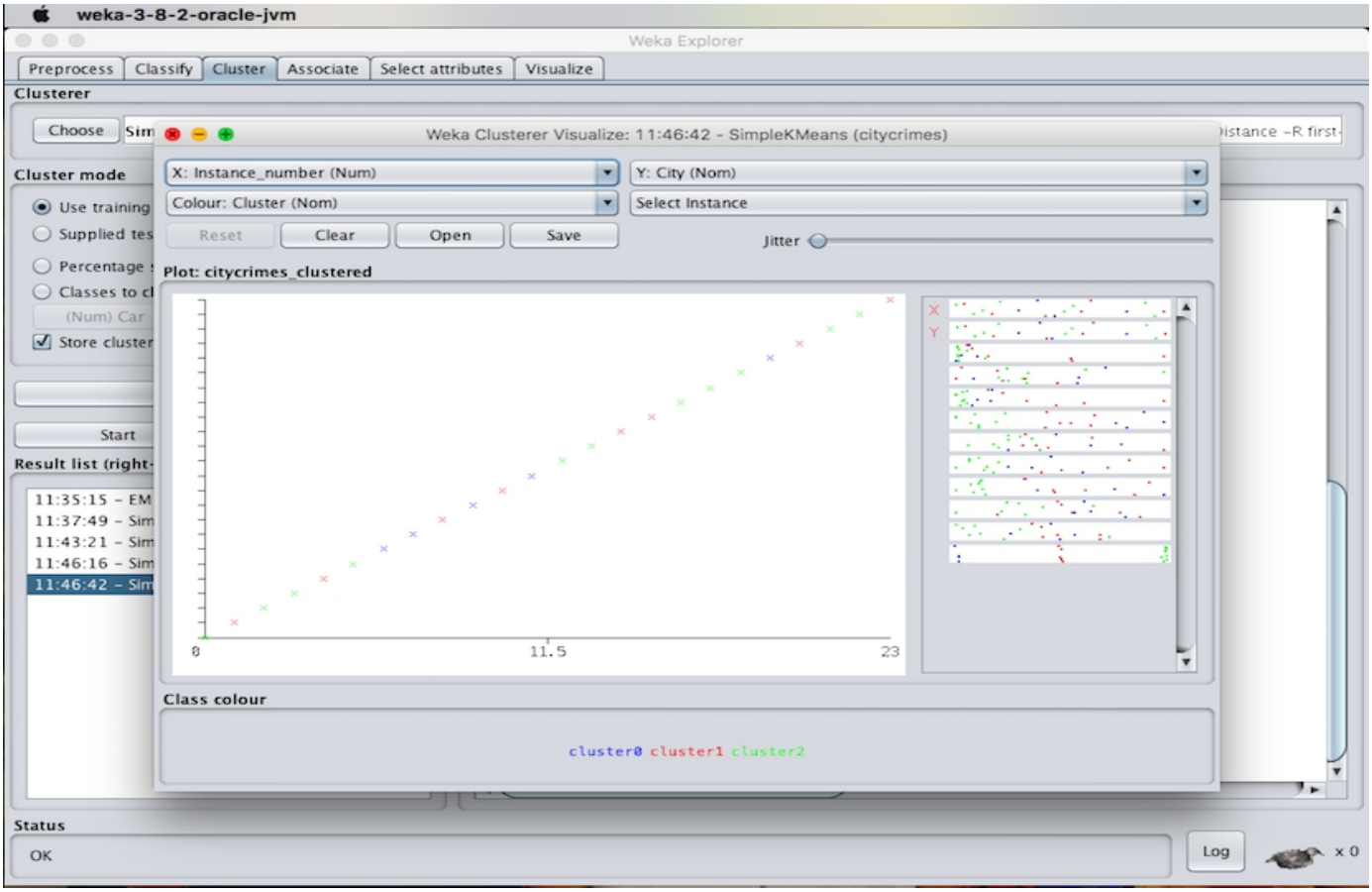


B. Experiment with atleast 2 different number of clusters but with same seed values:

STEPS INVOLVED :

- Compare the two different clusters but with the same seed values.
- Change the number of clusters value and need not to change the seed value.
- Apply the K-means algorithm and start executing the algorithm.





C. **Try with the different seed values . Explain what is the seed value controls**.

It was observed that when there is an increase in the seed value from the standard seed value 10 to higher ranges. The number of iterations will be reduced. In the case of seed value 10, for the given citycrime.csv dataset. It generated '6' iterations whereas for seed value 100 it has generated only '2' iterations.

Finally, there will be the change in the sequence of the tuples in the output and in clustered instances percentage will be changed. Seed value 100 controls the number of iterations.

❖ **USING R-TOOL :**

STEPS INVOLVED :

- Choose the dataset and import the dataset into the R-tool.
- View the dataset and start inserting queries for the k means clustering algorithm.

QUERIES :

- set.seed(20)
- clusters <- kmeans(citycrimes[,2:3], 5)
- citycrimes$Borough <- as.factor(clusters$cluster)
- str(clusters)

```
>
> clusters <- kmeans(citycrimes[,2:3], 5)
>
> citycrimes$Borough <- as.factor(clusters$cluster)
> str(clusters)
```

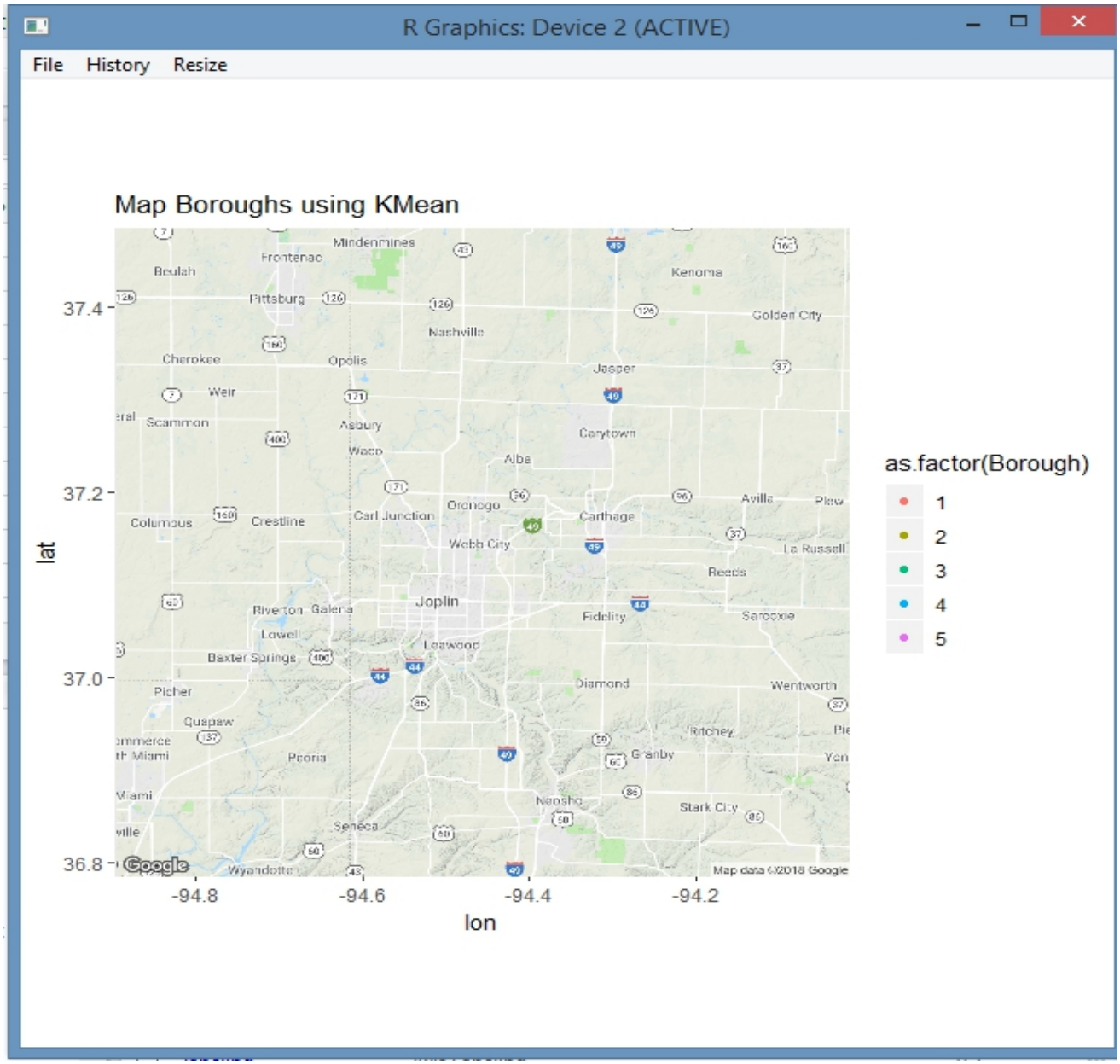
```
Console ~/
List of 9
 $ cluster     : int [1:24] 4 1 3 4 2 4 1 4 5 1 ...
 $ centers     : num [1:5, 1:2] 2079 8513 2908 1391 4509 ...
  ..- attr(*, "dimnames")=List of 2
  .. ..$ : chr [1:5] "1" "2" "3" "4" ...
  .. ..$ : chr [1:2] "Pop" "WC"
 $ totss       : num 1.39e+08
 $ withinss    : num [1:5] 315574 13643048 67068 52387 191844
 $ tot.withinss: num 14269920
 $ betweenss   : num 1.25e+08
 $ size        : int [1:5] 7 3 3 9 2
 $ iter        : int 3
 $ ifault      : int 0
 - attr(*, "class")= chr "kmeans"
> library(ggmap)
```

- library(ggmap)
- Map <- get_map("citycrimes:,zoom=10)
- ggmap(Map) + geom_point(aes(x = Pop[] , y = WC[] , colour = as.factor(Borough)) , data = citycrimes) +
  ggtitle("Map Boroughs using KMean")

```
Console ~/
<environment: namespace:stats>
> ggmap(Map) + geom_point(aes(x = Pop[], y = WC[], colour = as.factor(Borough)),data = c
itycrimes)
Warning message:
Removed 24 rows containing missing values (geom_point).
> ggtitle("Map Boroughs using KMean")
$title
[1] "Map Boroughs using KMean"

$subtitle
NULL

attr(,"class")
[1] "labels"
>
>
```



RESULT :

Thus, the K-means clustering analyzing using both the weka tool and R- tool has been successfully completed. In case of weka tool, the change in seed values lead to the decrease in the number of iterations. In case of R-tool, there are only 3 number of iterations.

EX.No: 07
Date :

DATA SEGMENTATION BY EXPECTATION MAXIMISATION ALGORITHM THROUGH WEKA

PROBLEM STATEMENT :

Analyze the dataset using Expectation Maximization algorithm(EM) by setting the minimum standard deviation for normal density calculation and compare the results with the simple K-means algorithm.

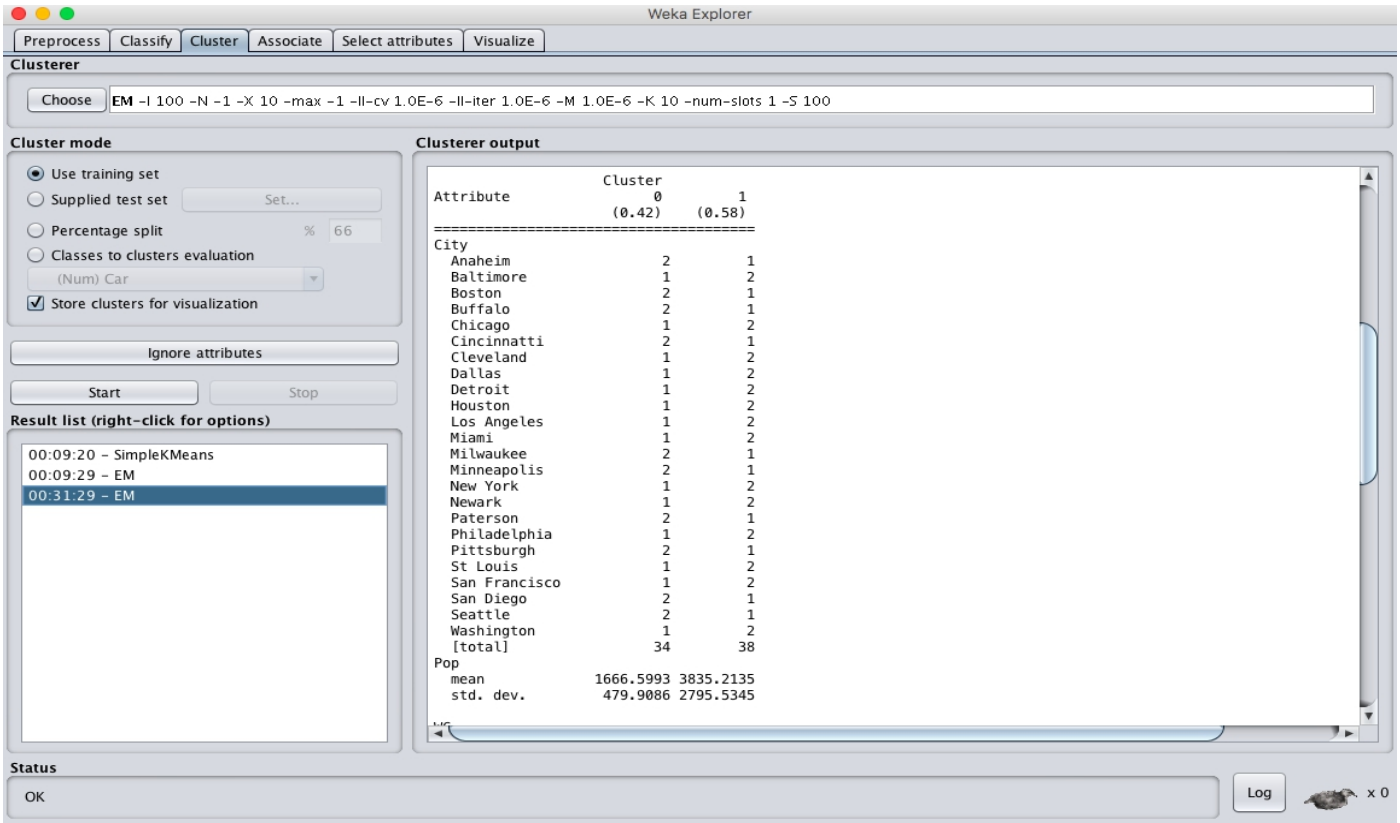
DESCRIPTION :

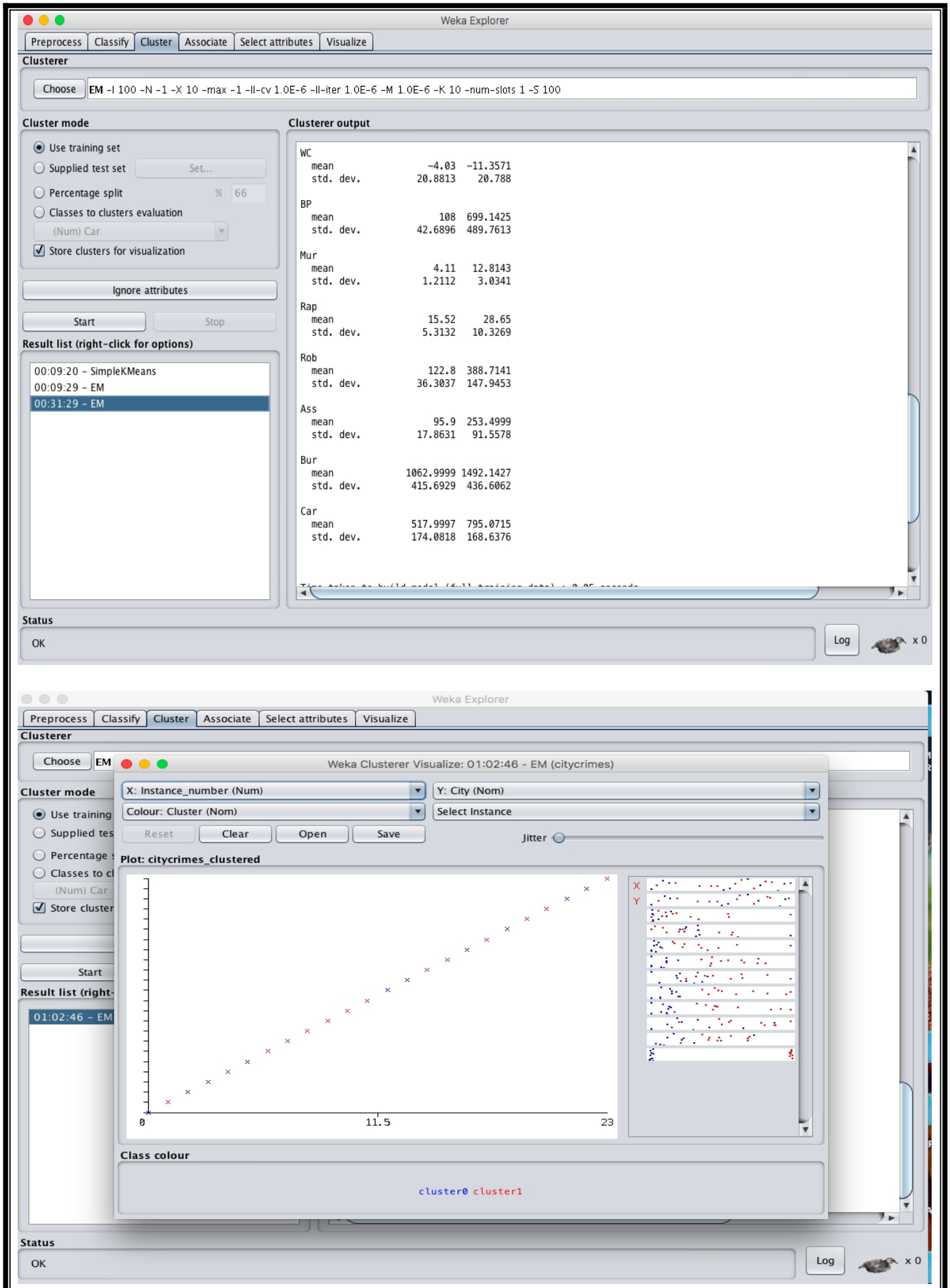
Consider a dataset of citycrimes.csv file of which it contains the attributes are City, Pop, WC, BP, Mur, Rap, Rob, Ass, Bus and car for the performance of the dataset by applying the K-means algorithm in weka and as well using R- tool.

When the clustering is been made through the expectation maximization algorithm by setting minimum standard deviation values then the results will be of the following :

- **Steps Involved :**

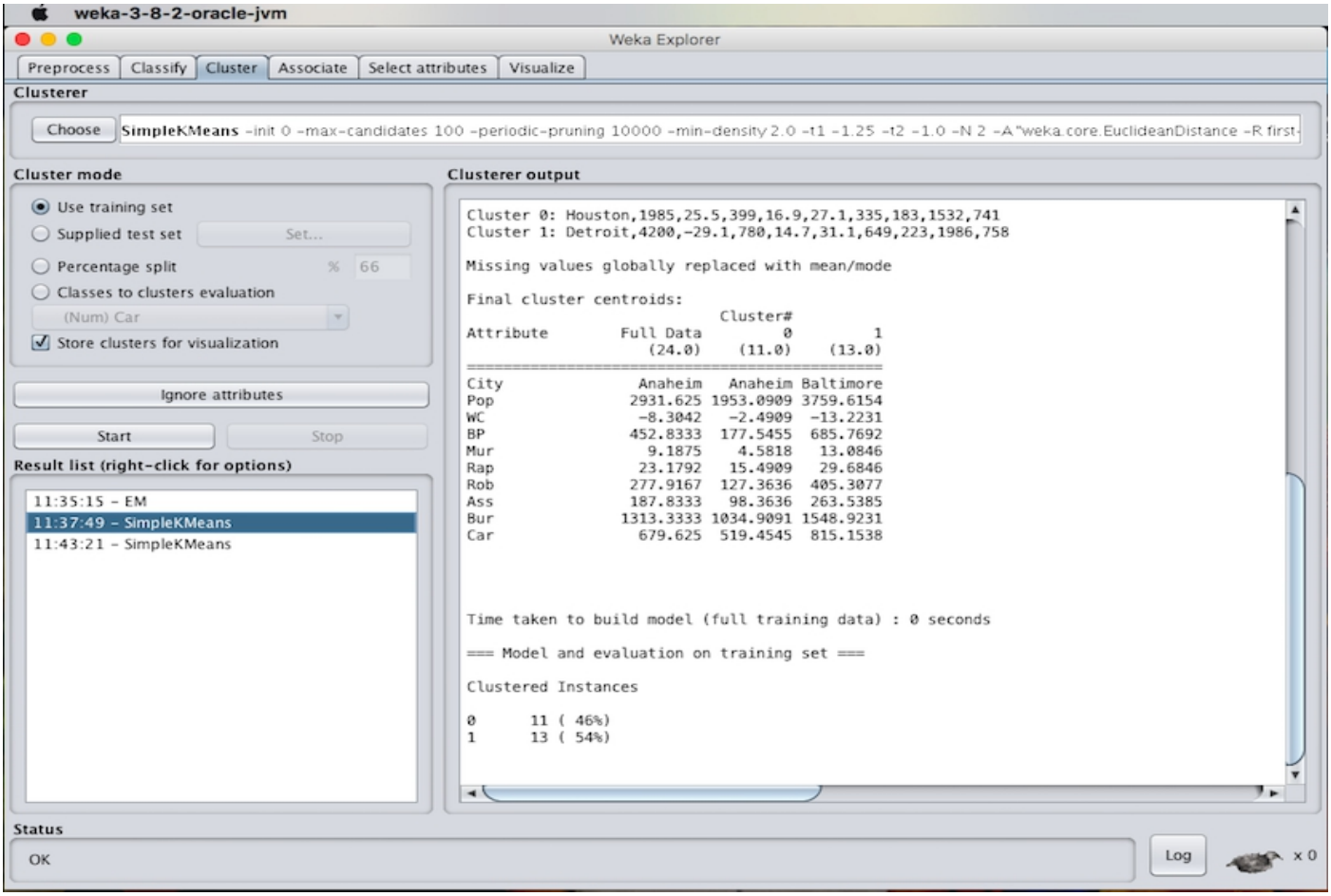
  - Initially, load the dataset into the weka tool and check for all the attributes present in the dataset.
  - Then move to cluster panel and apply the EM algorithm technique for the datasheet.
  - Finally, Observe the results that are obtained.




Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Clusterer
Choose
EM -I 100 -N -1 -X 10 -max -1 -ll-cv 1.0E-6 -ll-iter 1.0E-6 -M 1.0E-6 -K 10 -num-slots 1 -S 100
Cluster mode
Use training set
Supplied test set
Set...
Percentage split
%
66
Classes to clusters evaluation
(Num) Car
Store clusters for visualization
Ignore attributes
Start
Stop
Result list (right-click for options)
00:09:20 - SimpleKMeans
00:09:29 - EM
00:31:29 - EM
Clusterer output
WC
mean -4.03 -11.3571
std. dev. 20.8813 20.788
BP
mean 108 699.1425
std. dev. 42.6896 489.7613
Mur
mean 4.11 12.8143
std. dev. 1.2112 3.0341
Rap
mean 15.52 28.65
std. dev. 5.3132 10.3269
Rob
mean 122.8 388.7141
std. dev. 36.3037 147.9453
Ass
mean 95.9 253.4999
std. dev. 17.8631 91.5578
Bur
mean 1062.9999 1492.1427
std. dev. 415.6929 436.6062
Car
mean 517.9997 795.0715
std. dev. 174.0818 168.6376
Status
OK
Log
x 0

Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Clusterer
Choose
EM
Weka Clusterer Visualize: 01:02:46 - EM (citycrimes)
X: Instance_number (Num)
Y: City (Nom)
Colour: Cluster (Nom)
Select Instance
Reset
Clear
Open
Save
Jitter
Plot: citycrimes_clustered
0
11.5
23
Class colour
cluster0 cluster1
Cluster mode
Use training
Supplied tes
Percentage
Classes to cl
(Num) Car
Store cluster
Start
Result list (right-
01:02:46 - EM
Status
OK
Log
x 0

❖ K- MEANS ALGORITHM:



COMPARISION :

When compared to both the algorithms for the same dataset there will be a change in time taken to build model will be little longer in EM than when compared to K-means and there will be a percentage change in the clustered instances values.

RESULT :

Thus, the data analysis by the expectation maximization algorithm using weka has been analyzed and observed properly .

EX.No: 08
Date :

DATA SEGMENTATION BY COBWEB – HIERARCHIAL CLUSTERING ALGORITHM USING WEKA TOOL

PROBLEM STATEMENT :

For the given data file find the following using weka:

A. what are the most alive groups in terms of number of people involved, cluster together.

B. what are the most active community.

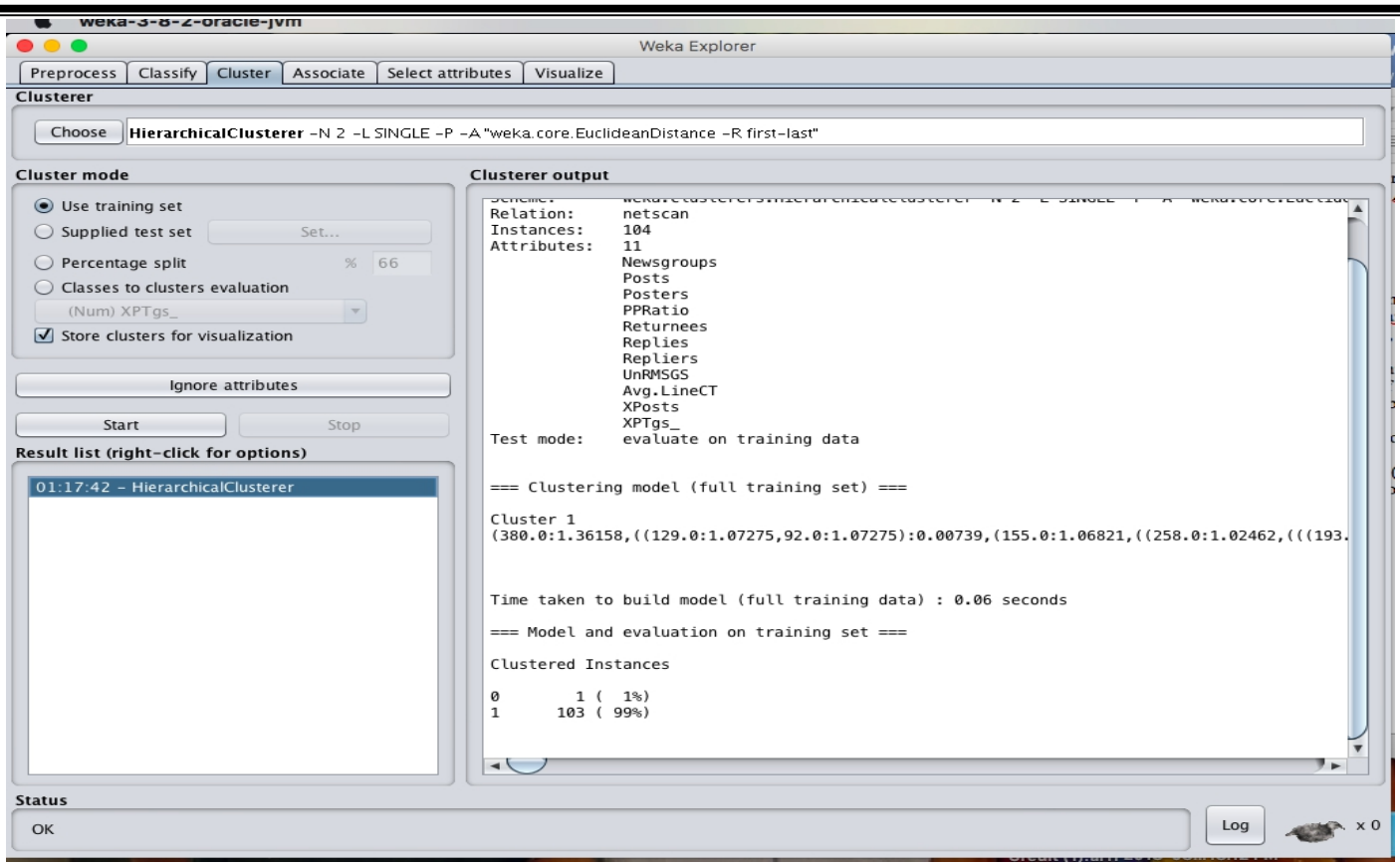
DESCRIPTION :

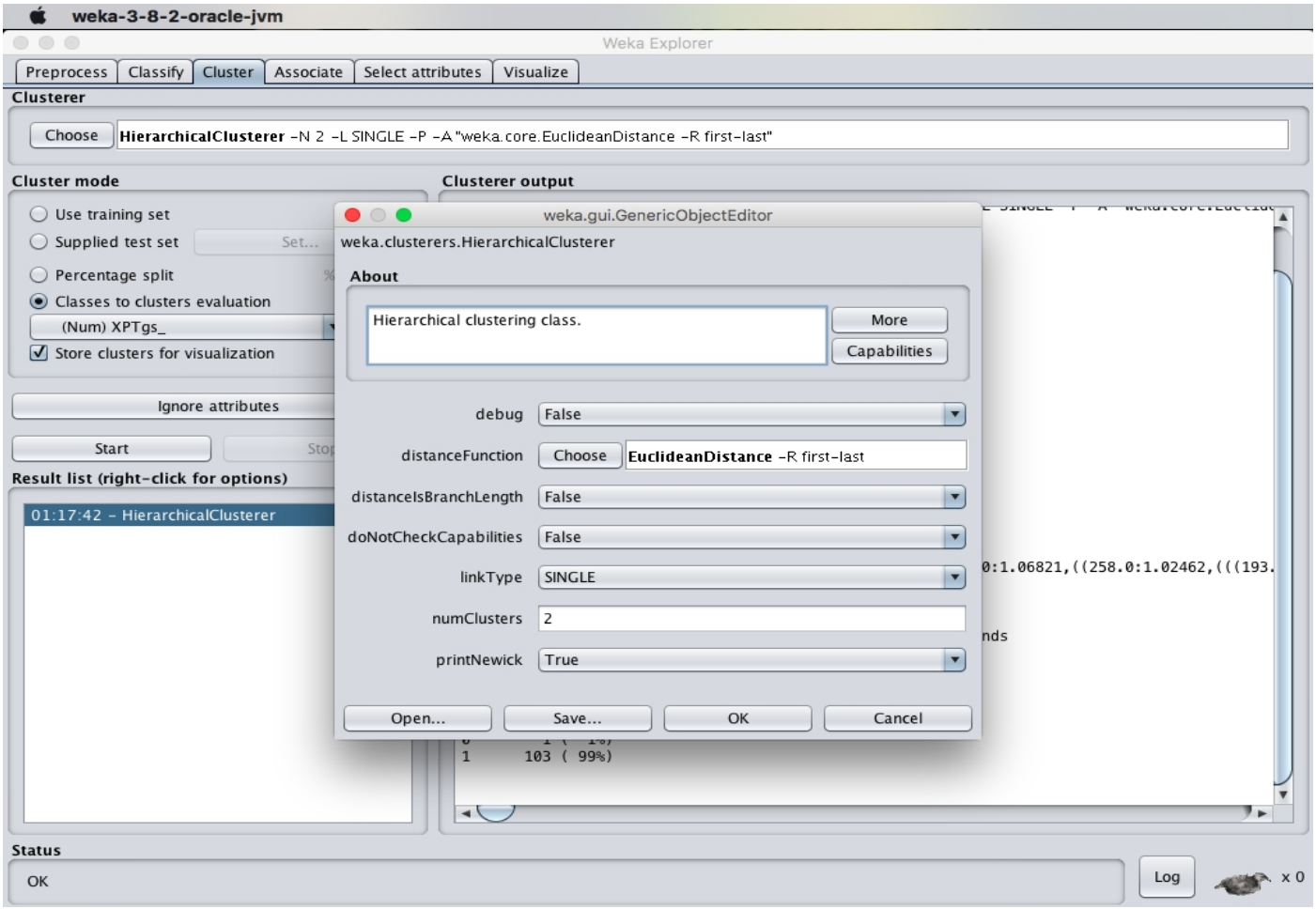
Consider a dataset netscan.csv where it contains the attributes of Newsgroups, posts, posters, PPRatio, Returnees, Replies, Repliers, UnRMsgs, Avg.LineCT, Xports, XPTgs. Each attribute will have different types of the meanings.

OBSERVATIONS :

A. The most active groups in terms of the number of people involved cluster together. Those groups - microsoft.public.windowsxp.perform_maintain, microsoft.public.windowsxp.network_web, microsoft.public.windowsxp.security_admin, microsoft.public.windowsxp.hardware - are all advanced user groups.
    1. They look like very active communities. (large number of posters, repliers, posts, and etc.)
    2. However, there are large number of isolated messages that might be questions with no answers yet, or might be questions ignored because they look like uninteresting to the advance users in those groups.

B. microsoft.public.es.* groups tightly cluster together except for the .windowsxp group. They share the followings.
    1. Relatively large number of XPosts (crosspostings) : reference many postings in other groups.
    2. Low PPRatio : Small number of posters post large number of postings.
    3.





RESULT :

Thus, the data analysis of cobweb hierarchial clustering algorithm using weka tools has been analyzed and observed successfully.

EX.No: 09
Date :

FREQUENT PATTERN MINING USING ASSOCIATION RULE THROUGH WEKA AND R TOOLS

PROBLEM STATEMENT :

Run the Apriori algorithm, and explore the association rules by changing the following parameters:

a) Upper bound min_sup
b) Lower bound min_sup
c) Metric type
d) Output itemsets

Implement the aprirori algorithm through R Tool and compare the results obtained through the weka.

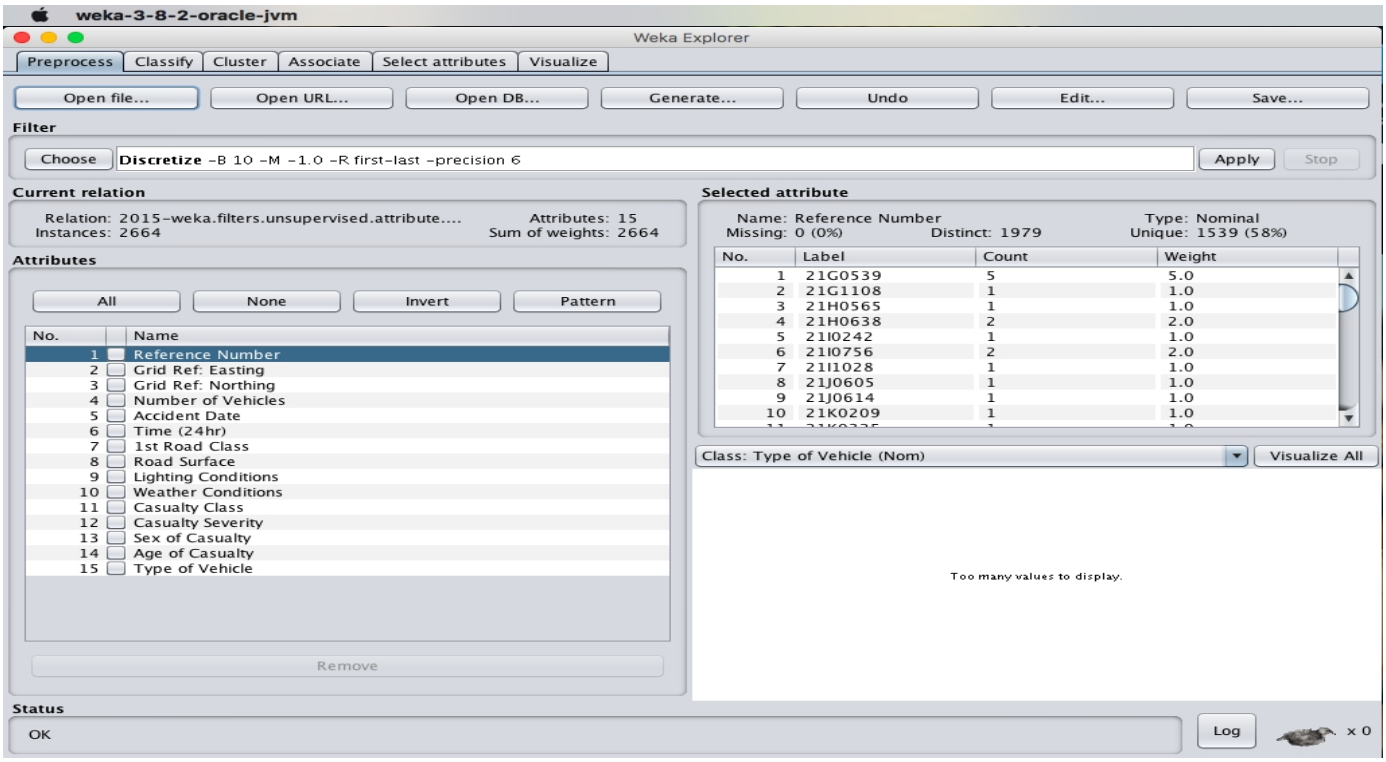
DESCRIPTION :

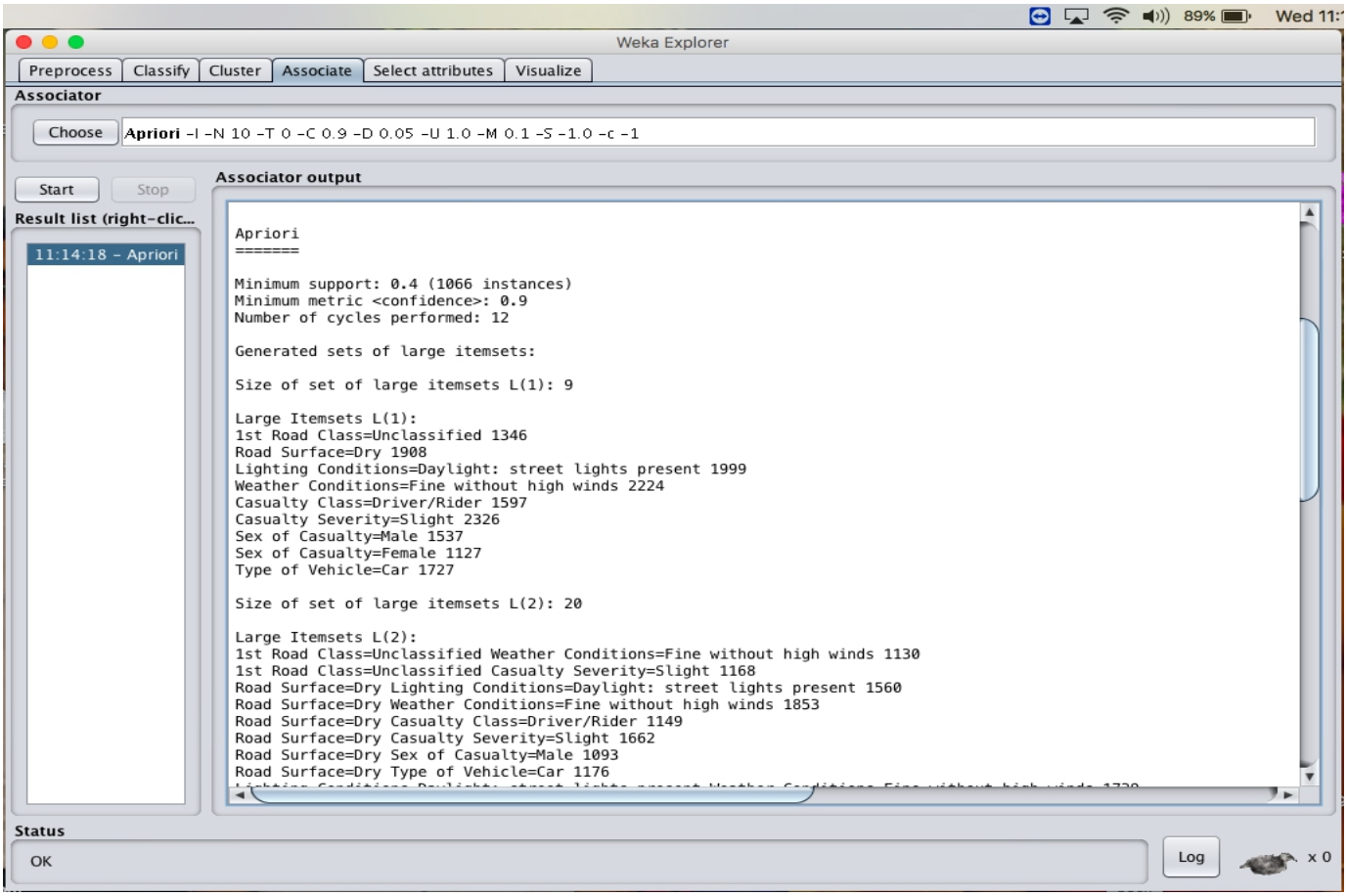
Consider a dataset of 2015.csv file of which it contains the attributes are Reference Number, Grid ref: Easting, Grid Ref: Northing, Number of vehicles, Accident date, Time(24 hr), 1st Road class, Road Surface, Lighting conditions, Weather conditions, casuality class, Sex of casuality, Age of casuality, Type of casuality for the performance of the dataset by applying the Apriori algorithm in weka and as well using R-tool.

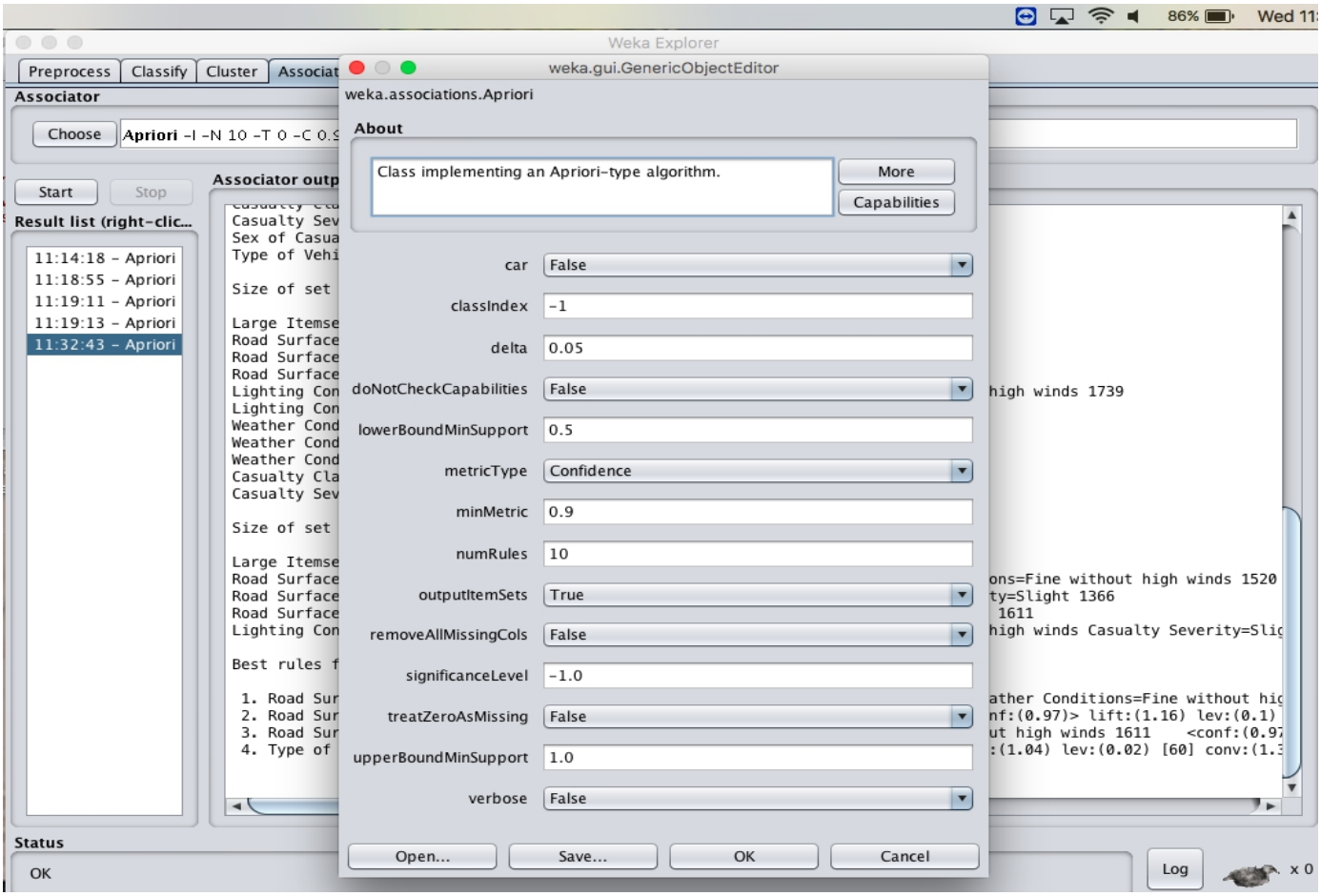
❖ USING WEKA TOOL :

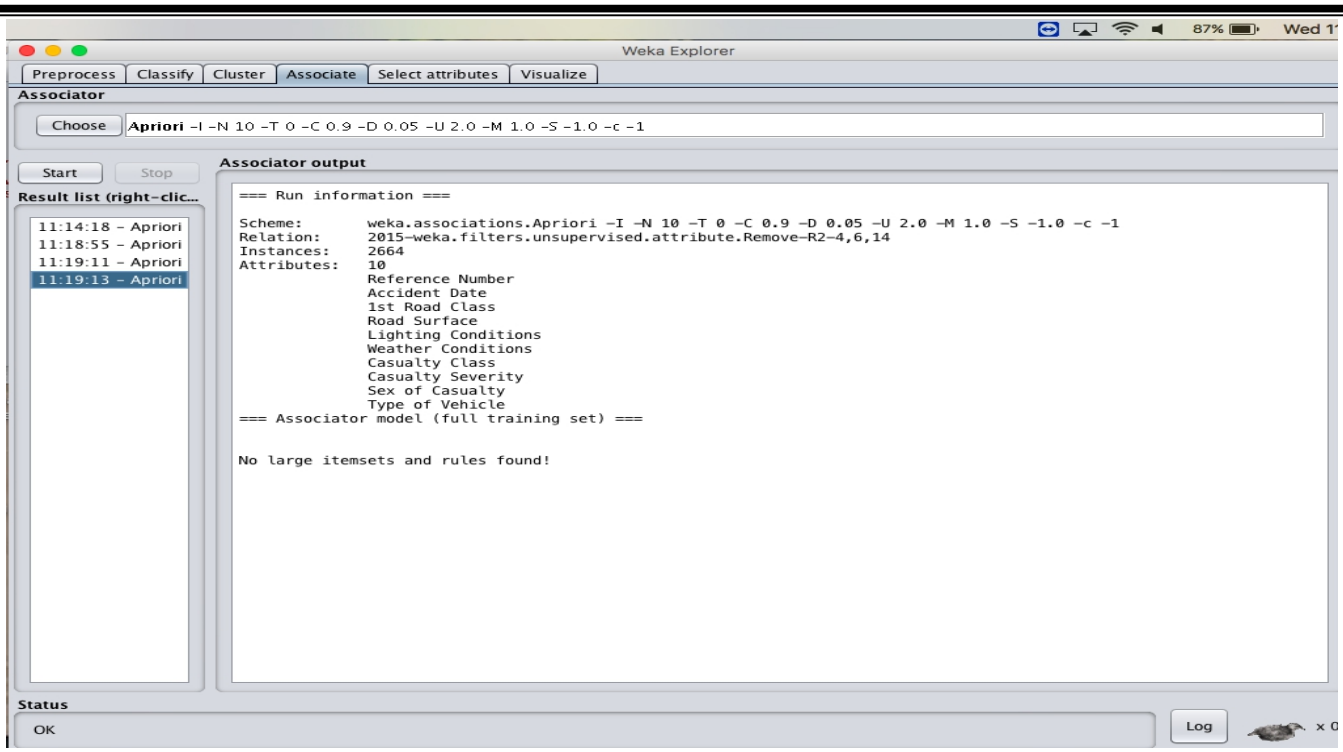
STEPS INVOLVED :

- Choose a set of attributes for clustering and for giving a motivation.
- Choose the dataset and import the dataset into Weka tool.
- Discretize the attributes from numeric to nominal to perform the algorithm.
- Cluster the dataset and choose simple Apriori algorithm.
- Set the Upper bound min_sup and lower bound min_sup values.




Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Associator
Choose
Apriori -I -N 10 -T 0 -C 0.9 -D 0.05 -U 1.0 -M 0.1 -S -1.0 -c -1
Start
Stop
Result list (right-clic...
11:14:18 - Apriori
Associator output
Apriori
=======
Minimum support: 0.4 (1066 instances)
Minimum metric <confidence>: 0.9
Number of cycles performed: 12
Generated sets of large itemsets:
Size of set of large itemsets L(1): 9
Large Itemsets L(1):
1st Road Class=Unclassified 1346
Road Surface=Dry 1908
Lighting Conditions=Daylight: street lights present 1999
Weather Conditions=Fine without high winds 2224
Casualty Class=Driver/Rider 1597
Casualty Severity=Slight 2326
Sex of Casualty=Male 1537
Sex of Casualty=Female 1127
Type of Vehicle=Car 1727
Size of set of large itemsets L(2): 20
Large Itemsets L(2):
1st Road Class=Unclassified Weather Conditions=Fine without high winds 1130
1st Road Class=Unclassified Casualty Severity=Slight 1168
Road Surface=Dry Lighting Conditions=Daylight: street lights present 1560
Road Surface=Dry Weather Conditions=Fine without high winds 1853
Road Surface=Dry Casualty Class=Driver/Rider 1149
Road Surface=Dry Casualty Severity=Slight 1662
Road Surface=Dry Sex of Casualty=Male 1093
Road Surface=Dry Type of Vehicle=Car 1176
Status
OK
Log
x 0


Weka Explorer
weka.gui.GenericObjectEditor
weka.associations.Apriori
About
Class implementing an Apriori-type algorithm.
More
Capabilities
car
False
classIndex
-1
delta
0.05
doNotCheckCapabilities
False
lowerBoundMinSupport
0.5
metricType
Confidence
minMetric
0.9
numRules
10
outputItemSets
True
removeAllMissingCols
False
significanceLevel
-1.0
treatZeroAsMissing
False
upperBoundMinSupport
1.0
verbose
False
Open...
Save...
OK
Cancel
Result list (right-clic...
11:14:18 - Apriori
11:18:55 - Apriori
11:19:11 - Apriori
11:19:13 - Apriori
11:32:43 - Apriori
Status
OK
Log
x 0



❖ USING R-TOOL :

STEPS INVOLVED :

- Choose the dataset and import the dataset into the R-tool.
- View the dataset and start inserting queries for the Apriori algorithm.

QUERIES :

- Data=read.csv("C:/users/prasanthiemani/Desktop/2015.csv")
- View(Data)
- a = apriori(data, parameter = list(sup=0.3, conf=0.9)

```
> a=apriori(data,parameter = list(sup=0.3,conf=0.9))
Apriori

Parameter specification:
 confidence minval smax arem  aval originalSupport maxtime support minlen maxlen
        0.9    0.1    1 none FALSE            TRUE       5     0.3      1     10
 target   ext
  rules FALSE

Algorithmic control:
 filter tree heap memopt load sort verbose
    0.1 TRUE TRUE  FALSE TRUE    2    TRUE

Absolute minimum support count: 799

set item appearances ...[0 item(s)] done [0.00s].
set transactions ...[2393 item(s), 2664 transaction(s)] done [0.00s].
sorting and recoding items ... [10 item(s)] done [0.00s].
creating transaction tree ... done [0.02s].
checking subsets of size 1 2 3 4 5 done [0.00s].
writing ... [30 rule(s)] done [0.00s].
creating S4 object  ... done [0.00s].
>
```

RESULT :

Thus, the Apriori algorithm analyzing using both the weka tool and R- tool has been successfully completed. In case of weka tool, the change in upper bound and lower bound values lead to the increase and decrease of number of itemsets and rules . In case of R-tool, there is an increase in absolute minimum support count value.

EX.No: 10
Date :

FREQUENT PATTERN MINING USING FP GROWTH THROUGH WEKA TOOL

PROBLEM STATEMENT :

Run the FP growth algorithm, and explore the association rules by changing the following parameters:

a) Upper bound min_sup
b) Lower bound min_sup
c) Metric type

DESCRIPTION :

Consider a dataset of 2015.csv file of which it contains the attributes are Reference Number, Grid ref: Easting, Grid Ref: Northing, Number of vehicles, Accident date, Time(24 hr), 1st Road class, Road Surface, Lighting conditions, Weather conditions, casuality class, Sex of casuality, Age of casuality, Type of casuality for the performance of the dataset by applying the FP algorithm in weka tool.

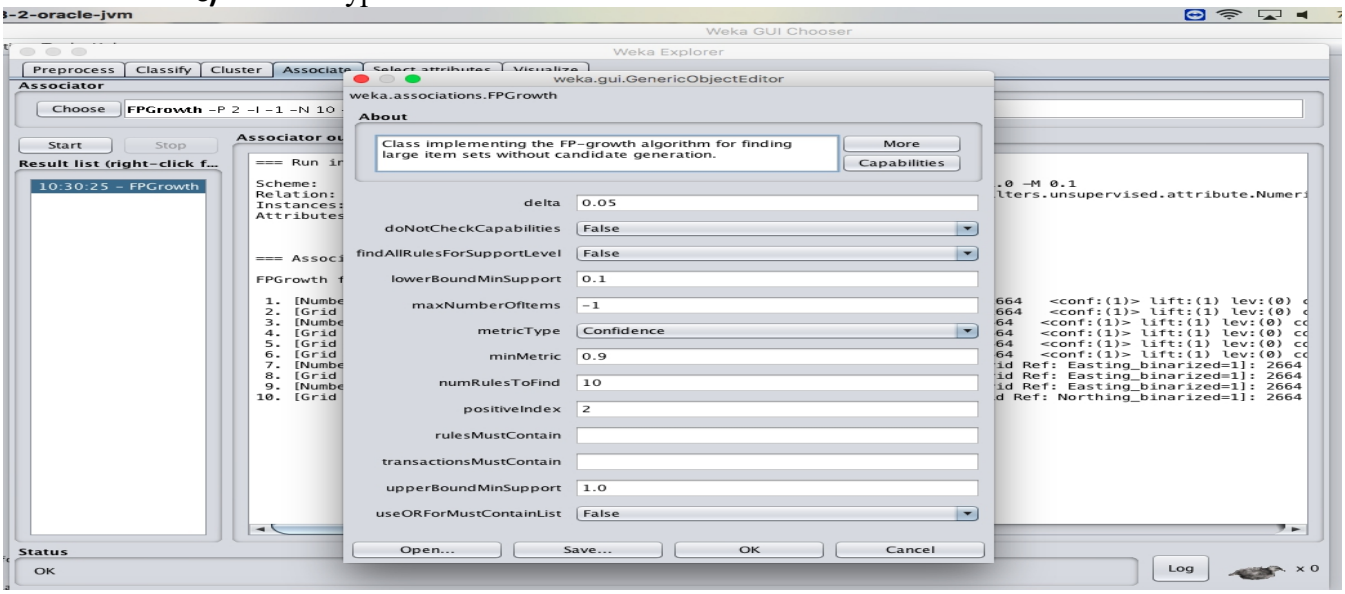
❖ USING WEKA TOOL :

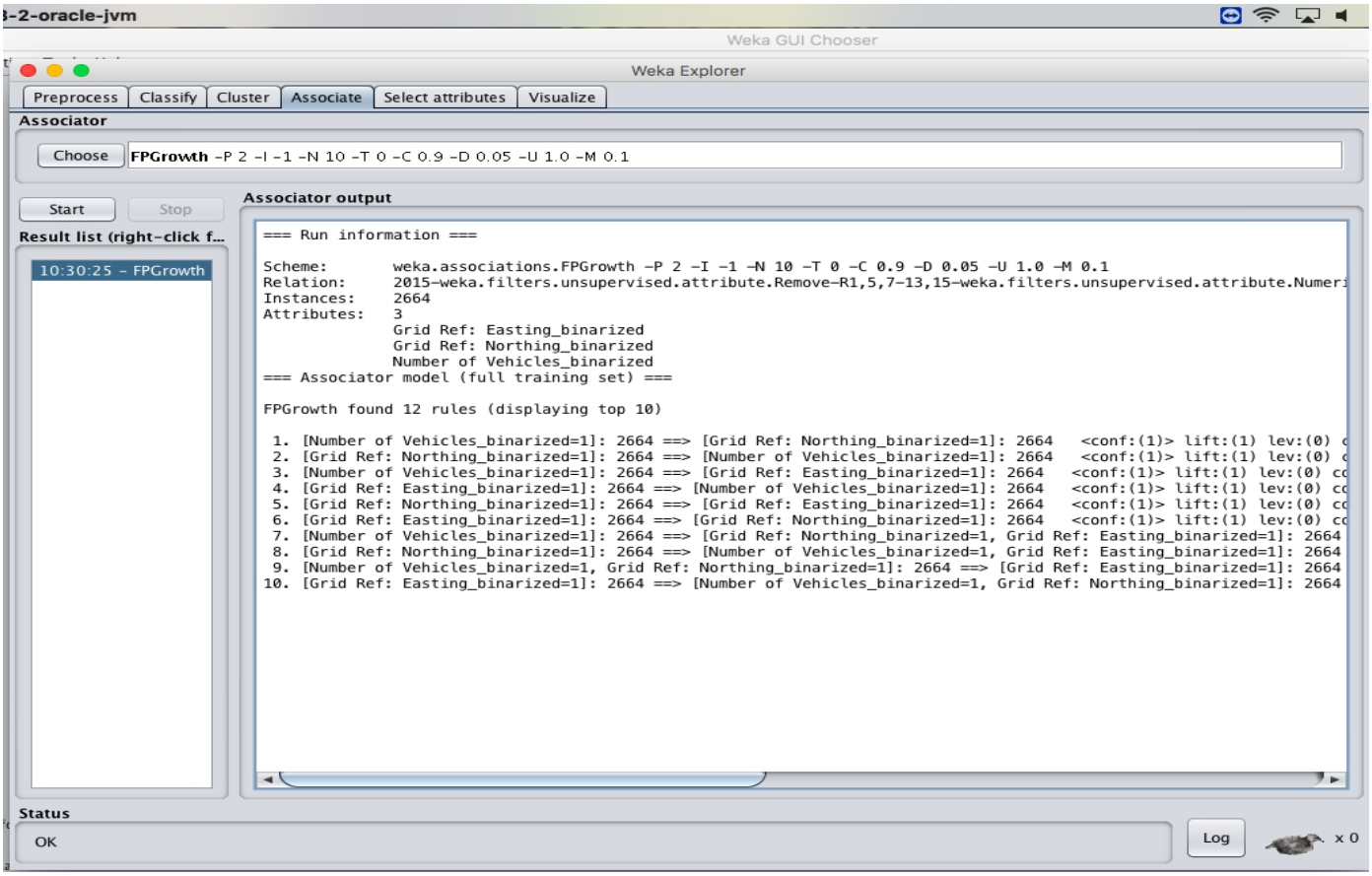
STEPS INVOLVED :

- Choose a set of attributes for clustering and for giving a motivation.
- Choose the dataset and import the dataset into Weka tool.
- Discretize the attributes from all data types to nominal to perform the algorithm.
- Associate the attributes with the FP growth algorithm.
- Set the Upper bound min_sup and lower bound min_sup values.

OBSERVATIONS :

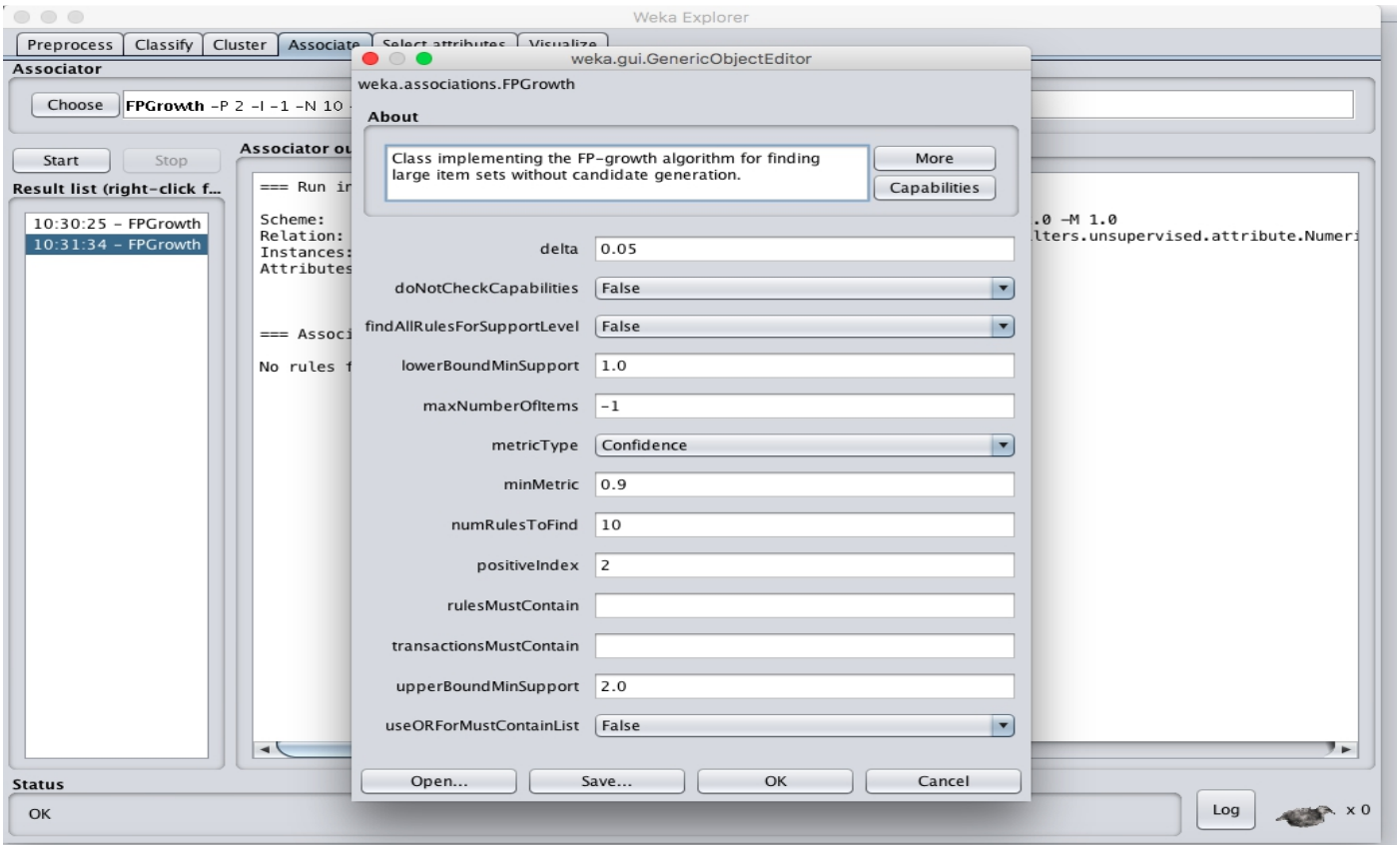
1) When the association rules are of values:
   a) Upper bound min_sup = 1.0
   b) Lower bound min_sup = 0.1
   c) Metric type = confidence.

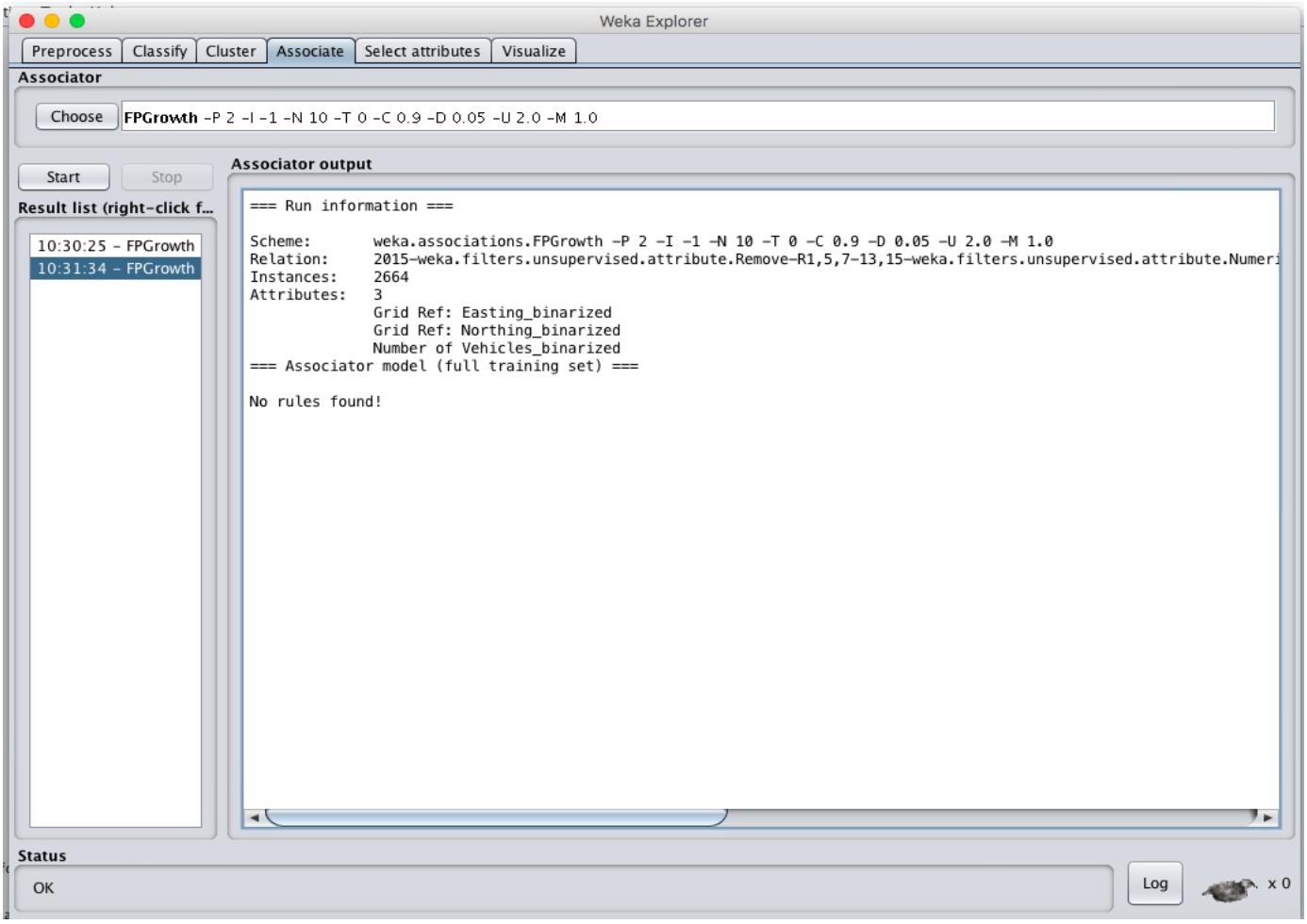




2) When the association rules are of values:

a) Upper bound min_sup = 2.0
b) Lower bound min_sup = 1.0
c) Metric type = confidence.





Through the comparision of both the cases 1 and 2, the rules will be changed according to the values of upper bound minimum support and lower bound minimum support. If the upperbound minimum support and lower bound minimum support are high then the number of rules will be very less and its vice-versa in the case of less upper bound and lower bound minimum support values.

RESULT :

Thus, the analysis of FP growth algorithm using weka tool has been successfully completed. In case of changing the upper bound and lower bound values there is a change in the number of rules that are found.

EX.No: 11
Date :

PREDICTION OF CATEGORICAL DATA USING DECISION TREE ALGORTIHM THROUGH WEKA

PROBLEM STATEMENT :

Actual historical credit data is not always easy to come by because of confidentiality rules. Here is one such dataset, consisting of 1000 actual cases collected in Germany. credit dataset (original) Excel spreadsheet version of the German credit data (Down load from web). In spite of the fact that the data is German, you should probably make use of it for this assignment. (Unless you really can consult a real loan officer !)

1) What attributes do you think might be crucial in making the credit assessement ? Come up with some simple rules in plain English using your selected attributes.
2) One type of model that you can create is a Decision Tree - train a Decision Tree using the complete dataset as the training data. Report the model obtained after training.
3) Suppose you use your above model trained on the complete dataset, and classify credit good/bad for each of the examples in the dataset. What % of examples can you classify correctly ? (This is also called testing on the training set) Why do you think you cannot get 100 % training accuracy ?
4) Is testing on the training set as you did above a good idea ? Why or Why not ?
5) One approach for solving the problem encountered in the previous question is using cross - validation ? Describe what is cross-validation briefly. Train a Decistion Tree again using cross-validation and report your results. Does your accuracy increase/decrease ? Why ?

DESCRIPTION :

1. **What attributes do you think might be crucial in making the credit assessement ? Come up with some simple rules in plain English using your selected attributes.**

The attributes that might be crucial in making the credit assessment are :

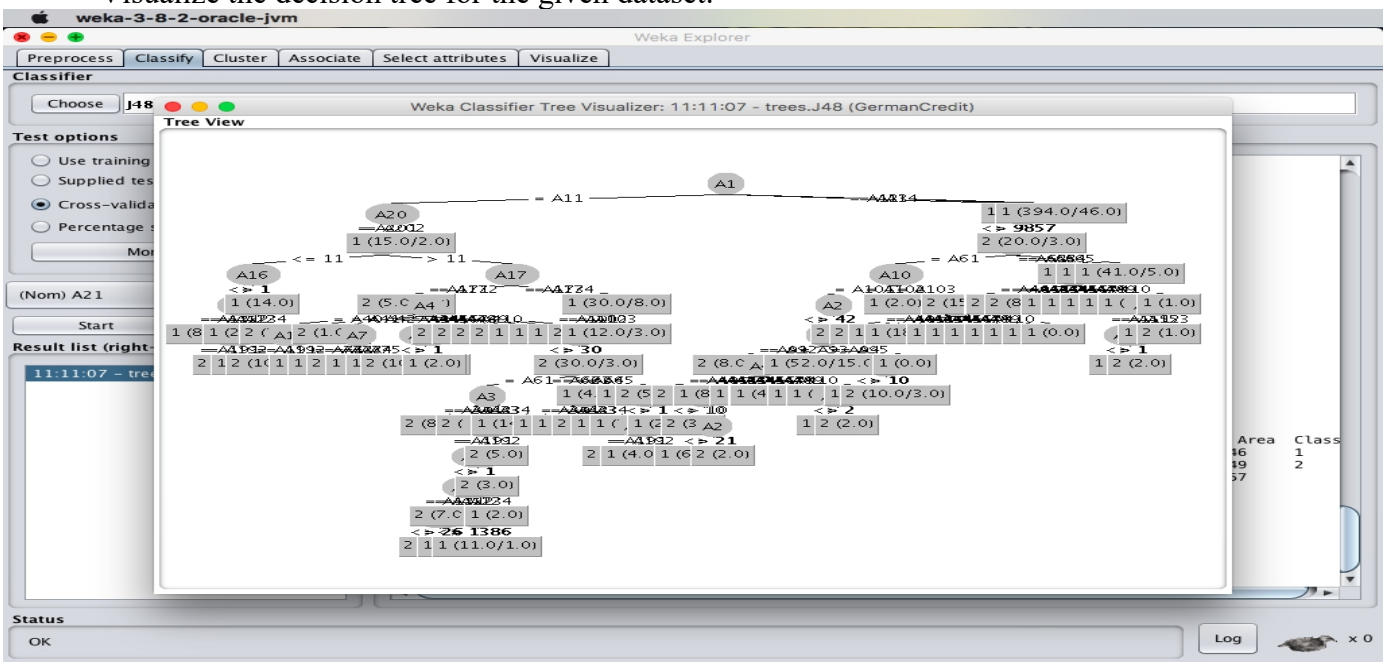
- Numerical attributes
- Nominal attributes

➢ Nominal and numeric attributes are the capabilities of the given dataset.

2. **One type of model that you can create is a Decision Tree - train a Decision Tree using the complete dataset as the training data. Report the model obtained after training.**

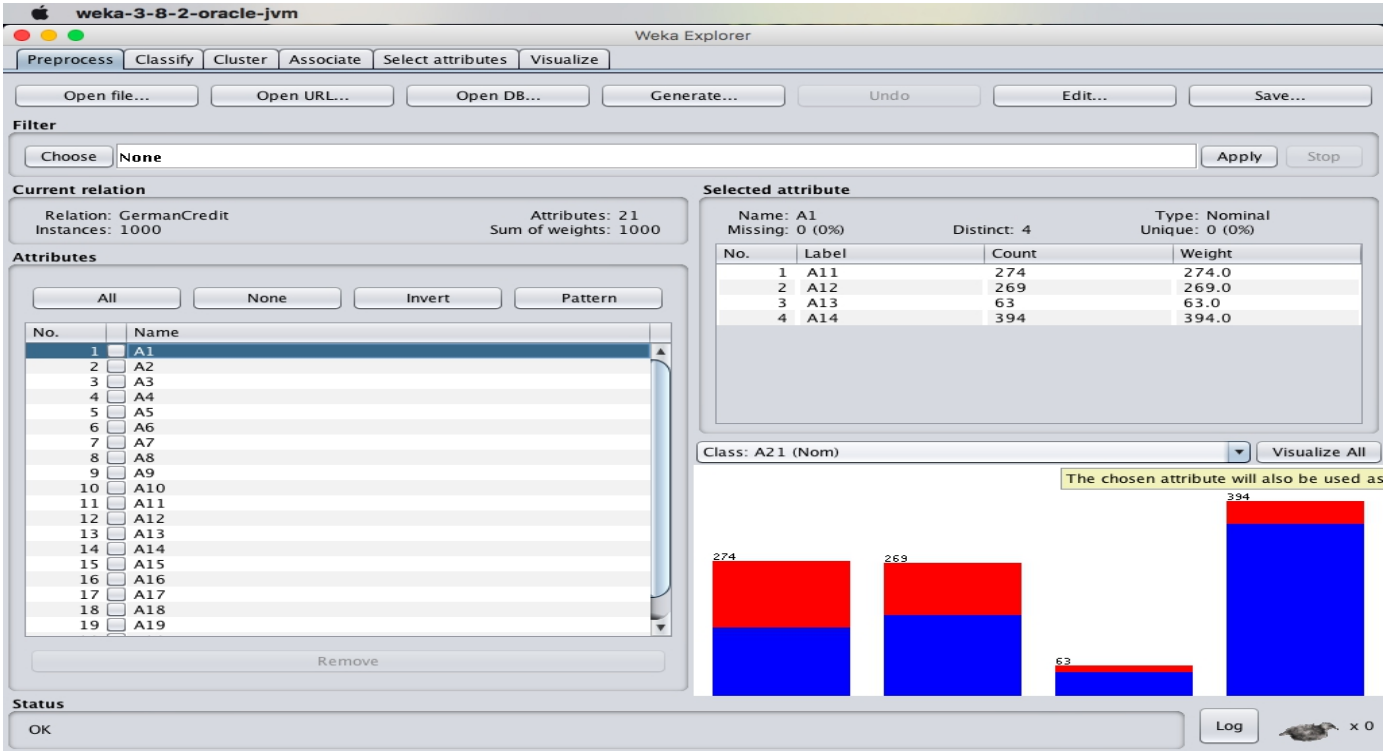
❖ **Decision Tree :**

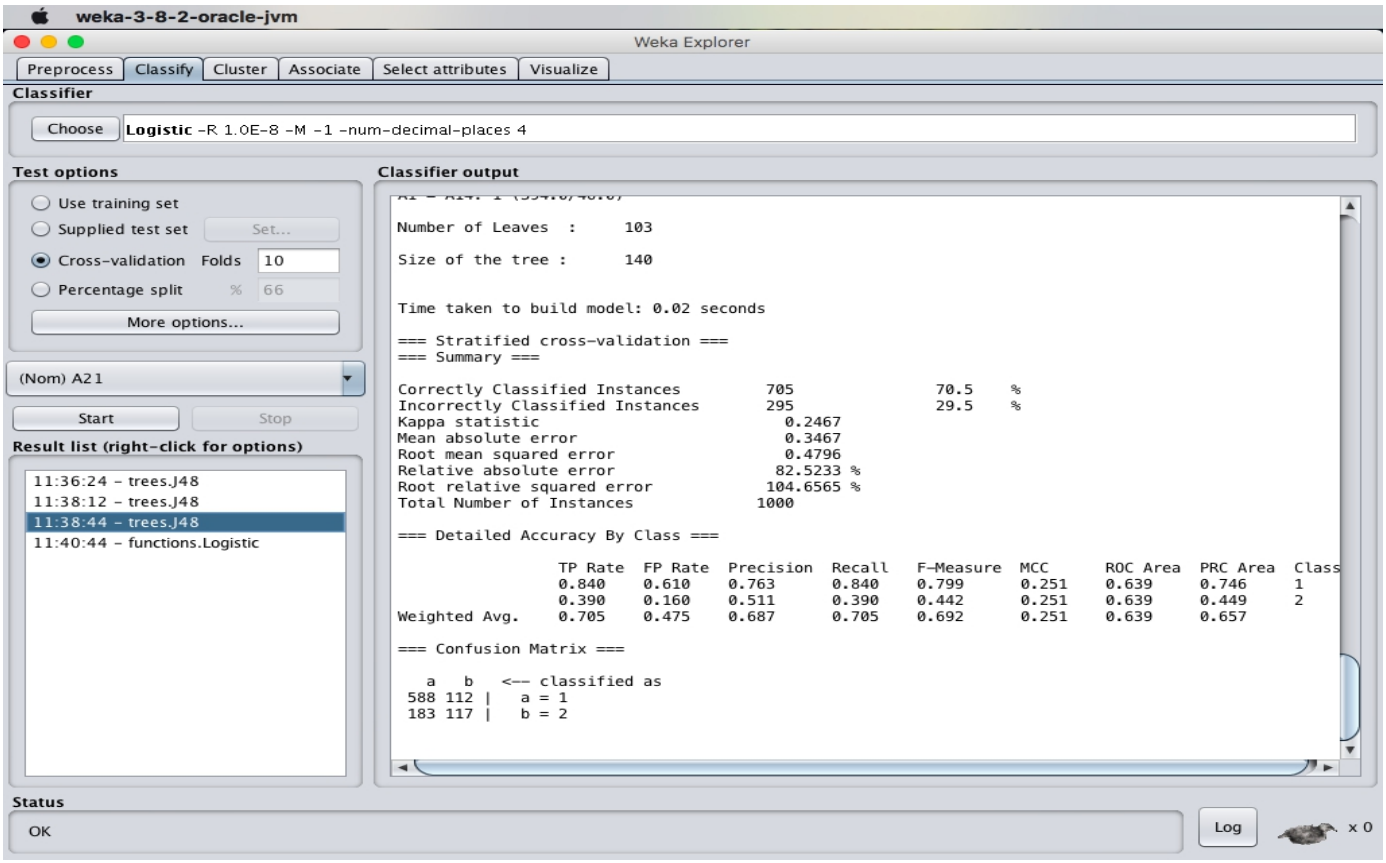
Visualize the decision tree for the given dataset.



3. **Suppose you use your above model trained on the complete dataset, and classify credit good/bad for each of the examples in the dataset. What % of examples can you classify correctly ? (This is also called testing on the training set) Why do you think you cannot get 100 % training accuracy ?**

Usually training dataset will be evaluated by the test dataset. This can be evaluated and analyzed by the instances of the dataset and error like Mean absolute error, Root mean square error, Relative absolute error, Root relative squared error.





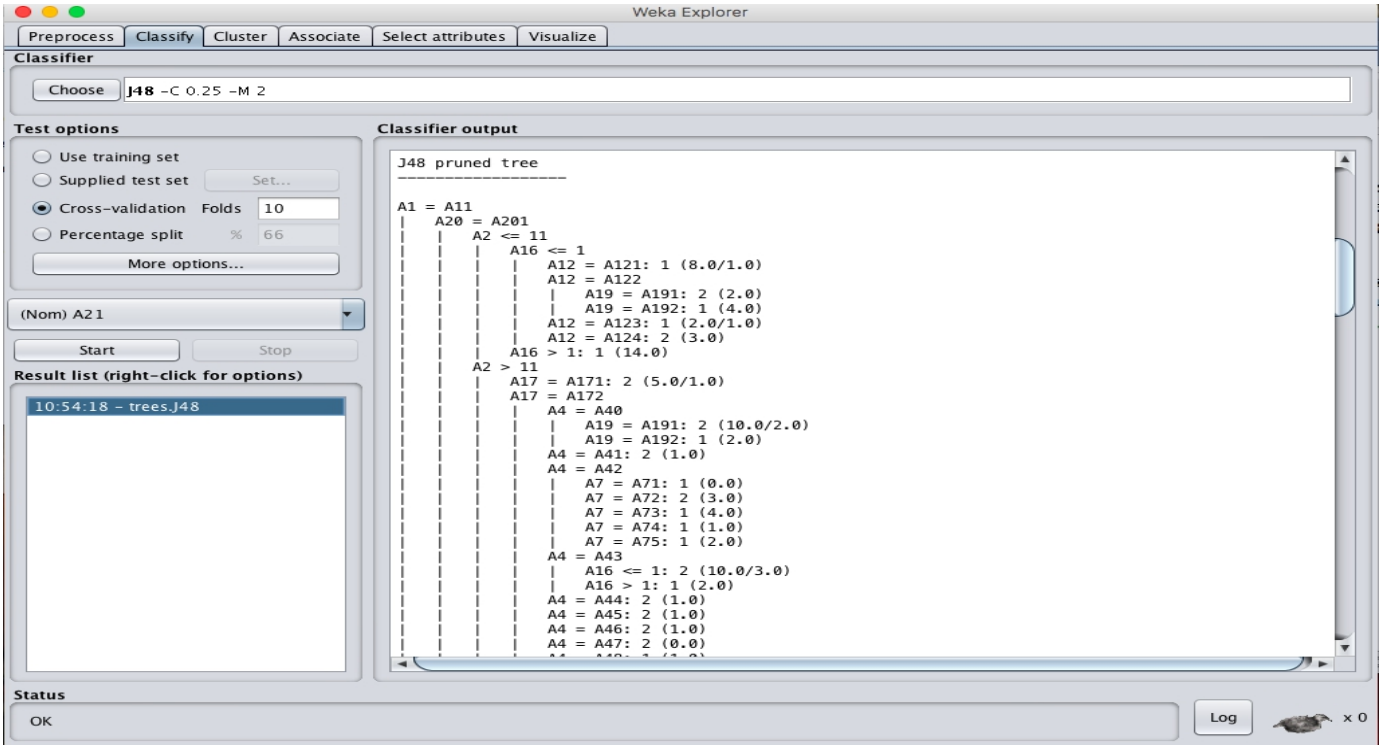
4. Is testing on the training set as you did above a good idea ? Why or Why not ?

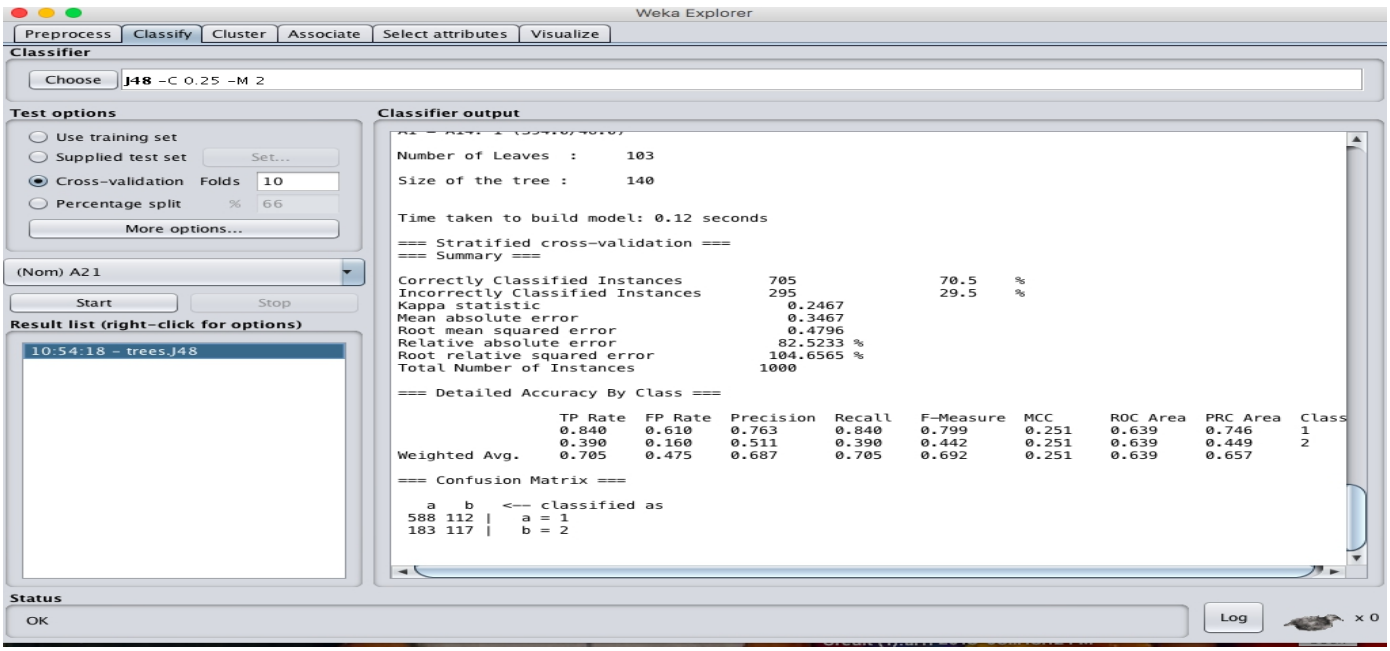
Testing on the training dataset is a good idea where it helps to increase the accuracy of your model by decreasing its complexity. In this case of the dataset, you can prune tree after training. This will decrease the amount of specification in the specific training dataset and increase generalisation on unseen data.

5. One approach for solving the problem encountered in the previous question is using cross - validation ? Describe what is cross-validation briefly. Train a Decistion Tree again using cross-validation and report your results. Does your accuracy increase/decrease ? Why ?

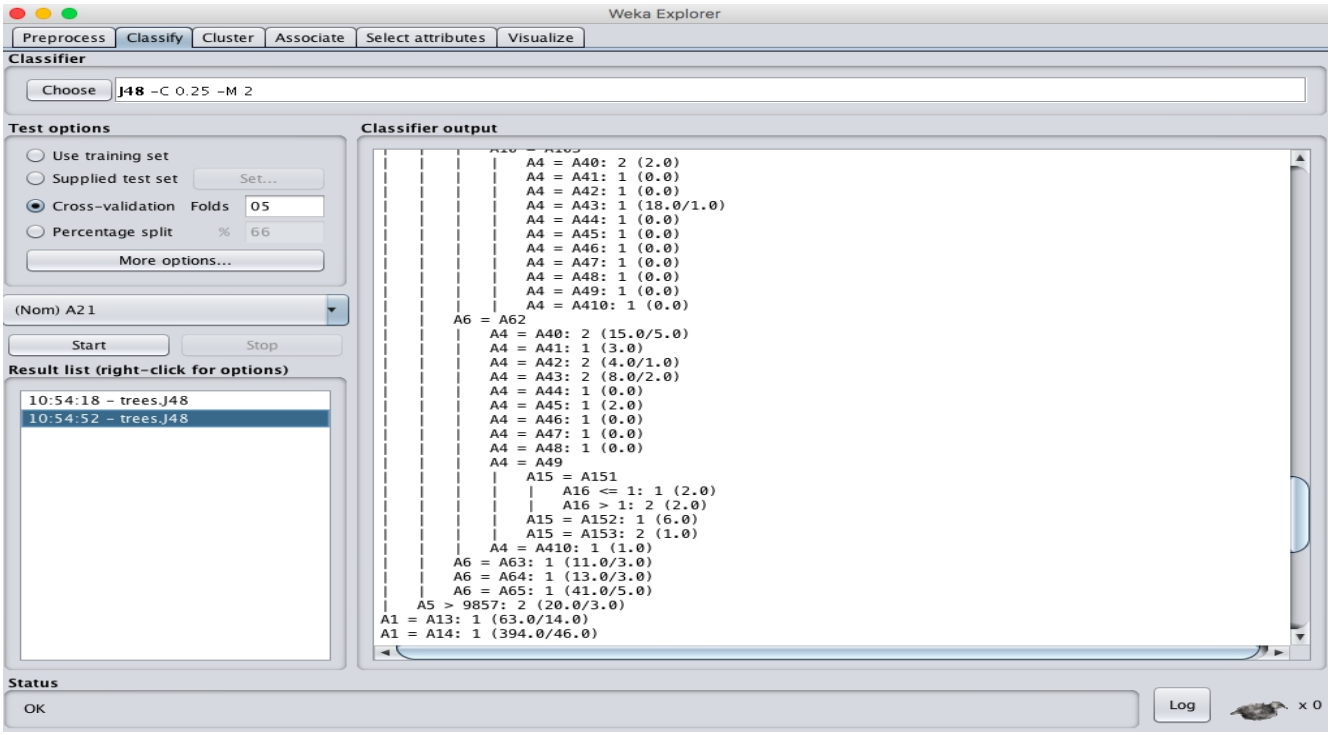
➢ **CROSS VALIDATION ANALYSIS :**

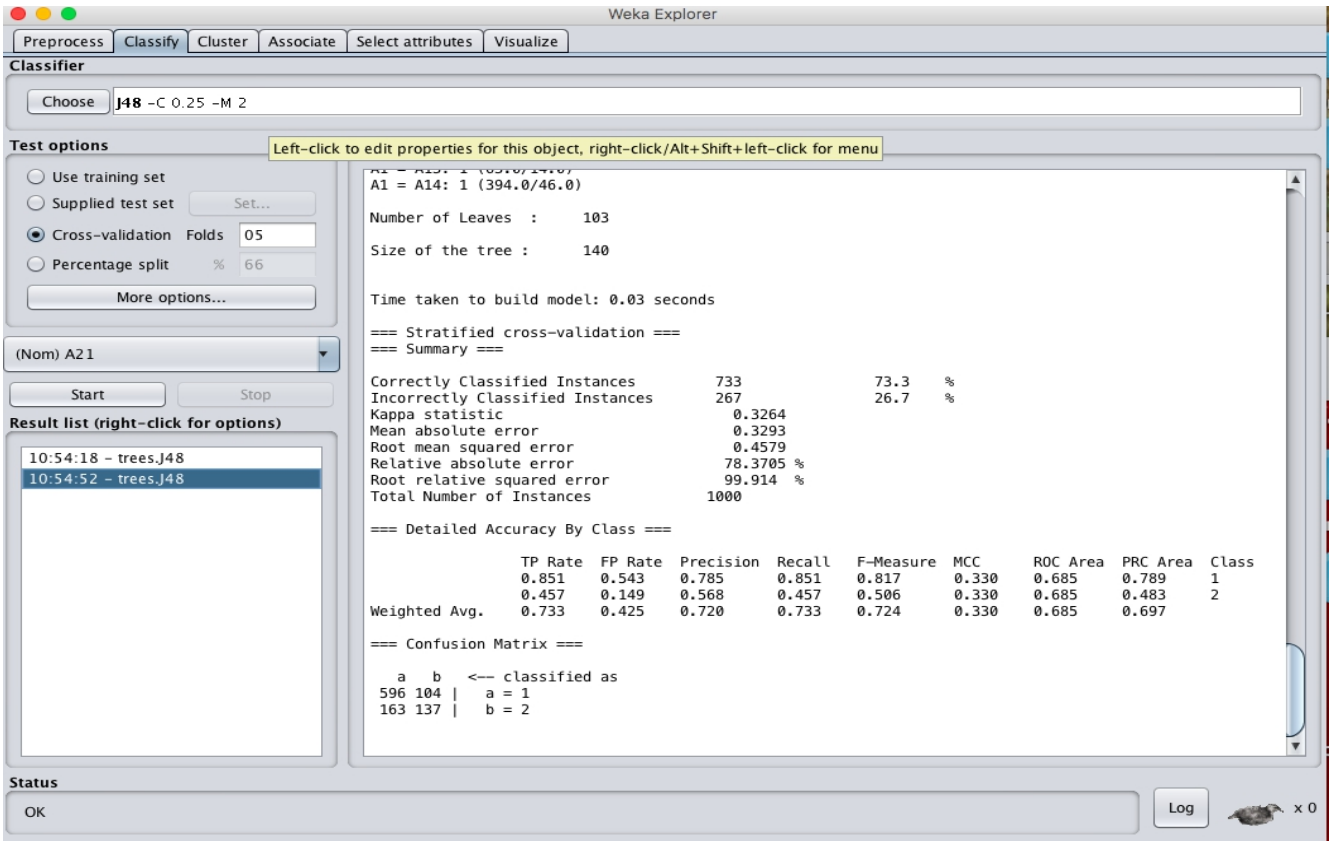
- **When cross validation folds are 10 :**





- **When cross validation folds are : 05 :-**





The change in the cross validation folds leads to the change in the stratified cross validation summary which contains the correctly classified instances and incorrectly classified instances.

RESULT :

Thus, the observations and evaluations done on the german_credit dataset are analyzed. The decision tree has been successfully visualized. Various evaluations and comparisons done through the cross validation folds change. Which lead to the change of values in confusion matrix.

EX.No: 12
Date :

PREDICTION OF CATEGORICAL DATA USING SMO ALGORTIHM THROUGH WEKA

PROBLEM STATEMENT :

Use the german credit dataset download from UCI repository and analyze the given task.

1. Do you really need to input so many attributes to get good results?
2. Compare the results obtained by decision tree and SMO ?
3. Set the cost sensitive evaluation and compare the obtained results.
4. What is the significance of the following parameters :
   a) Mean Absolute Error
   b) Root Mean Square Error
   c) Relative Absolute Error
   d) Total Number of Instances

DESCRIPTION :

Consider the german credit dataset which can be downloaded from the UCI repository.

ANALYSIS :

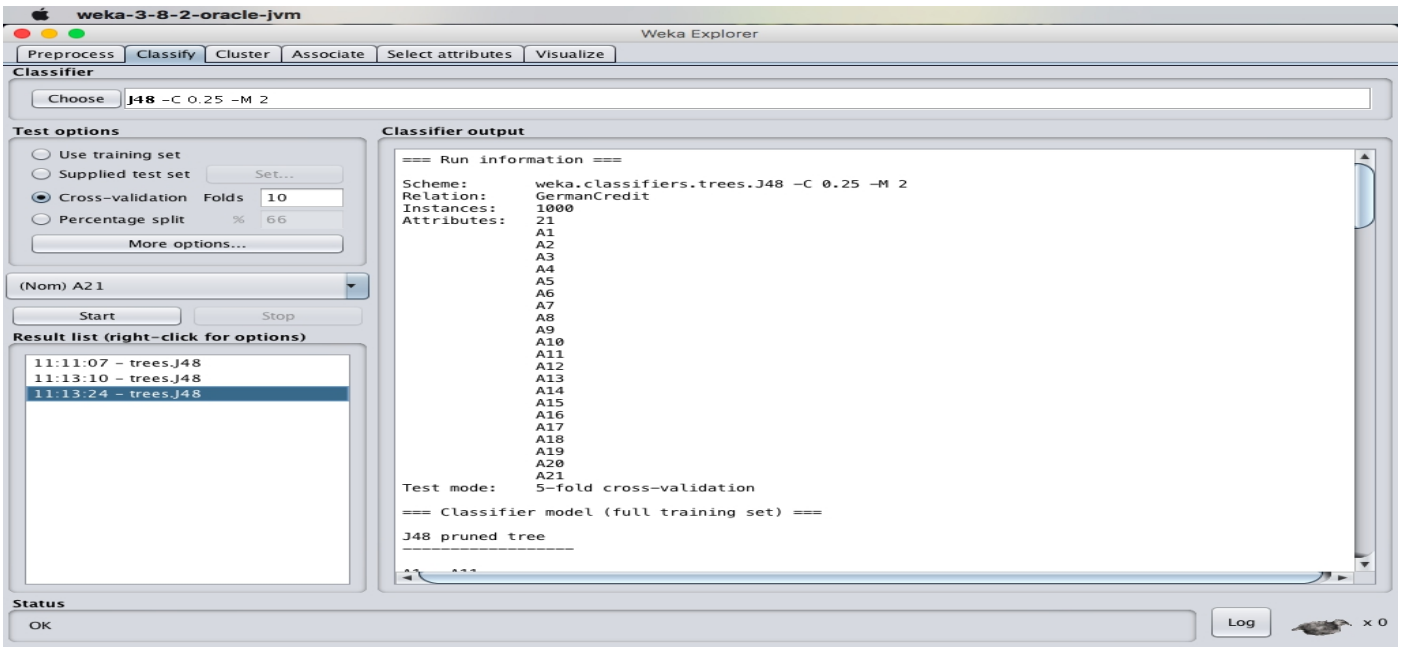
1. **Do you really need to input so many attributes to get good results?**

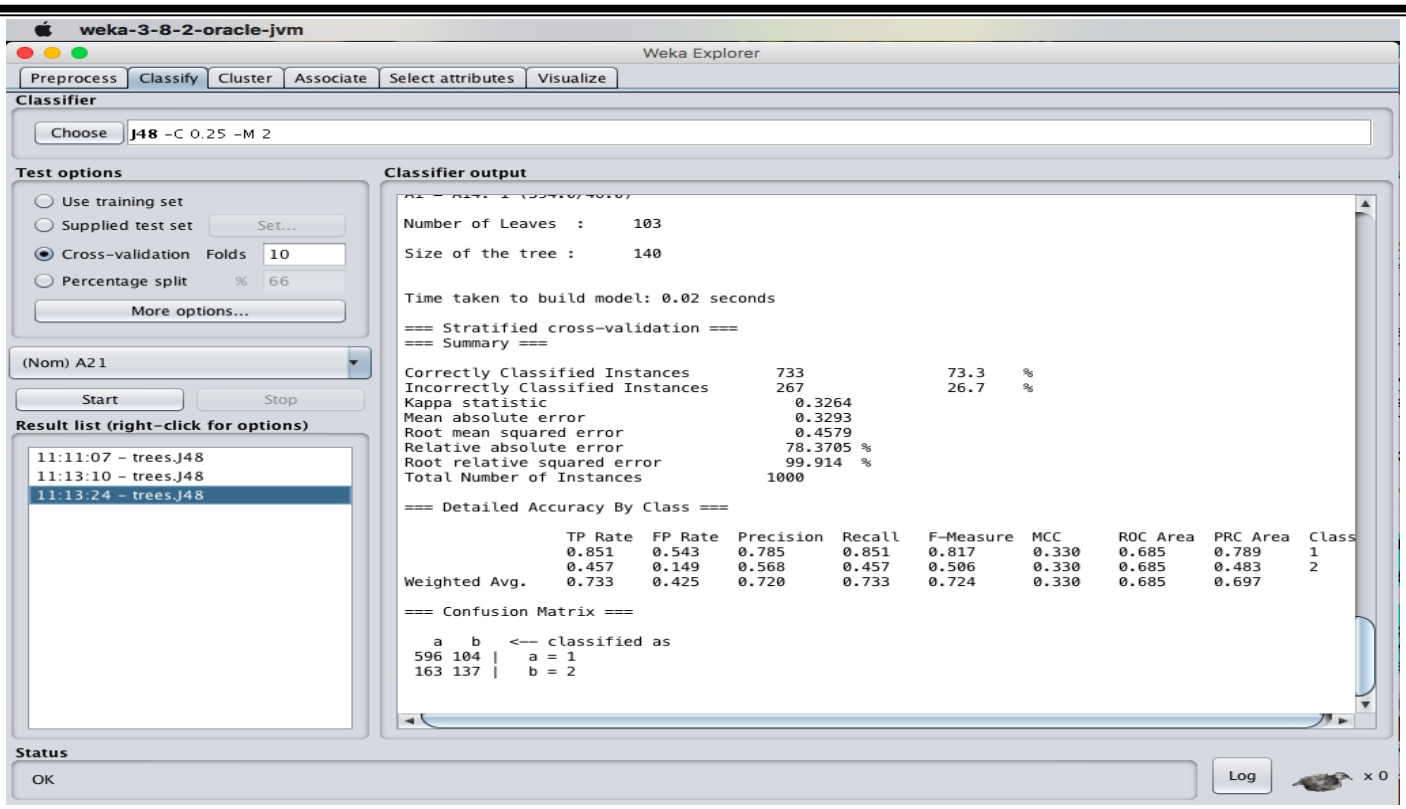
Yes, Having many attributes as the input leads to the good results regarding the dataset. Having more attributes leads to have more and different types of the evaluations.

2. **Compare the results obtained by decision tree and SMO ?**

❖ **DECISION TREE :**

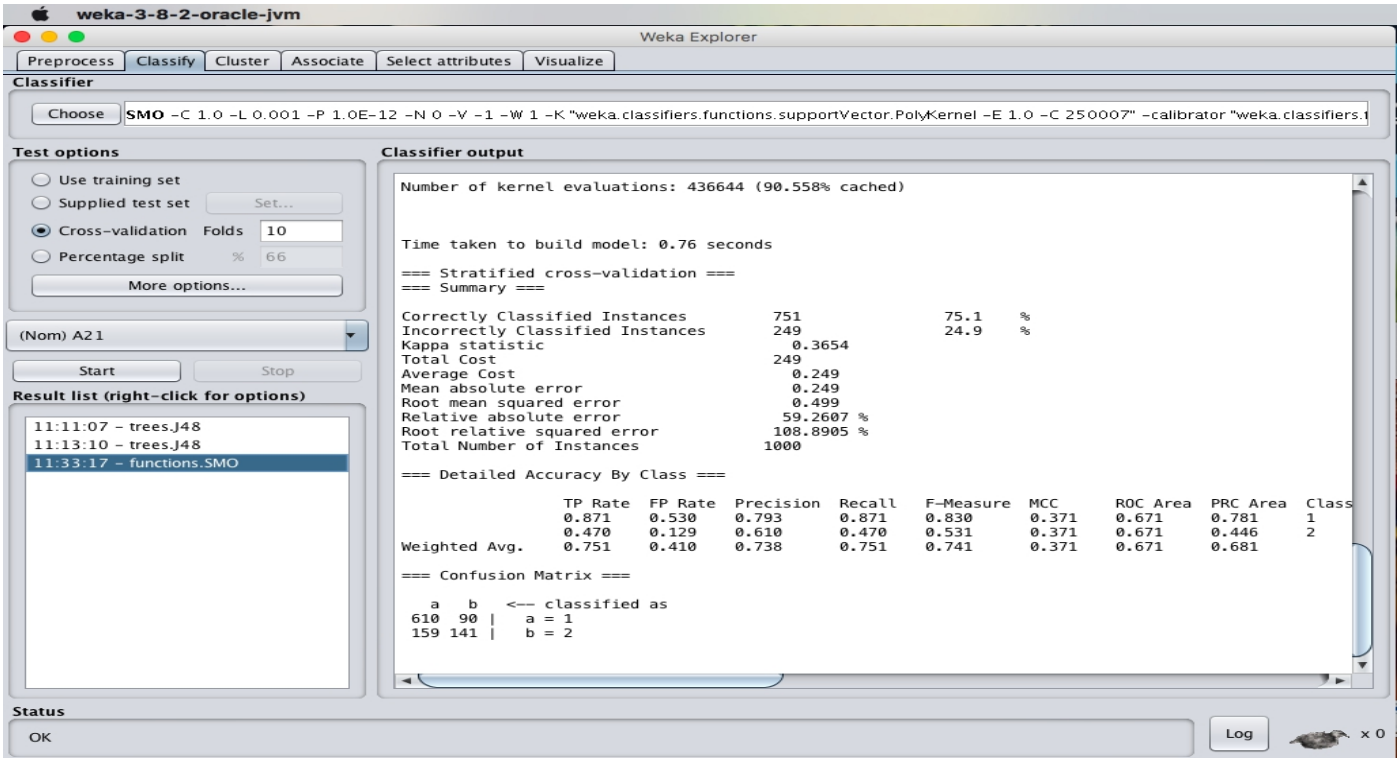
A tree has many analogies in real life, and turns out that it has influenced a wide area of machine learning, covering both classification and regression. In decision analysis, a decision tree can be used to visually and explicitly represent decisions and decision making. As the name goes, it uses a tree-like model of decisions. Though a commonly used tool in data mining for deriving a strategy to reach a particular goal, its also widely used in machine learning, which will be the main focus of this article.





❖ SMO ALGORITHM:

The iterative algorithm Sequential Minimal Optimization (SMO) is used for solving quadratic programming (QP) problems. One example where QP problems are relevant is during the training process of support vector machines (SVM). The SMO algorithm is used to solve in this example a constraint optimization problem. John Platt proposed this algorithm in 1998 and it was successfully used since then. We describe here the basics of the algorithm in the light of big data.



When compared to both the Decision Tree and SMO. SMO is taking more time for building the model. Also there will be a change in instances.

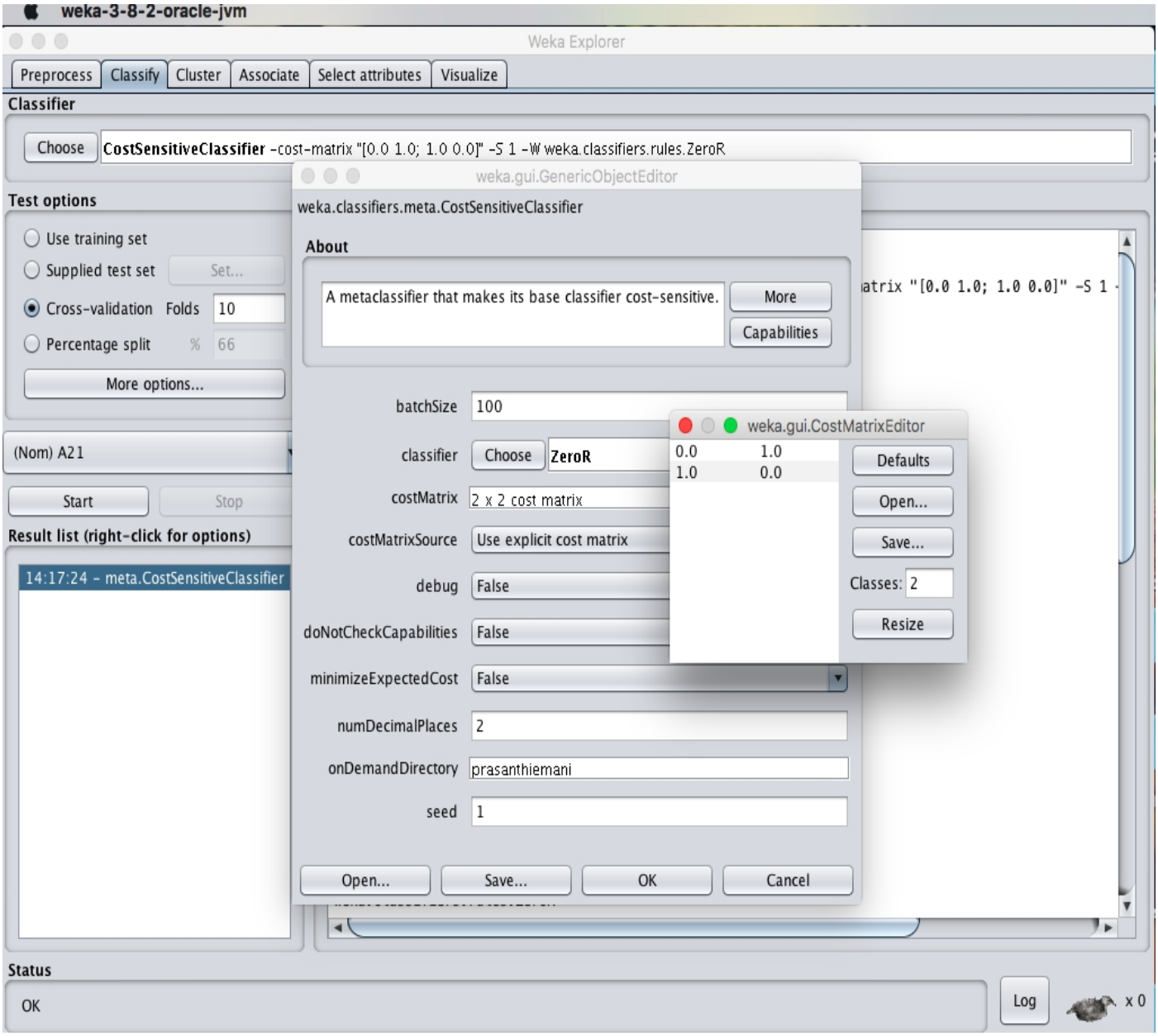
3. Set the cost sensitive evaluation and compare the obtained results.

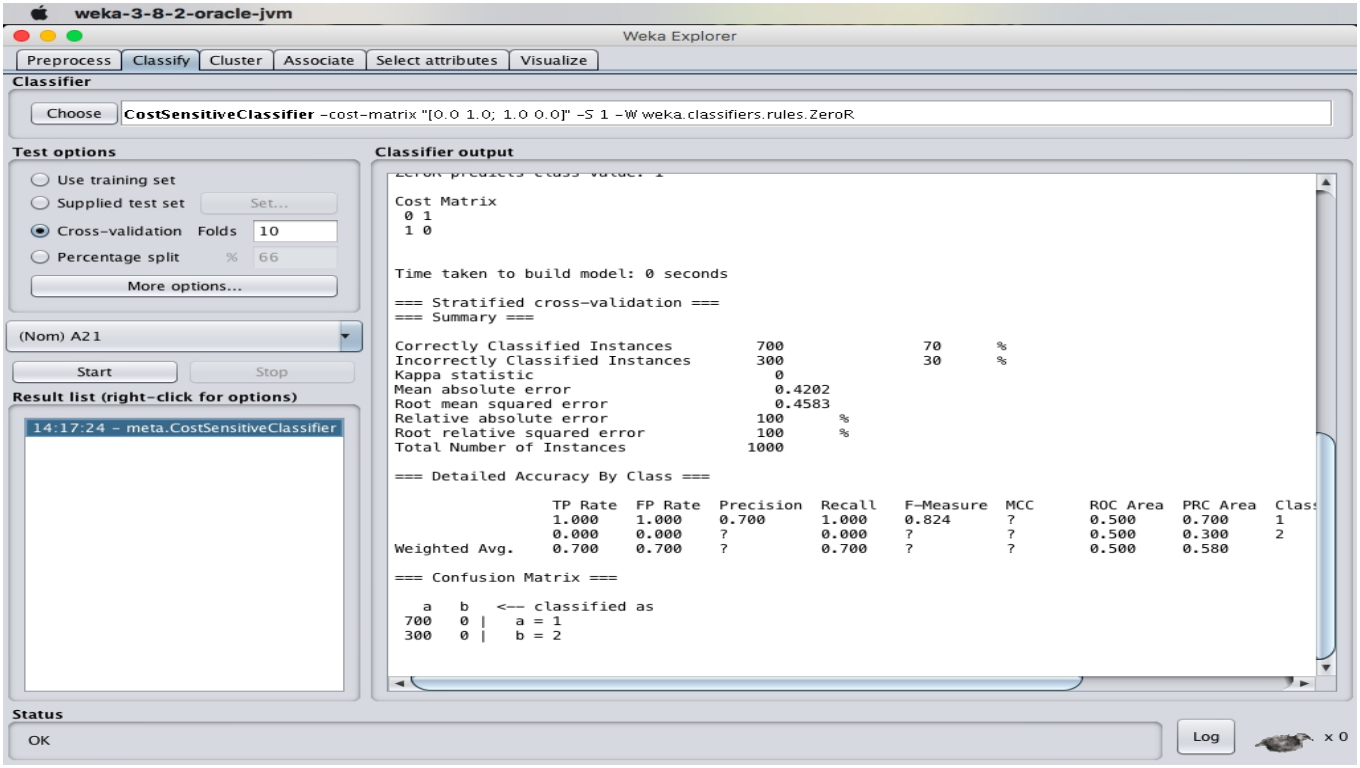
Cost-Sensitive Learning is a type of learning in data mining that takes the misclassification costs (and possibly other types of cost) into consideration. The goal of this type of learning is to minimize the total cost. The key difference between cost-sensitive learning and cost-insensitive learning is that cost-sensitive learning treats the different misclassifications differently. Cost in sensitive learning does not take the misclassification costs into consideration. The goal of this type of learning is to pursue a high accuracy of classifying examples into a set of known classes.

STEPS :

- Classify the dataset with the cost sensitive classifier technique.
- Change the cost matrix to 2*2 matrix and execute.

ANALYSIS :





4. What is the significance of the following parameters :

a) Mean Absolute Error :

Mean Absolute Error (MAE) is similar to the Mean Squared Error, but it uses absolute values instead of squaring. This measure is not as popular as MSE, though its meaning is more intuitive (the "average error").

b) Root Mean Square Error :

Root Mean Square Error (RMSE) is the standard deviation of the residuals (prediction errors). Residuals are a measure of how far from the regression line data points are; RMSE is a measure of how spread out these residuals are. In other words, it tells you how concentrated the data is around the line of best fit. Root mean square error is commonly used in climatology, forecasting, and regression analysis to verify experimental results.

c) **Relative Absolute Error :**

The relative absolute error is very similar to the relative squared error in the sense that it is also relative to a simple predictor, which is just the average of the actual values. In this case, though, the error is just the total absolute error instead of the total squared error. Thus, the relative absolute error takes the total absolute error and normalizes it by dividing by the total absolute error of the simple predictor.

d) Total Number of Instances :

The data present consists of various instances of the class. In the case of german_credit dataset, the total number of instances present in the german credit dataset are 1000 instances.

RESULT :

Thus, the observations and evaluations done on the german_credit dataset are analyzed. The comparison between decision tree and Sequential Minimal Optimization (SMO) has been successfully visualized. In addition to that cost sensitive classifier is been used to analyze few things.

EX.No: 13
Date :

EVALUATING ACCURACY OF THE CLASSIFIERS

PROBLEM STATEMENT :

Compare the confusion matrix generated using weka for the german_credit dataset(download from the UCI repository).

a) Logistic Regression
b) Naïve Bayes Algorithm
c) J48
d) K-Nearest Neighbor
e) SMO Algorithm

DESCRIPTION :

Consider the german credit dataset which can be downloaded from the UCI repository.

ANALYSIS :

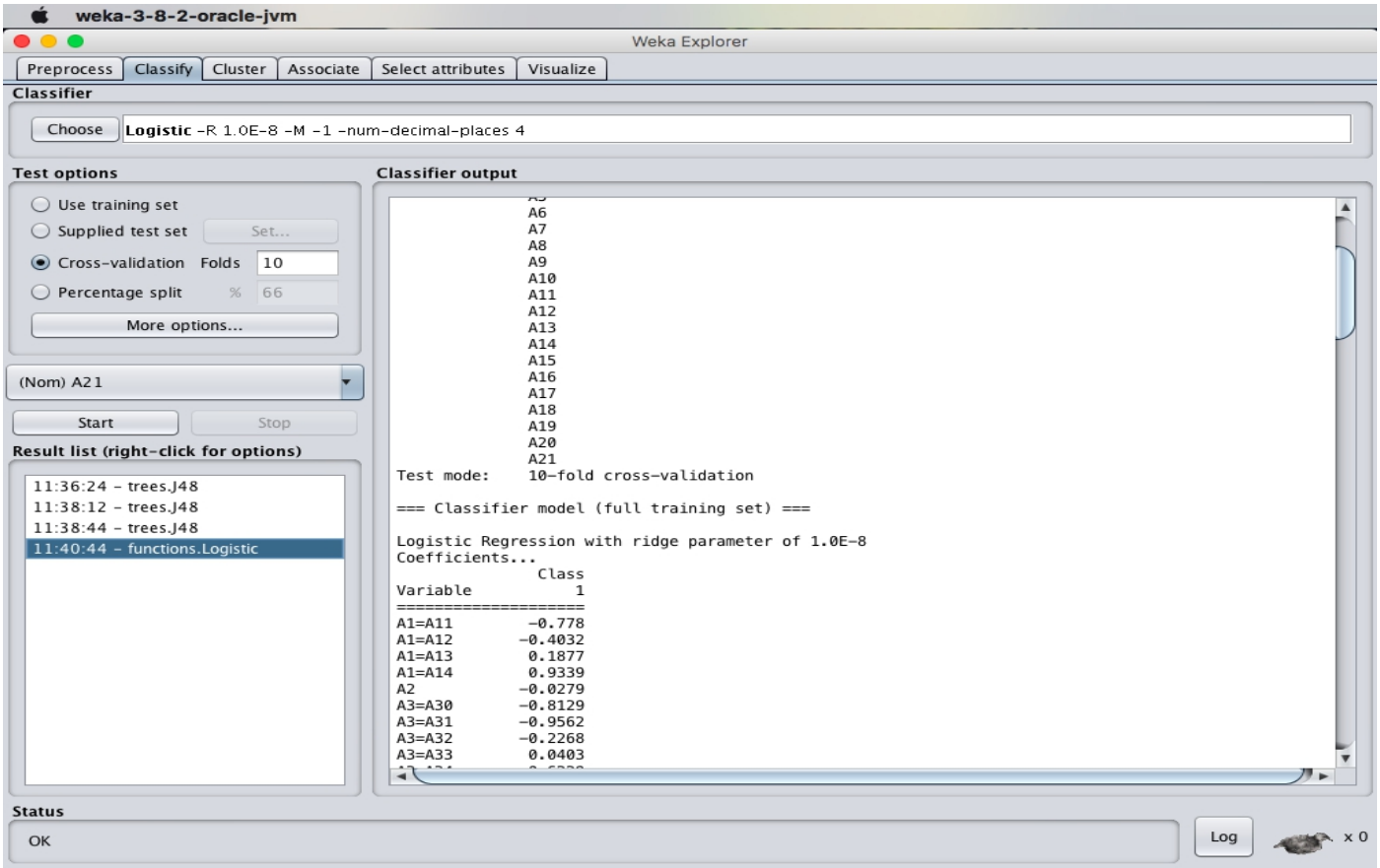
A) Logistic Regression :

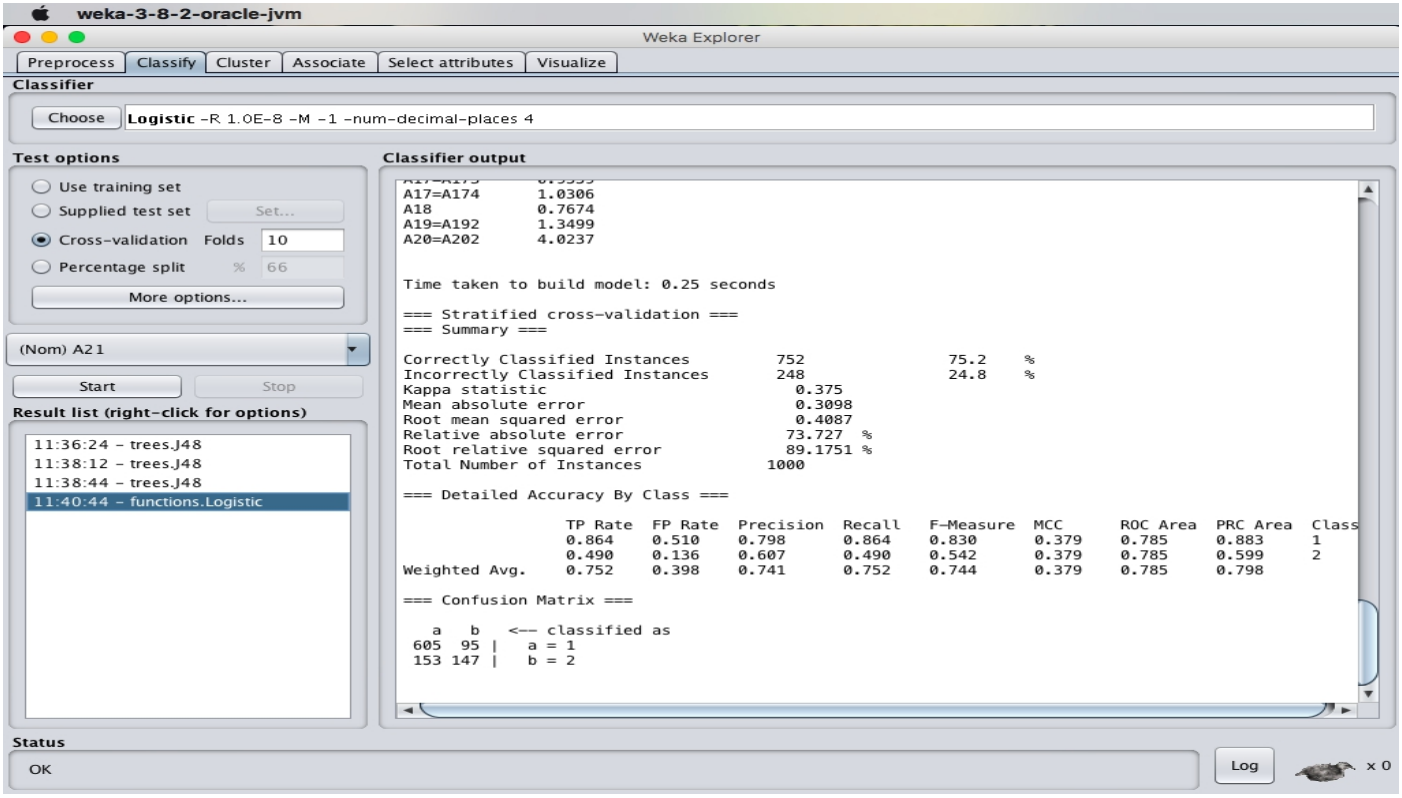
Logistic regression predicts the probability of an outcome that can only have two values (i.e. a dichotomy). The prediction is based on the use of one or several predictors (numerical and categorical).

Steps :

- Load the dataset into the weka tool and preprocess it.
- Apply the classification the logistic regression technique and execute for the result.

Output :





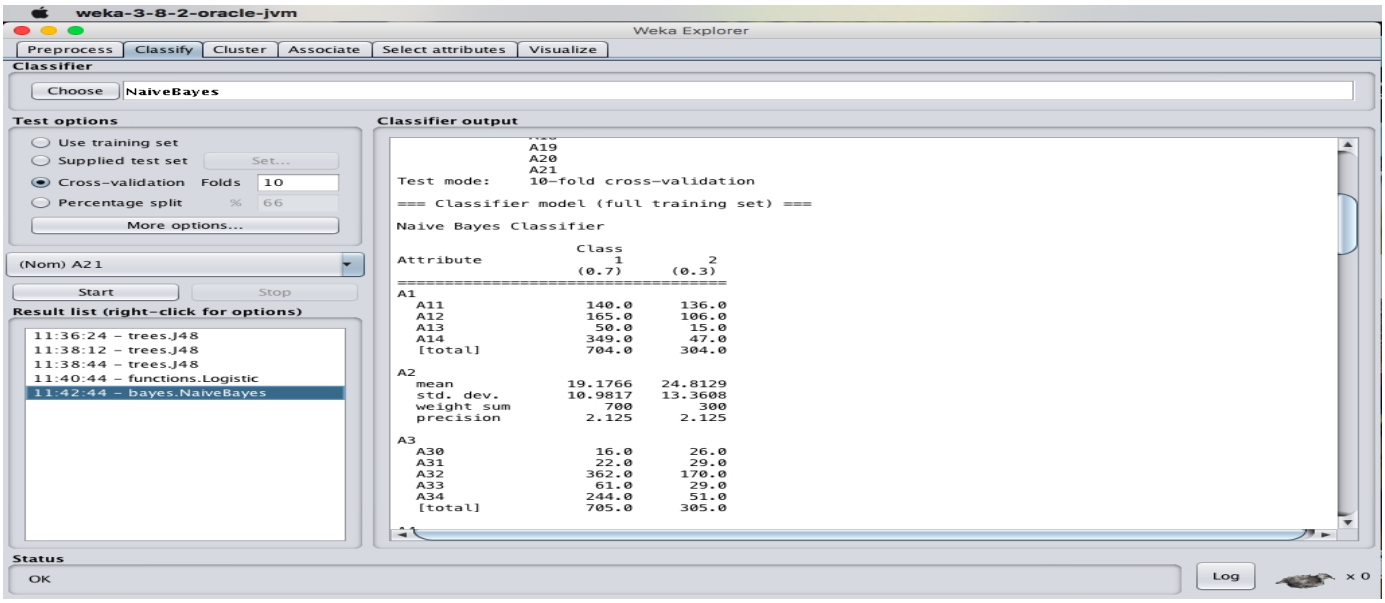
B) Naïve Bayes Algorithm :

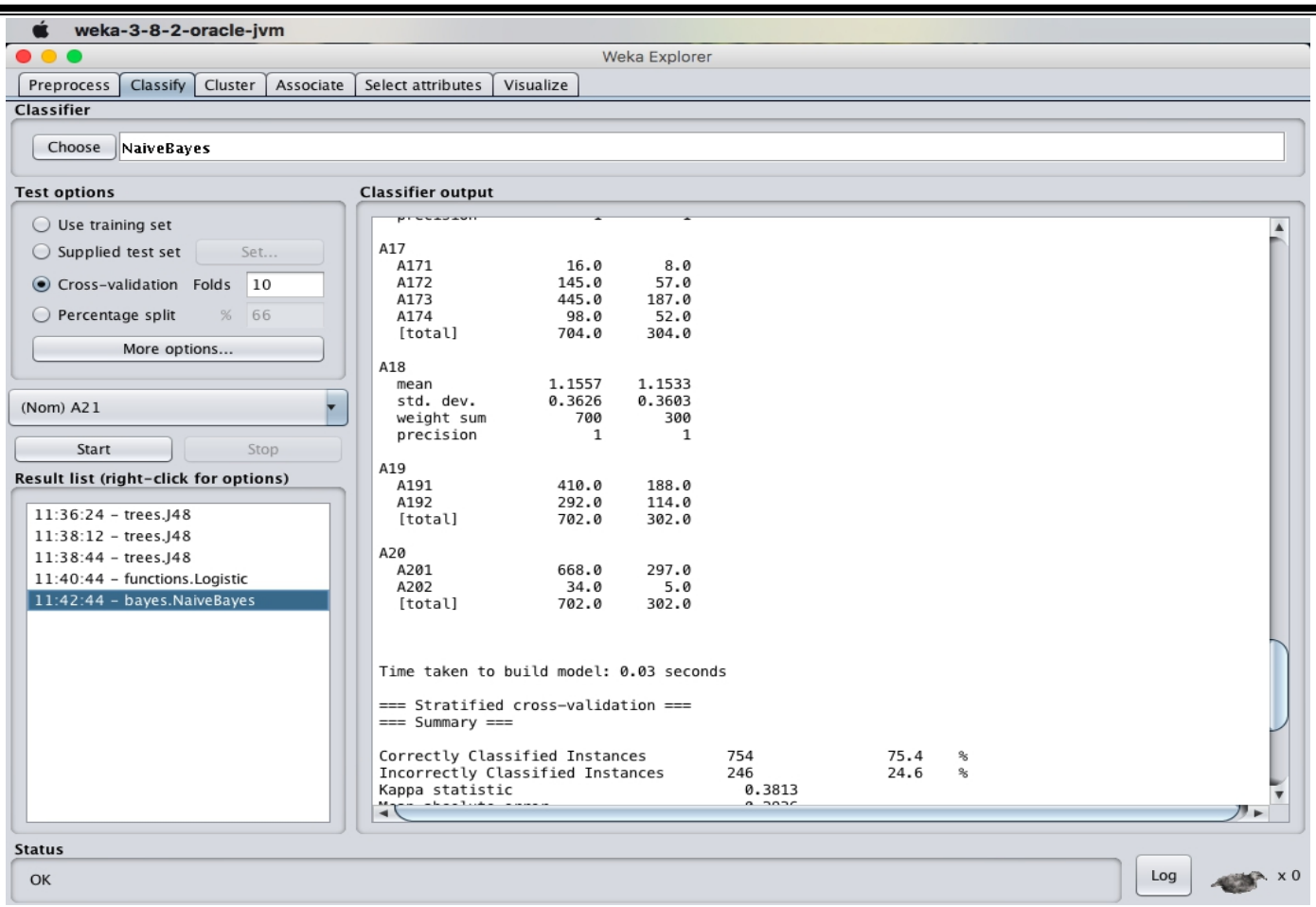
The Naive Bayesian classifier is based on Bayes' theorem with the independence assumptions between predictors. A Naive Bayesian model is easy to build, with no complicated iterative parameter estimation which makes it particularly useful for very large datasets. Despite its simplicity, the Naive Bayesian classifier often does surprisingly well and is widely used because it often outperforms more sophisticated classification methods.

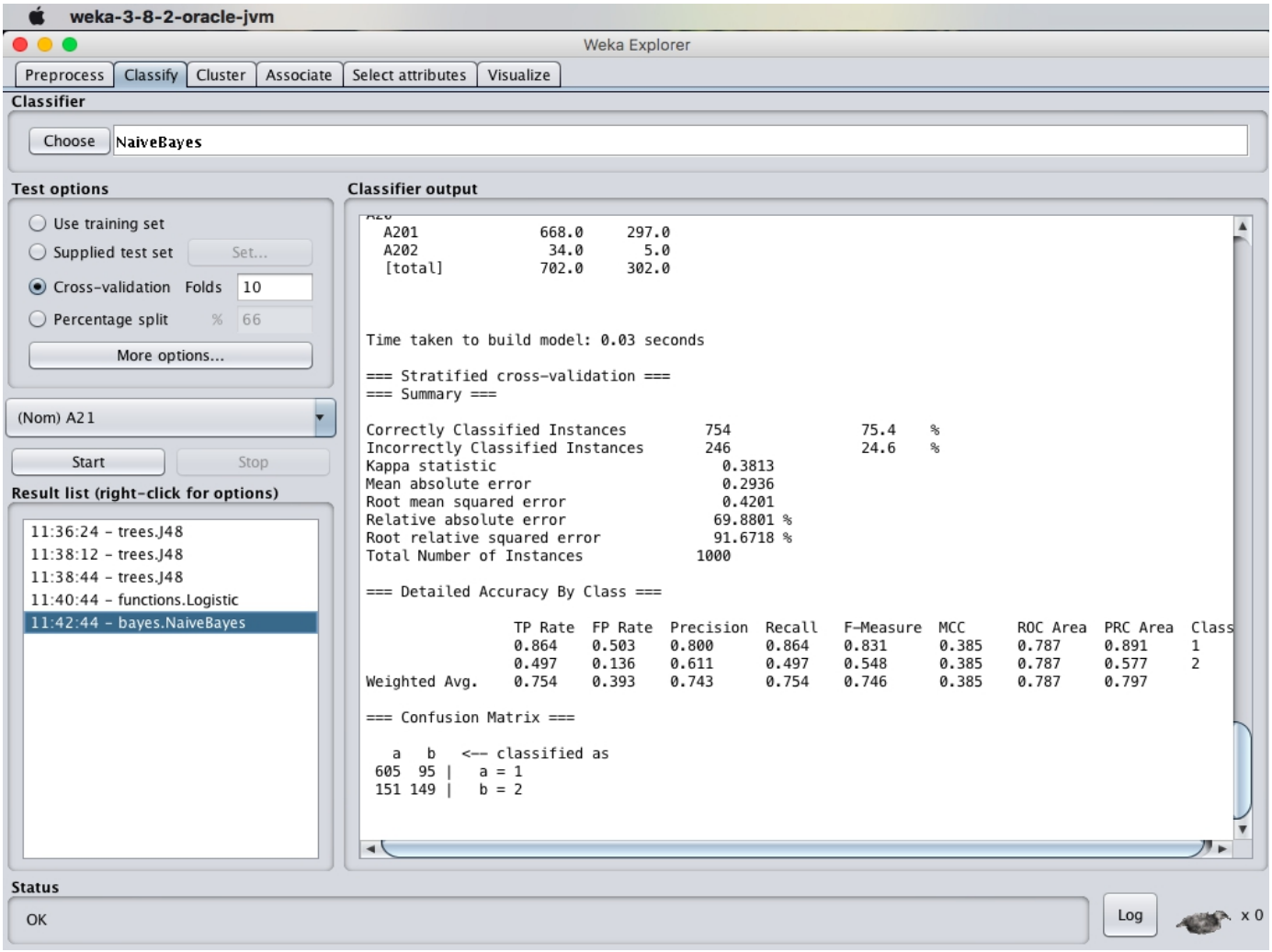
Steps :

- Load the dataset into the weka tool and preprocess it.
- Apply the classification the Naïve bayes technique and execute for the result.

Output :




weka-3-8-2-oracle-jvm
Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
NaiveBayes
Test options
Use training set
Supplied test set
Set...
Cross-validation
Folds
10
Percentage split
%
66
More options...
(Nom) A21
Start
Stop
Result list (right-click for options)
11:36:24 - trees.J48
11:38:12 - trees.J48
11:38:44 - trees.J48
11:40:44 - functions.Logistic
11:42:44 - bayes.NaiveBayes
Classifier output
A17
A171 16.0 8.0
A172 145.0 57.0
A173 445.0 187.0
A174 98.0 52.0
[total] 704.0 304.0
A18
mean 1.1557 1.1533
std. dev. 0.3626 0.3603
weight sum 700 300
precision 1 1
A19
A191 410.0 188.0
A192 292.0 114.0
[total] 702.0 302.0
A20
A201 668.0 297.0
A202 34.0 5.0
[total] 702.0 302.0
Time taken to build model: 0.03 seconds
=== Stratified cross-validation ===
=== Summary ===
Correctly Classified Instances 754 75.4 %
Incorrectly Classified Instances 246 24.6 %
Kappa statistic 0.3813
Status
OK
Log
x 0


weka-3-8-2-oracle-jvm
Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
NaiveBayes
Test options
Use training set
Supplied test set
Set...
Cross-validation
Folds
10
Percentage split
%
66
More options...
(Nom) A21
Start
Stop
Result list (right-click for options)
11:36:24 - trees.J48
11:38:12 - trees.J48
11:38:44 - trees.J48
11:40:44 - functions.Logistic
11:42:44 - bayes.NaiveBayes
Classifier output
A201 668.0 297.0
A202 34.0 5.0
[total] 702.0 302.0
Time taken to build model: 0.03 seconds
=== Stratified cross-validation ===
=== Summary ===
Correctly Classified Instances 754 75.4 %
Incorrectly Classified Instances 246 24.6 %
Kappa statistic 0.3813
Mean absolute error 0.2936
Root mean squared error 0.4201
Relative absolute error 69.8801 %
Root relative squared error 91.6718 %
Total Number of Instances 1000
=== Detailed Accuracy By Class ===
TP Rate FP Rate Precision Recall F-Measure MCC ROC Area PRC Area Class
0.864 0.503 0.800 0.864 0.831 0.385 0.787 0.891 1
0.497 0.136 0.611 0.497 0.548 0.385 0.787 0.577 2
Weighted Avg. 0.754 0.393 0.743 0.754 0.746 0.385 0.787 0.797
=== Confusion Matrix ===
a b <-- classified as
605 95 | a = 1
151 149 | b = 2
Status
OK
Log
x 0

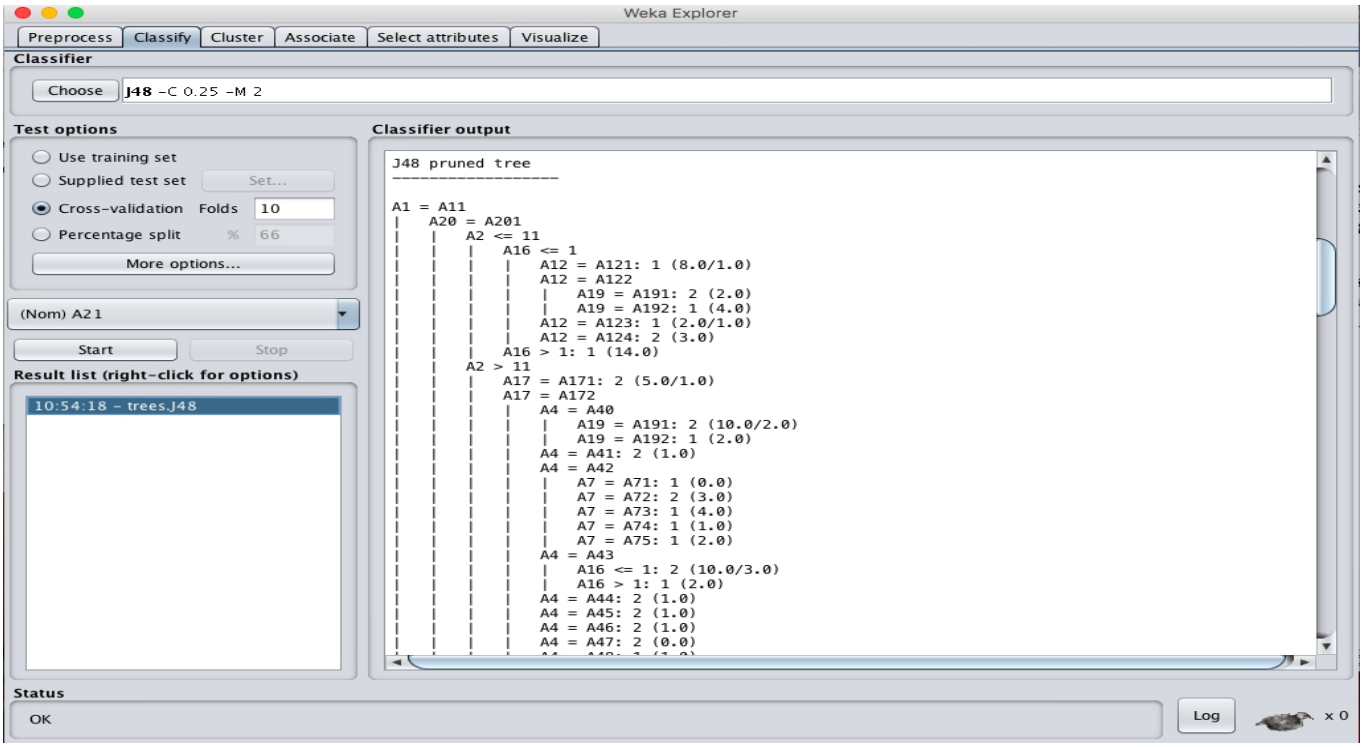
C) J48 Algorithm :

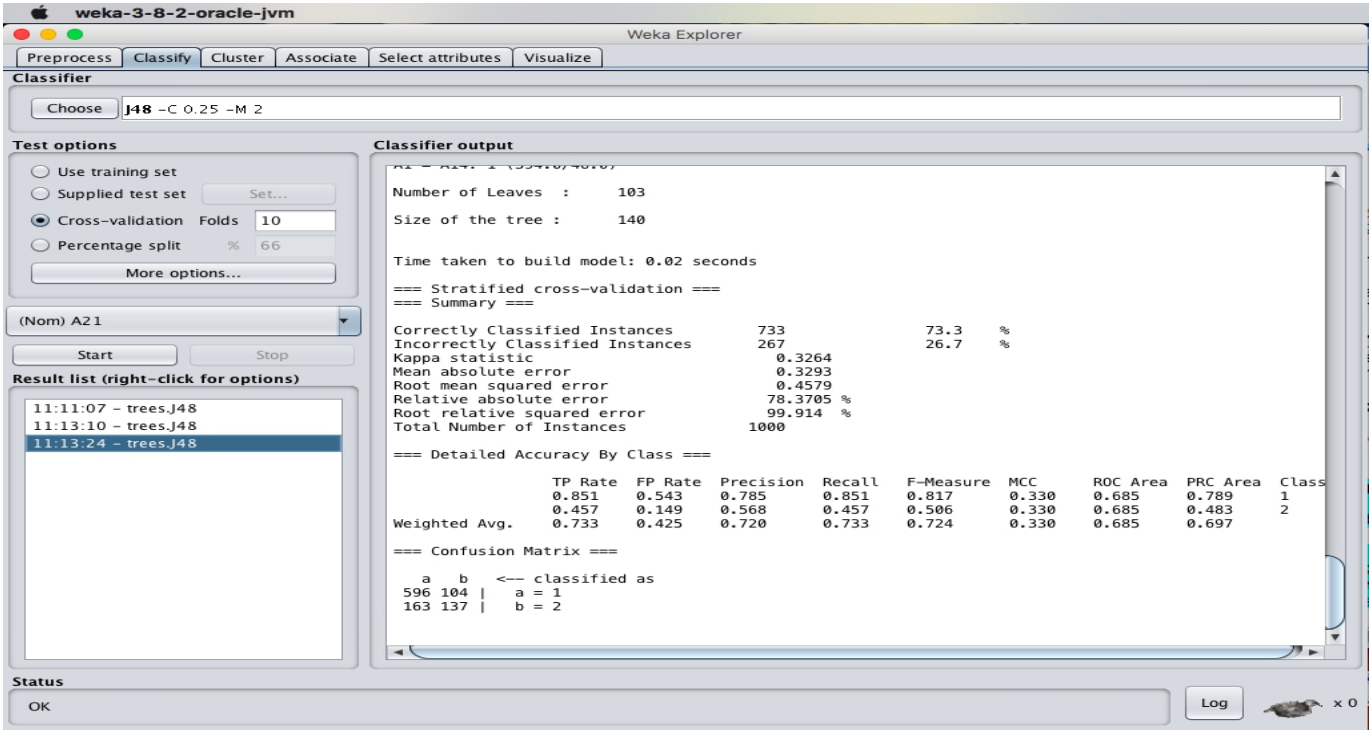
Classification is the process of building a model of classes from a set of records that contain class labels. Decision Tree Algorithm is to find out the way the attributes-vector behaves for a number of instances. Also on the bases of the training instances the classes for the newly generated instances are being found. This algorithm generates the rules for the prediction of the target variable. With the help of tree classification algorithm the critical distribution of the data is easily understandable.

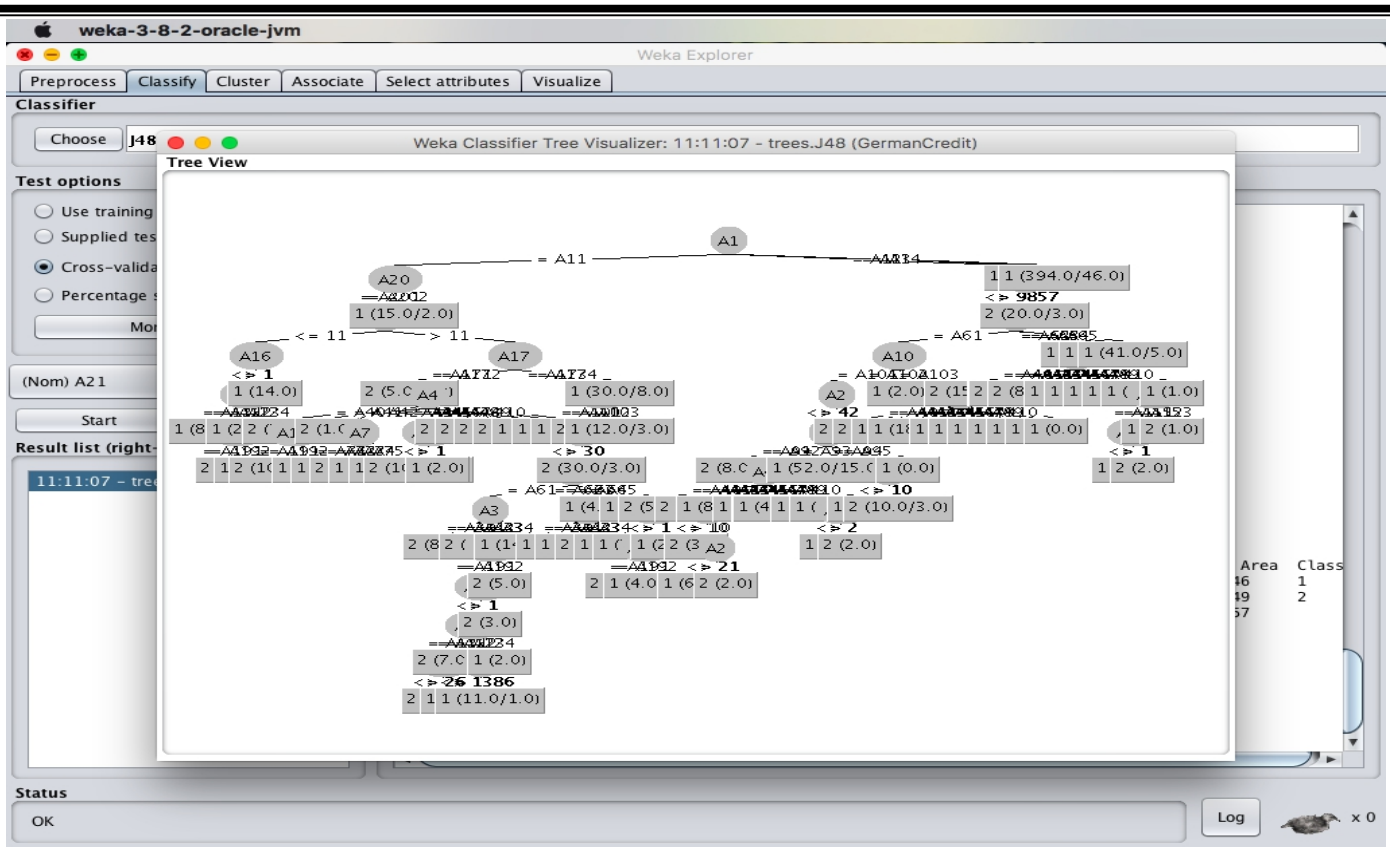
Steps :

- Load the dataset into the weka tool and preprocess it.
- Apply the classification the J48 technique and execute for the result.

Output :







D) K-Nearest Neighbor :

K-Nearest Neighbors is one of the most basic yet essential classification algorithms in Machine Learning. It belongs to the supervised learning domain and finds intense application in pattern recognition, data mining and intrusion detection.

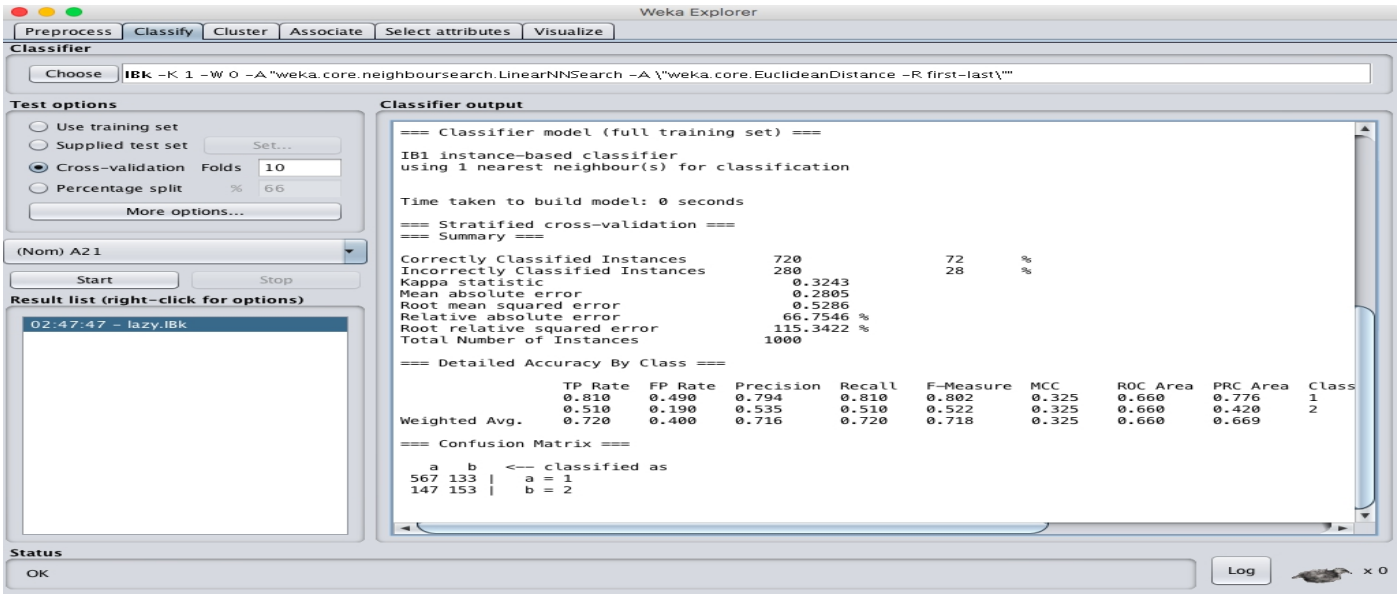
It is widely disposable in real-life scenarios since it is non-parametric, meaning, it does not make any underlying assumptions about the distribution of data (as opposed to other algorithms such as GMM, which assume a Gaussian distribution of the given data).

We are given some prior data (also called training data), which classifies coordinates into groups identified by an attribute.

Steps :

- Load the dataset into the weka tool and preprocess it.
- Apply the classification the K- Nearest Neighbor technique and execute for the result.
-

Output :

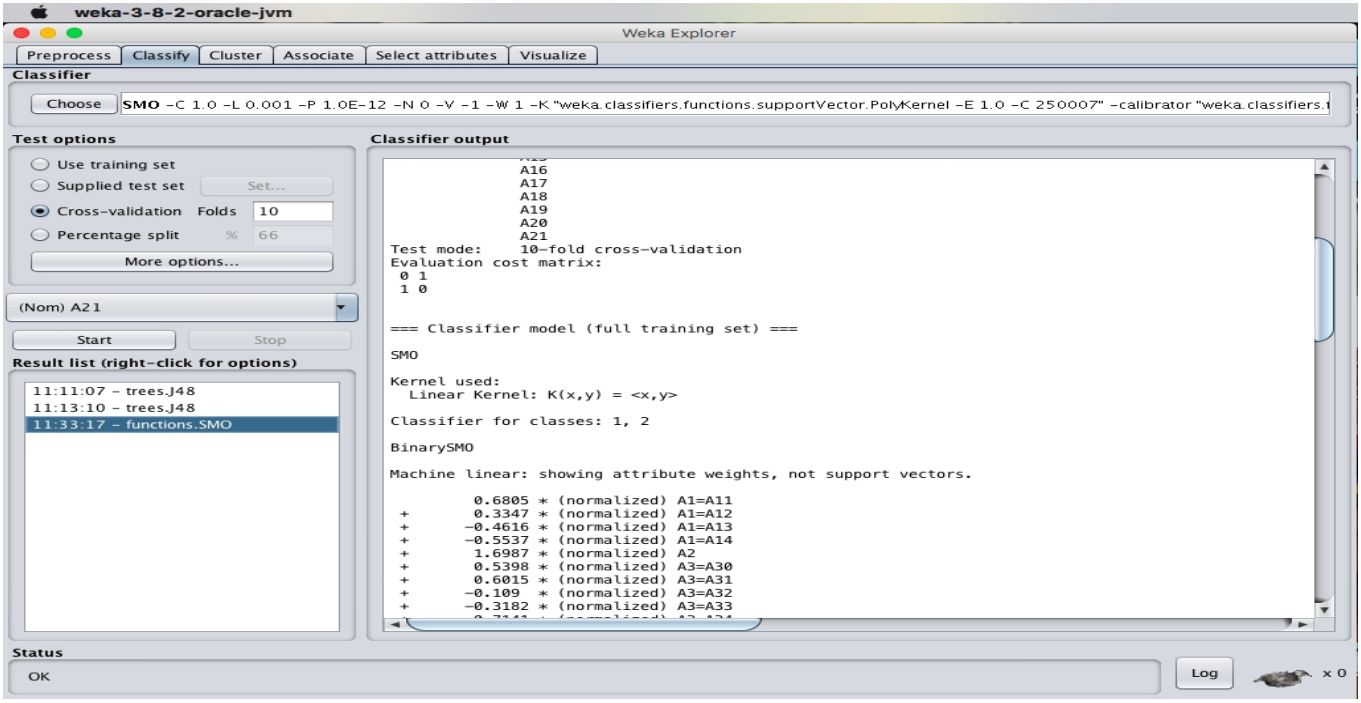


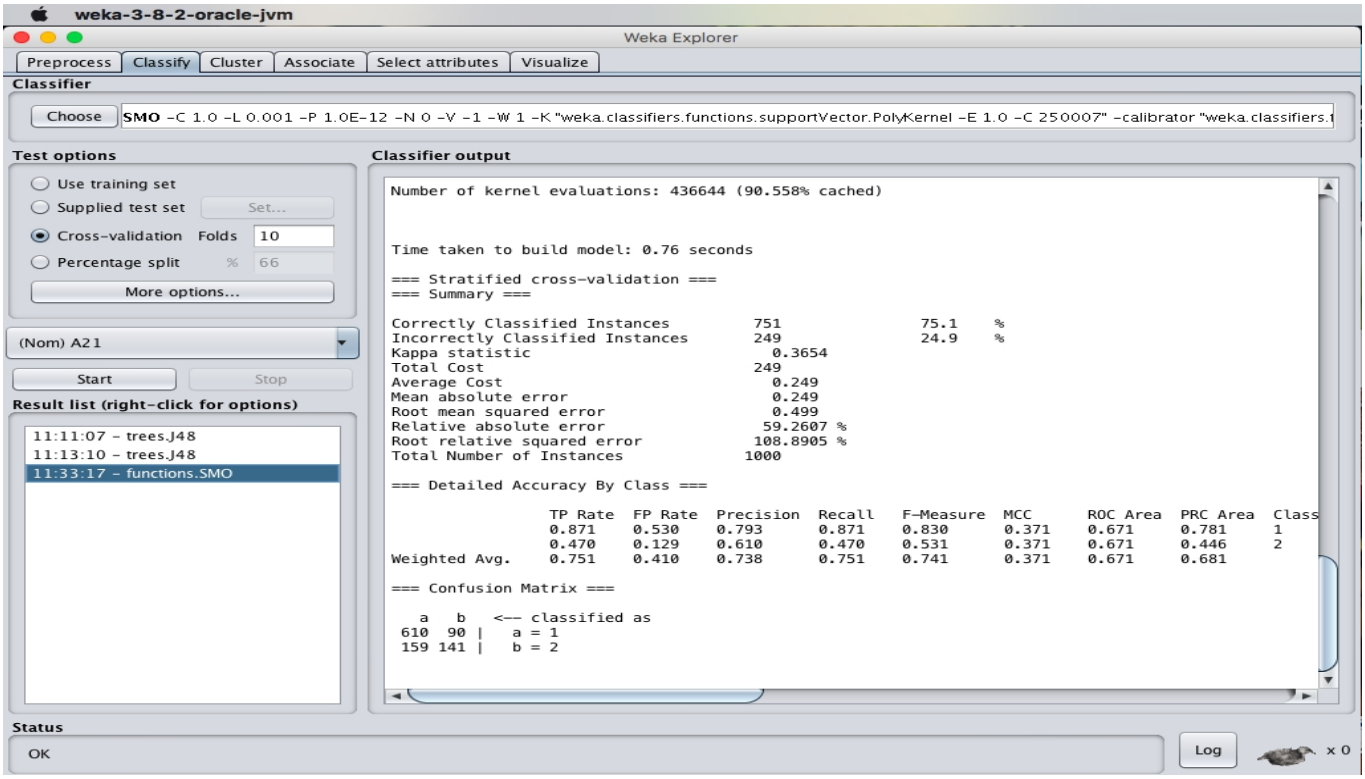
E) SMO Algorithm :

The iterative algorithm Sequential Minimal Optimization (SMO) is used for solving quadratic programming (QP) problems. One example where QP problems are relevant is during the training process of support vector machines (SVM). The SMO algorithm is used to solve in this example a constraint optimization problem. John Platt proposed this algorithm in 1998 and it was successfully used since then. We describe here the basics of the algorithm in the light of big data.

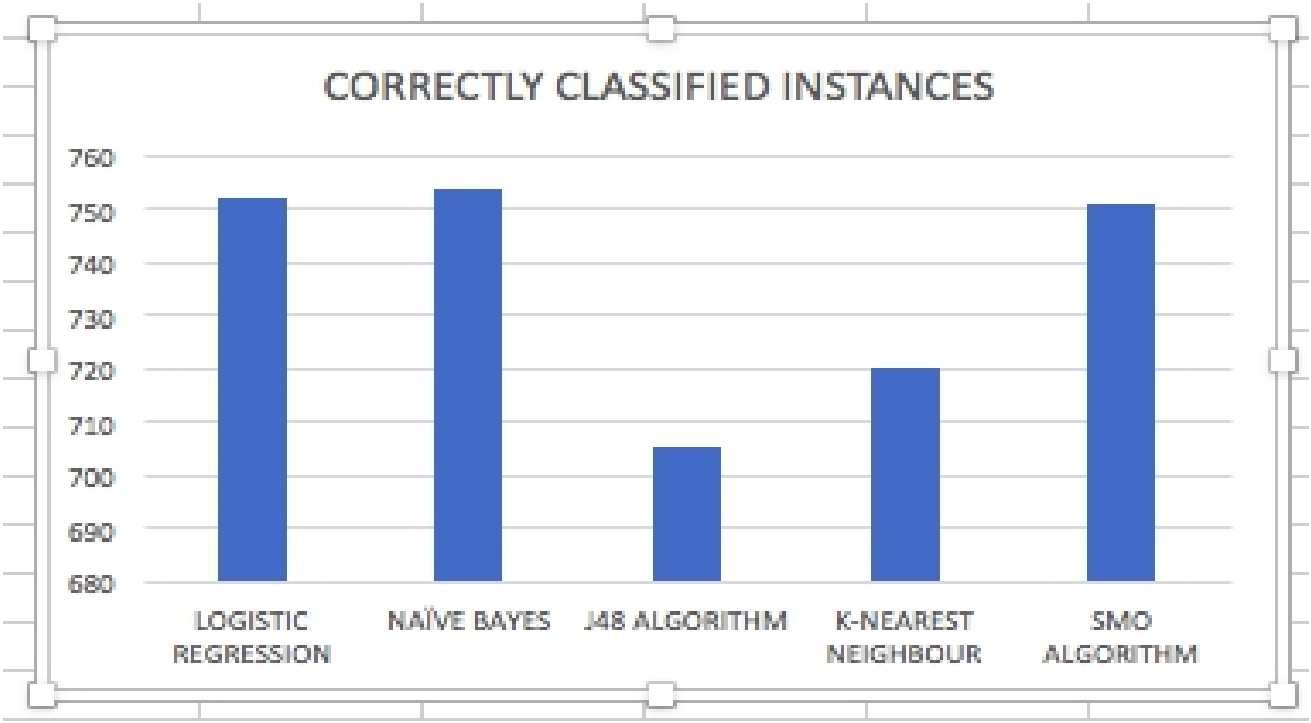
Steps :

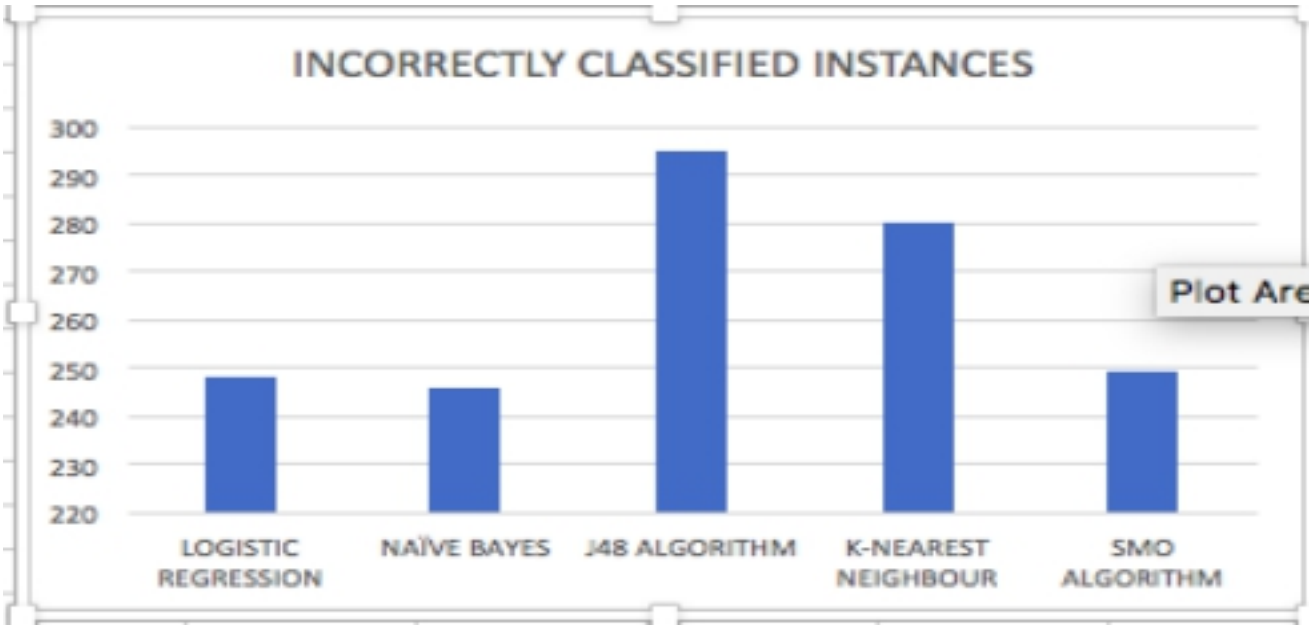
- Load the dataset into the weka tool and preprocess it.
- Apply the classification the Sequential Minimal Optimization (SMO) technique and execute for the result.

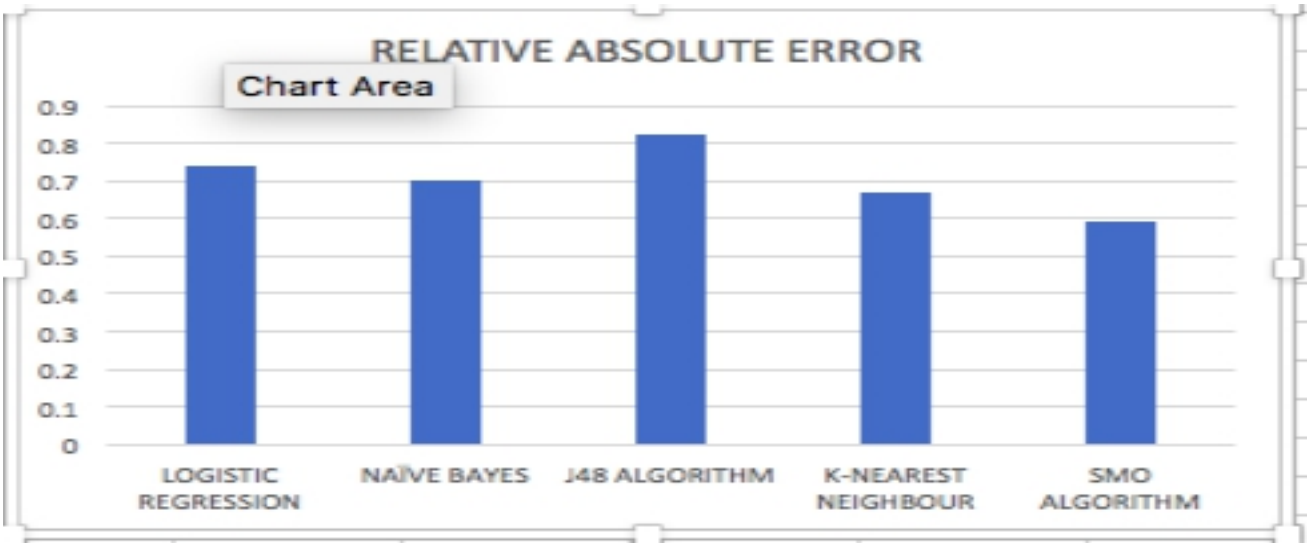


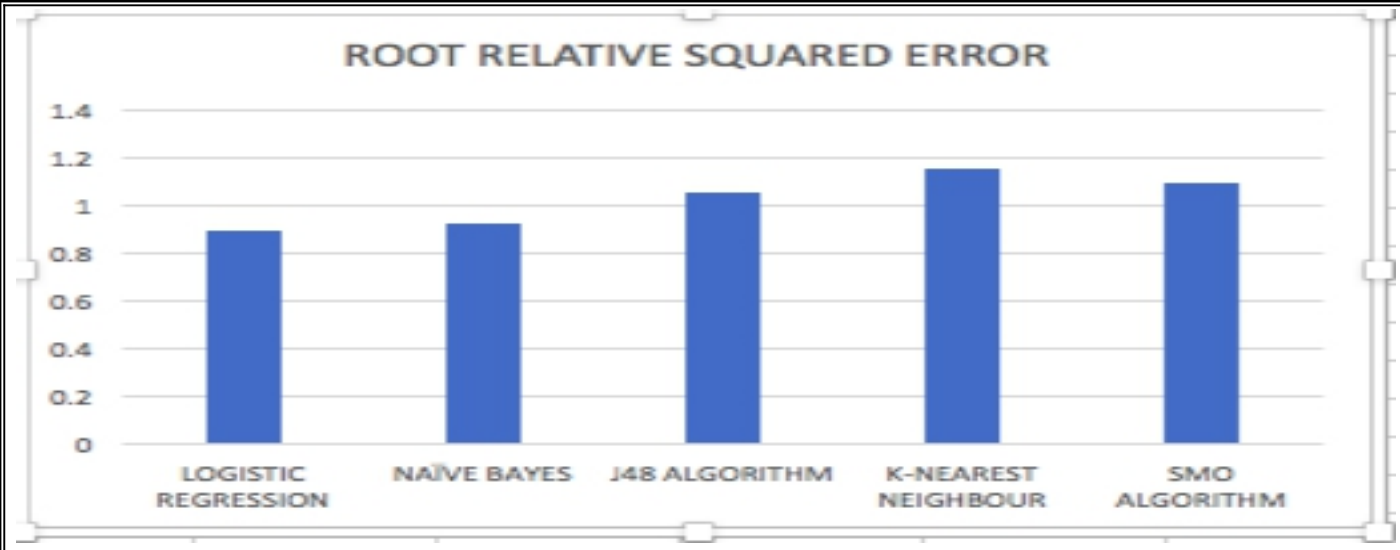


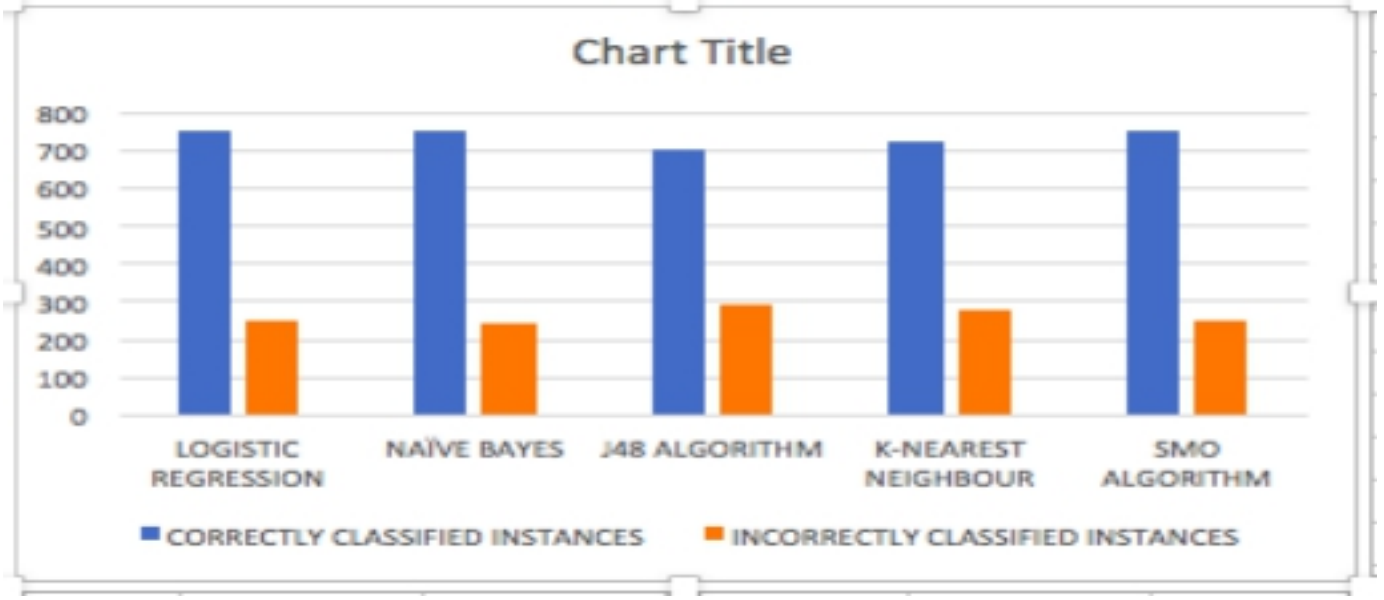
COMPARISION OF VALUES :

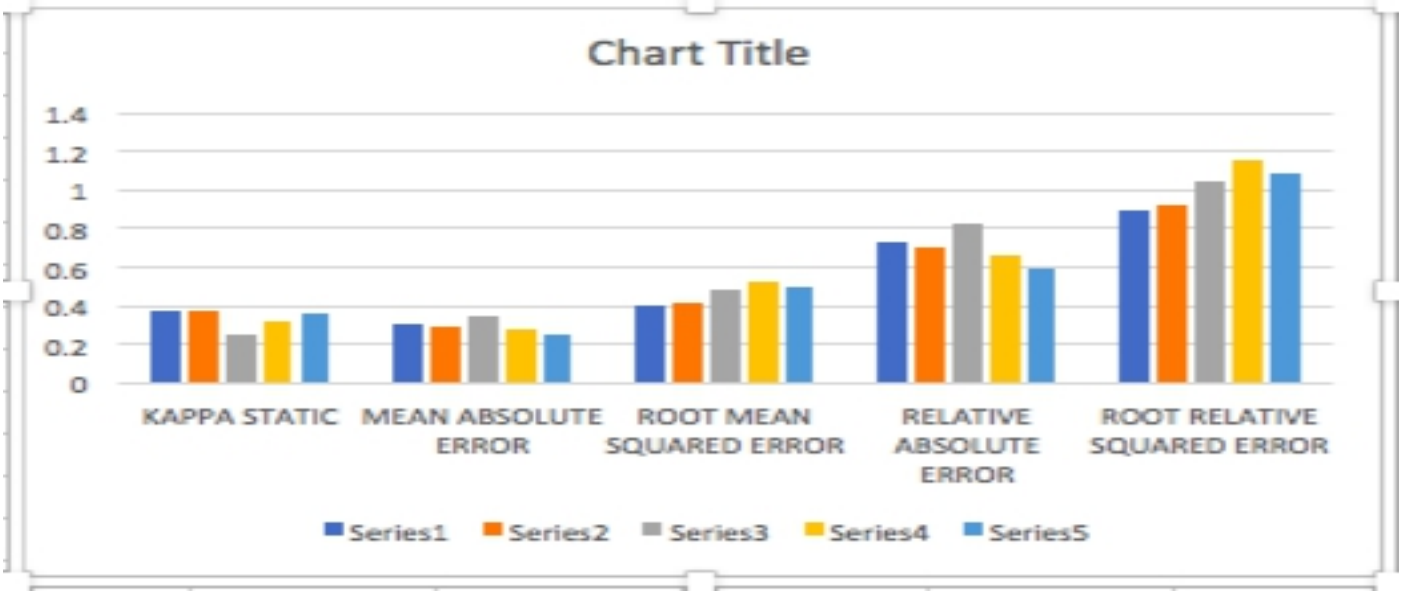












RESULT :

Thus, the comparison of the confusion matrix for all the methods and techniques. Out of the comparing matrix with all the techniques there is a change in instances. Naïve bayes has more number of correct instances than other but when compared to time K-nearest neighbor is best. The above graphs will show the variations of values in the parameters.

EX.No: 14
Date :

NUMERICAL PREDICTION ANALYSIS USING LINEAR REGRESSION THROUGH WEKA

PROBLEM STATEMENT :

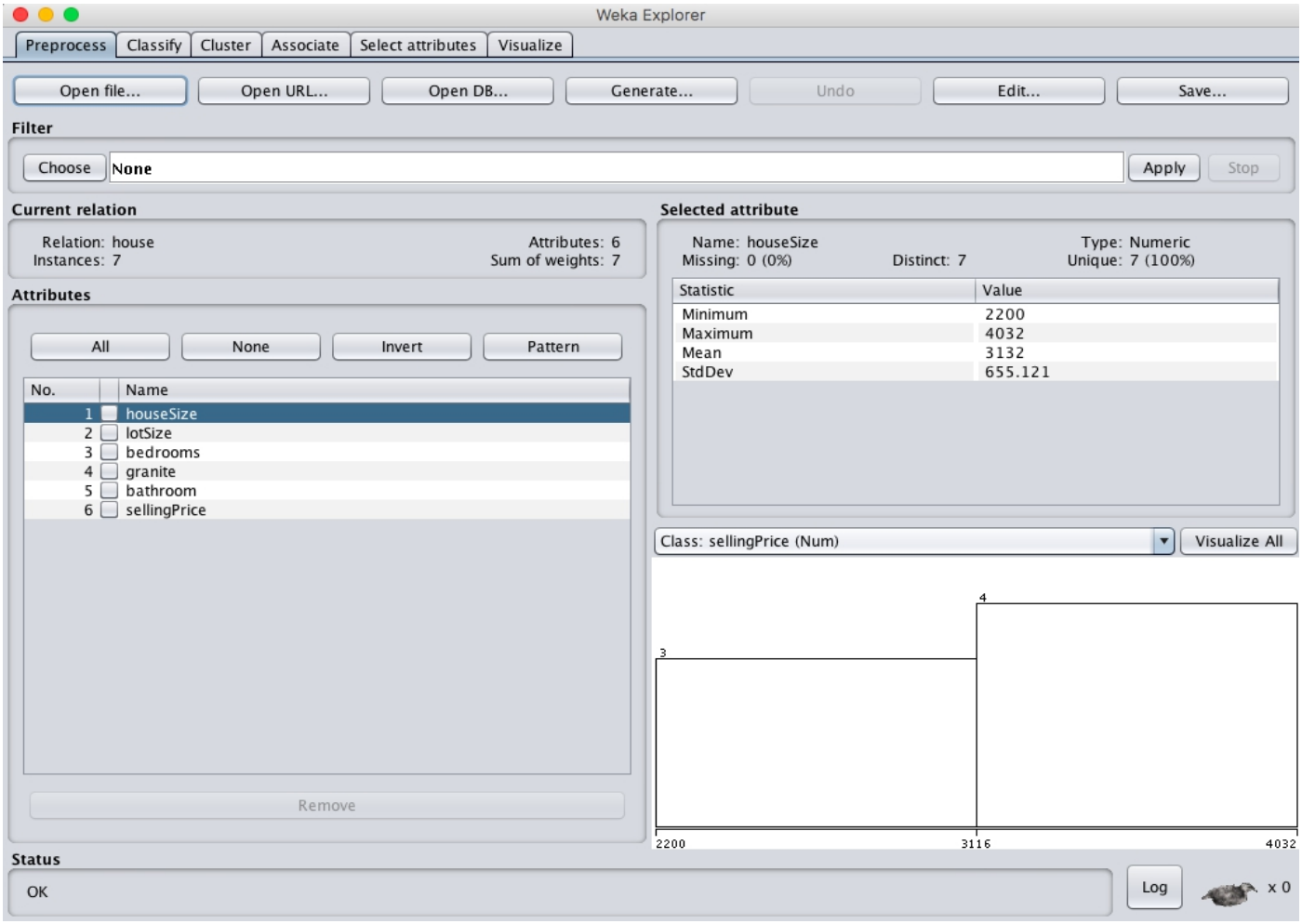
Using regression analysis create a model to calculate the price of the house. Create the model based on other comparable houses in the neighborhood and how much they sold for. Build the dataset for weka in arff file format and load the dataset into weka and finally create the regression model with weka.

DESCRIPTION :

Consider a dataset of house.arff where it contains the attributes as house size, lot size, bedrooms, granite, bathroom and the selling price.

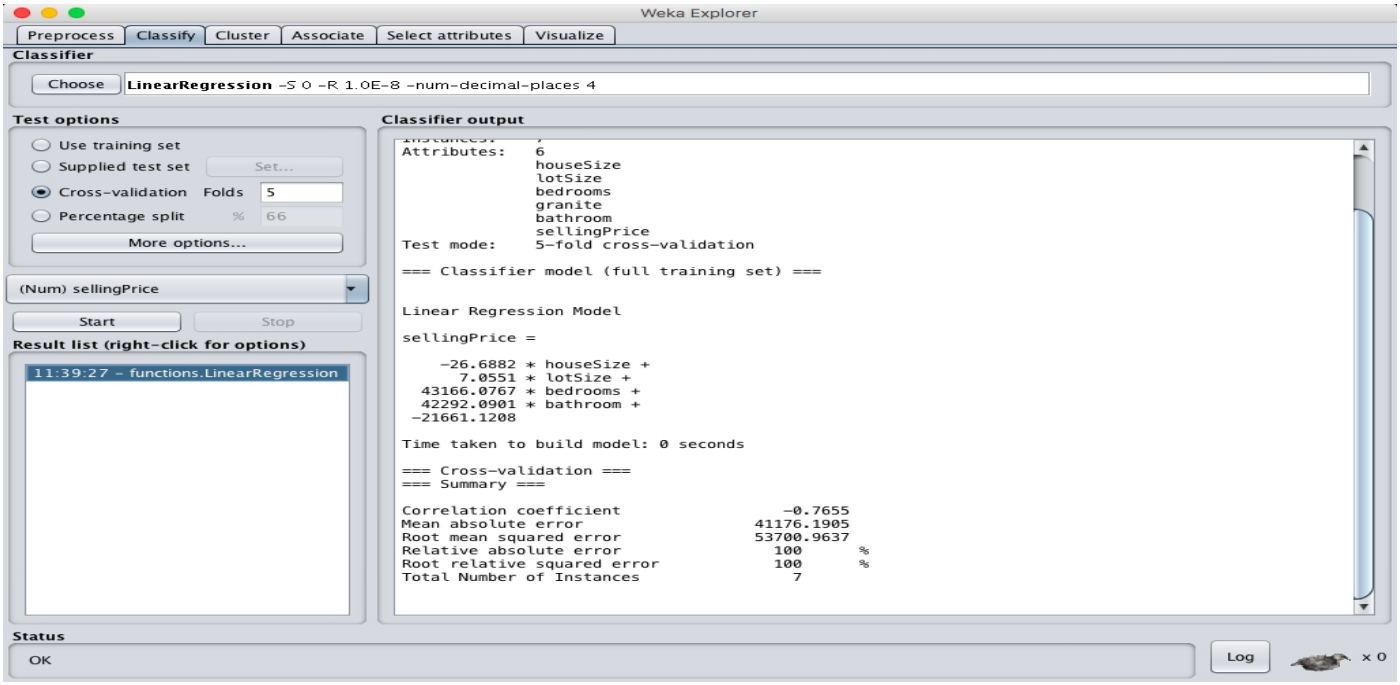
Steps :

- Load the dataset into the weka tool and check for the attributes.
- Classify the data using linear regression analysis method (or) technique.
- Check for the cross-validation folds where the value of the folds should be less than the value of the instances present in the dataset.
- Observe the cross validation summary after applying the linear regression technique for the price of the house.

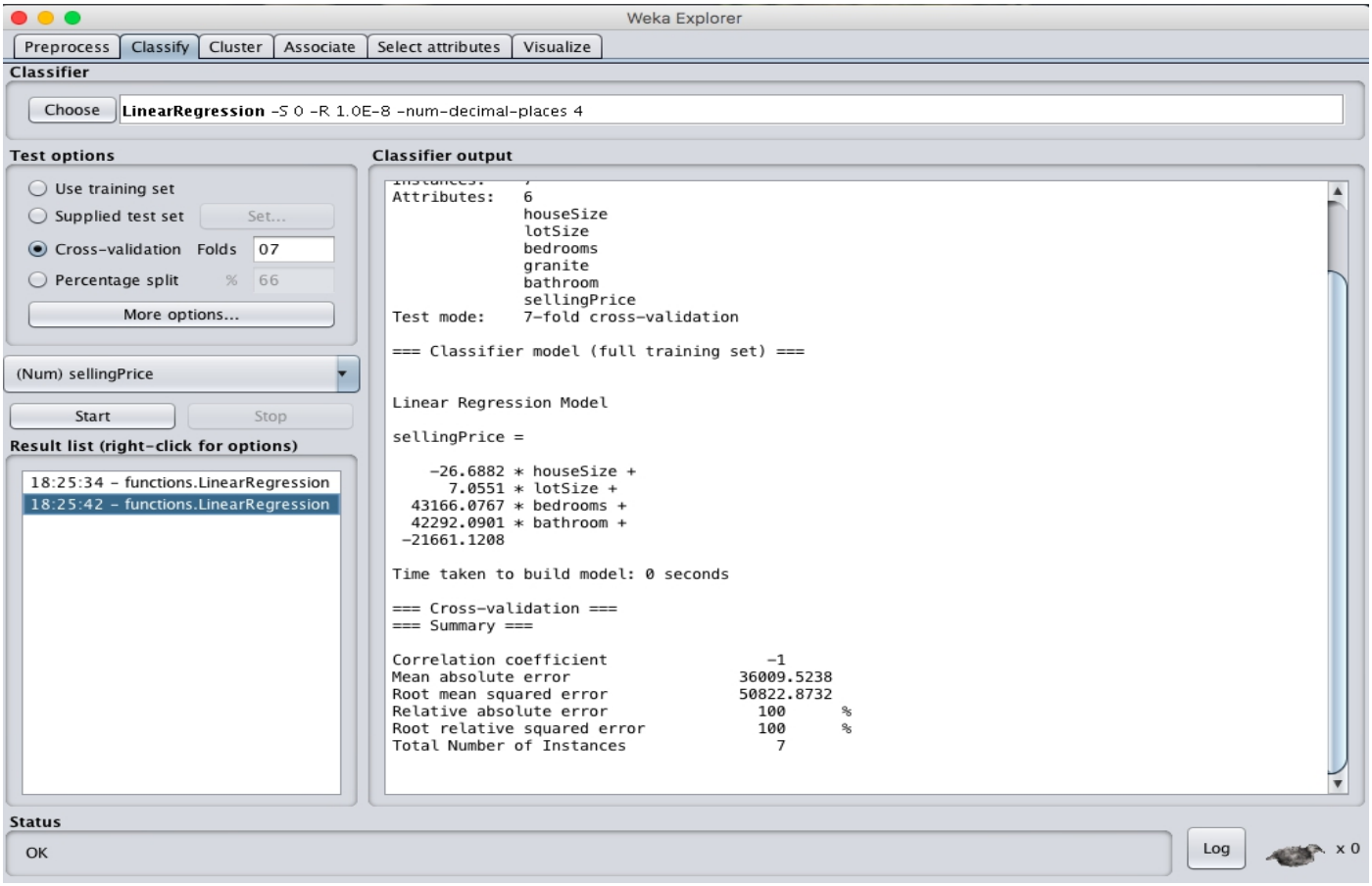


OBSERVATION :

- **When cross validation folds = 05 :**



- **When cross validation folds = 10 :**



RESULT :

Thus, the house selling price has been observed using linear regression model. If the value of cross validation folds decreases time for creating model will be less than when folds value high, and the mean absolute error and Root mean square error values decreases with increase in the cross validation folds value.

EX.No: 15
Date :

EXTRACT TRANSFORM LOAD (ETL) AND OLAP OPERATION USING KNIME TOOL

PROBLEM STATEMENT :

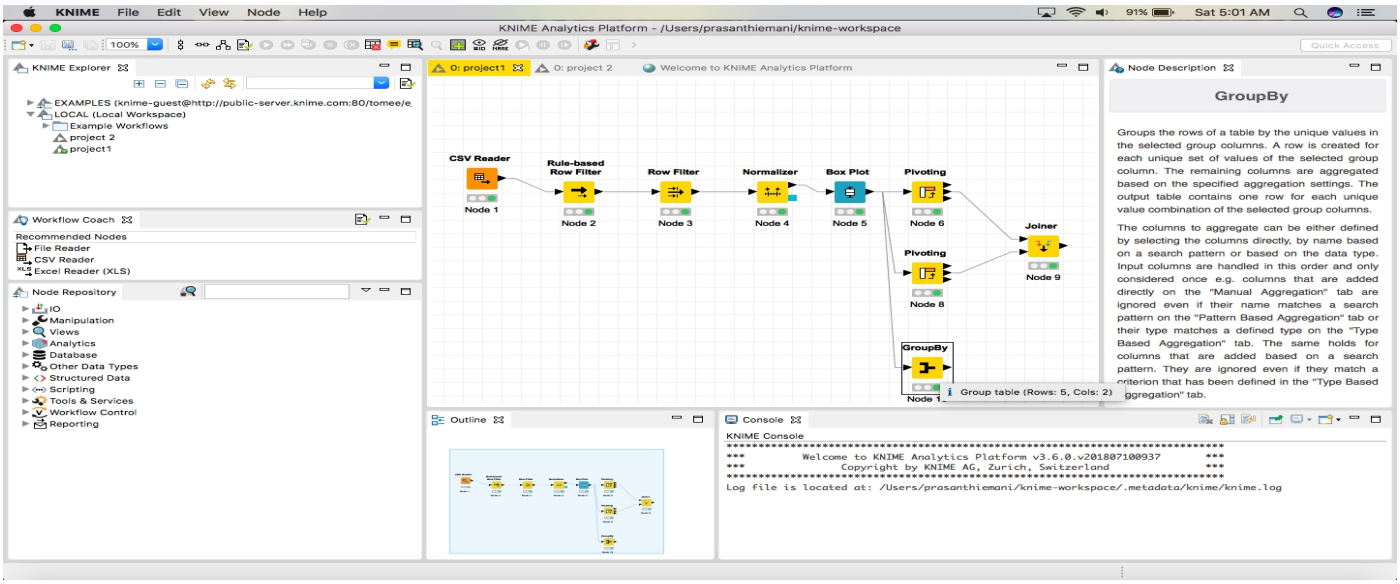
Extract the data from csv file, transform[use row filter and rule based filter] use pivot and group by] load the data for reporting (Visualization).

DESCRIPTION :

Consider a dataset movies.csv where it contains the attributes ad title, genre, director, year, duration, gross, budget, cast_facebook_likes, votes, reviews, rating.

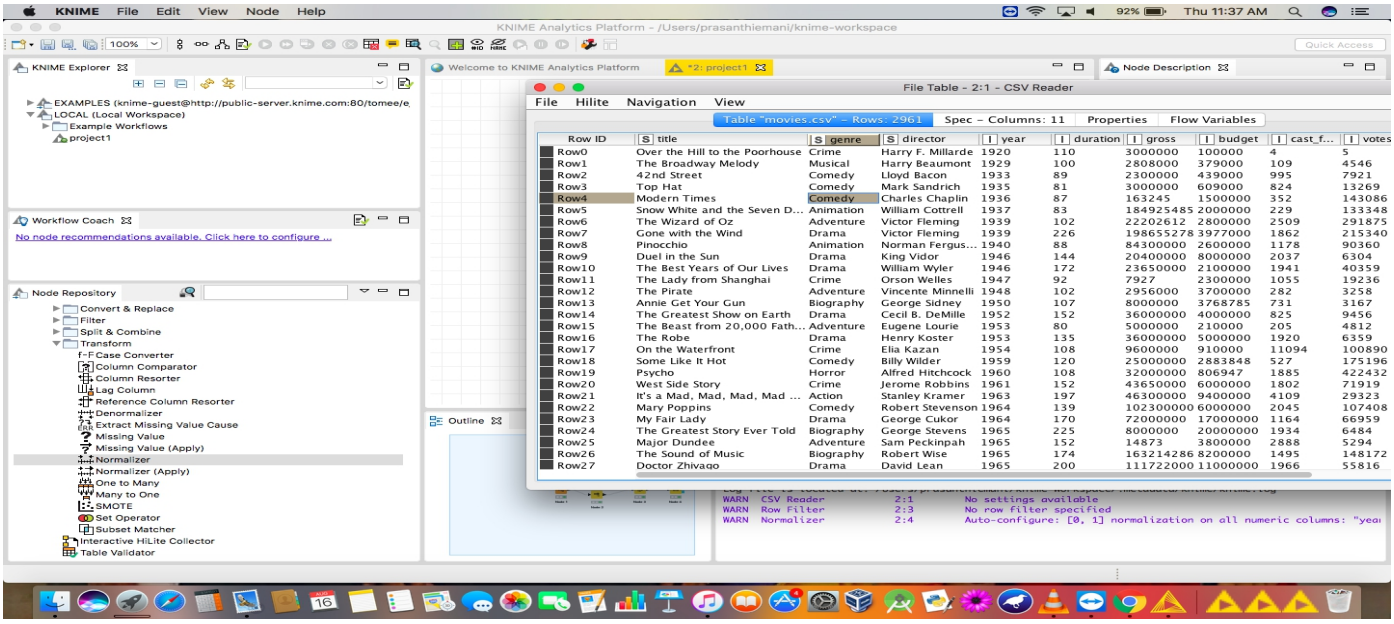
STEPS :

- Import the dataset into the knime tool using csv reader.



- **CSV Reader :**

Execute the CSV reader, look at the table of the loaded dataset.



Row ID	title	genre	director	year	duration	gross	budget	cast_f...	votes
Row0	Over the Hill to the Poorhouse	Crime	Harry F. Millarde	1920	110	3000000	100000	4	5
Row1	The Broadway Melody	Musical	Harry Beaumont	1929	100	2808000	379000	109	4546
Row2	42nd Street	Comedy	Lloyd Bacon	1933	89	2300000	439000	995	7921
Row3	Top Hat	Comedy	Mark Sandrich	1935	81	3000000	609000	824	13269
Row4	Modern Times	Comedy	Charles Chaplin	1936	87	163245	1500000	352	143086
Row5	Snow White and the Seven D...	Animation	William Cottrell	1937	83	184925485	2000000	229	133348
Row6	The Wizard of Oz	Adventure	Victor Fleming	1939	102	22202612	2800000	2509	291875
Row7	Gone with the Wind	Drama	Victor Fleming	1939	226	198655278	3977000	1862	215340
Row8	Pinocchio	Animation	Norman Fergus...	1940	88	84300000	2600000	1178	90360
Row9	Duel in the Sun	Drama	King Vidor	1946	144	20400000	8000000	2037	6304
Row10	The Best Years of Our Lives	Drama	William Wyler	1946	172	23650000	2100000	1941	40359
Row11	The Lady from Shanghai	Crime	Orson Welles	1947	92	7927	2300000	1055	19236
Row12	The Pirate	Adventure	Vincente Minnelli	1948	102	2956000	3700000	282	3258
Row13	Annie Get Your Gun	Biography	George Sidney	1950	107	8000000	3768785	731	3167
Row14	The Greatest Show on Earth	Drama	Cecil B. DeMille	1952	152	36000000	4000000	825	9456
Row15	The Beast from 20,000 Fath...	Adventure	Eugene Lourie	1953	80	5000000	210000	205	4812
Row16	The Robe	Drama	Henry Koster	1953	135	36000000	5000000	1920	6359
Row17	On the Waterfront	Crime	Elia Kazan	1954	108	9600000	910000	11094	100890
Row18	Some Like It Hot	Comedy	Billy Wilder	1959	120	25000000	2883848	527	175196
Row19	Psycho	Horror	Alfred Hitchcock	1960	108	32000000	806947	1885	422432
Row20	West Side Story	Crime	Jerome Robbins	1961	152	43650000	6000000	1802	71919
Row21	It's a Mad, Mad, Mad, Mad ...	Action	Stanley Kramer	1963	197	46300000	9400000	4109	29323
Row22	Mary Poppins	Comedy	Robert Stevenson	1964	139	102300000	6000000	2045	107408
Row23	My Fair Lady	Drama	George Cukor	1964	170	72000000	17000000	1164	66959
Row24	The Greatest Story Ever Told	Biography	George Stevens	1965	225	8000000	20000000	1934	6484
Row25	Major Dundee	Adventure	Sam Peckinpah	1965	152	14873	3800000	2888	5294
Row26	The Sound of Music	Biography	Robert Wise	1965	174	163214286	8200000	1495	148172
Row27	Doctor Zhivago	Drama	David Lean	1965	200	111722000	11000000	1966	55816

❖ **Rule-Based Row Filter :**

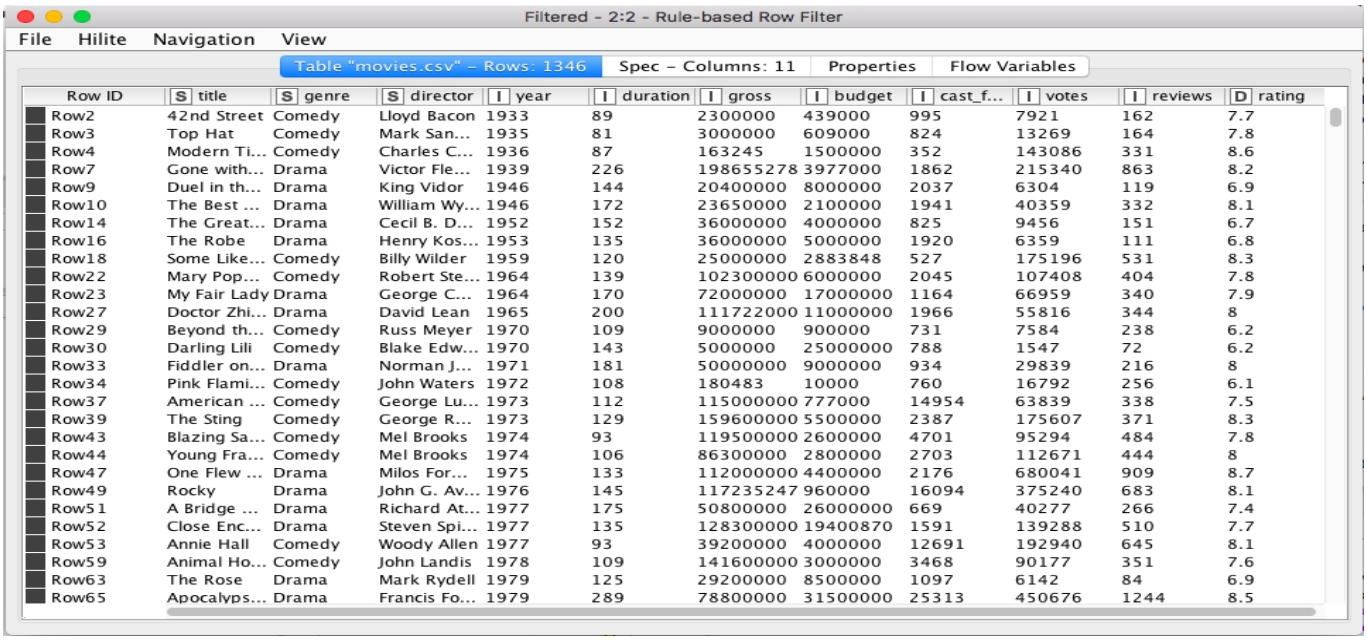
This node takes a list of user-defined rules and tries to match them to each row in the input table. If the first matching rule has a TRUE outcome, the row will be selected for inclusion. Otherwise (i.e. if the first matching rule yields FALSE) it will be excluded. If no rule matches the row will be excluded.

Steps :

- Make a connection between CSV reader and rule based row filter.
- Configure rule based row filter.
- Execute and Check out for the table after applying rule based row filter.

Code :

$genre$ = “Drama” XOR $genre$ = “Comedy” => TRUE



Row ID	title	genre	director	year	duration	gross	budget	cast_f...	votes	reviews	rating
Row2	42nd Street	Comedy	Lloyd Bacon	1933	89	2300000	439000	995	7921	162	7.7
Row3	Top Hat	Comedy	Mark San...	1935	81	3000000	609000	824	13269	164	7.8
Row4	Modern Ti...	Comedy	Charles C...	1936	87	163245	1500000	352	143086	331	8.6
Row7	Gone with...	Drama	Victor Fle...	1939	226	198655278	3977000	1862	215340	863	8.2
Row9	Duel in th...	Drama	King Vidor	1946	144	20400000	8000000	2037	6304	119	6.9
Row10	The Best ...	Drama	William Wy...	1946	172	23650000	2100000	1941	40359	332	8.1
Row14	The Great...	Drama	Cecil B. D...	1952	152	36000000	4000000	825	9456	151	6.7
Row16	The Robe	Drama	Henry Kos...	1953	135	36000000	5000000	1920	6359	111	6.8
Row18	Some Like...	Comedy	Billy Wilder	1959	120	25000000	2883848	527	175196	531	8.3
Row22	Mary Pop...	Comedy	Robert Ste...	1964	139	102300000	6000000	2045	107408	404	7.8
Row23	My Fair Lady	Drama	George C...	1964	170	72000000	17000000	1164	66959	340	7.9
Row27	Doctor Zhi...	Drama	David Lean	1965	200	111722000	11000000	1966	55816	344	8
Row29	Beyond th...	Comedy	Russ Meyer	1970	109	9000000	900000	731	7584	238	6.2
Row30	Darling Lili	Comedy	Blake Edw...	1970	143	5000000	25000000	788	1547	72	6.2
Row33	Fiddler on...	Drama	Norman J...	1971	181	50000000	9000000	934	29839	216	8
Row34	Pink Flami...	Comedy	John Waters	1972	108	180483	10000	760	16792	256	6.1
Row37	American ...	Comedy	George Lu...	1973	112	115000000	777000	14954	63839	338	7.5
Row39	The Sting	Comedy	George R...	1973	129	159600000	5500000	2387	175607	371	8.3
Row43	Blazing Sa...	Comedy	Mel Brooks	1974	93	119500000	2600000	4701	95294	484	7.8
Row44	Young Fra...	Comedy	Mel Brooks	1974	106	86300000	2800000	2703	112671	444	8
Row47	One Flew ...	Drama	Milos For...	1975	133	112000000	4400000	2176	680041	909	8.7
Row49	Rocky	Drama	John G. Av...	1976	145	117235247	960000	16094	375240	683	8.1
Row51	A Bridge ...	Drama	Richard At...	1977	175	50800000	26000000	669	40277	266	7.4
Row52	Close Enc...	Drama	Steven Spi...	1977	135	128300000	19400870	1591	139288	510	7.7
Row53	Annie Hall	Comedy	Woody Allen	1977	93	39200000	4000000	12691	192940	645	8.1
Row59	Animal Ho...	Comedy	John Landis	1978	109	141600000	3000000	3468	90177	351	7.6
Row63	The Rose	Drama	Mark Rydell	1979	125	29200000	8500000	1097	6142	84	6.9
Row65	Apocalyps...	Drama	Francis Fo...	1979	289	78800000	31500000	25313	450676	1244	8.5

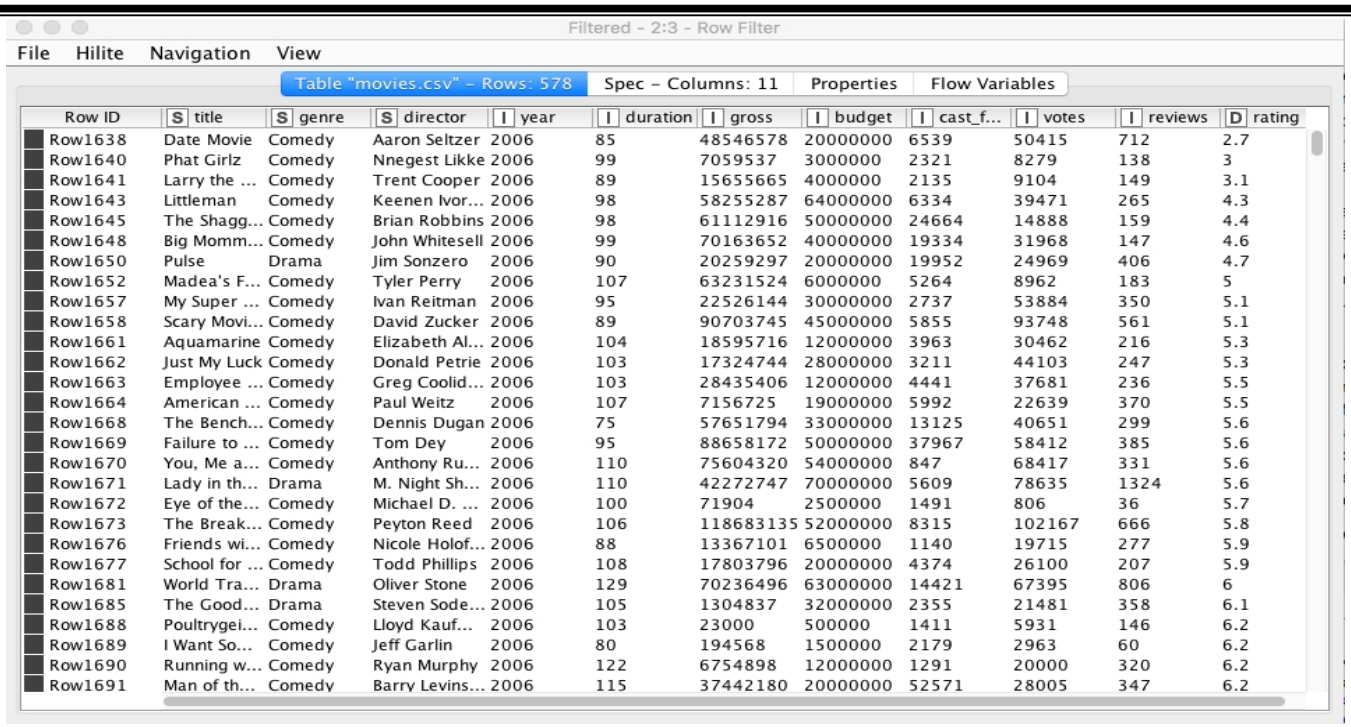
❖ **Row Filter :**

3 matching criteria on data colums: on String by full or partial pattern matching, on numbers by range, on missing values, all of them also on collection columns. 1 matching criterion on row numbers: from row number to row number. 1 matching criterion on RowID: full and partial patterm matching. Partial pattern matching is obtained through wild cards and RegEx. All matching criteria can be used in Include or Exclude mode. Include keeps the match results. Exclude excludes it.

Steps :

- Make a connection between rule based row filter and row filter.
- Configure row filter.
- Execute and Check out for the table after applying row filter.

➢ Use range checking :
Lower Bound : 2006
Upper Bound : 2016

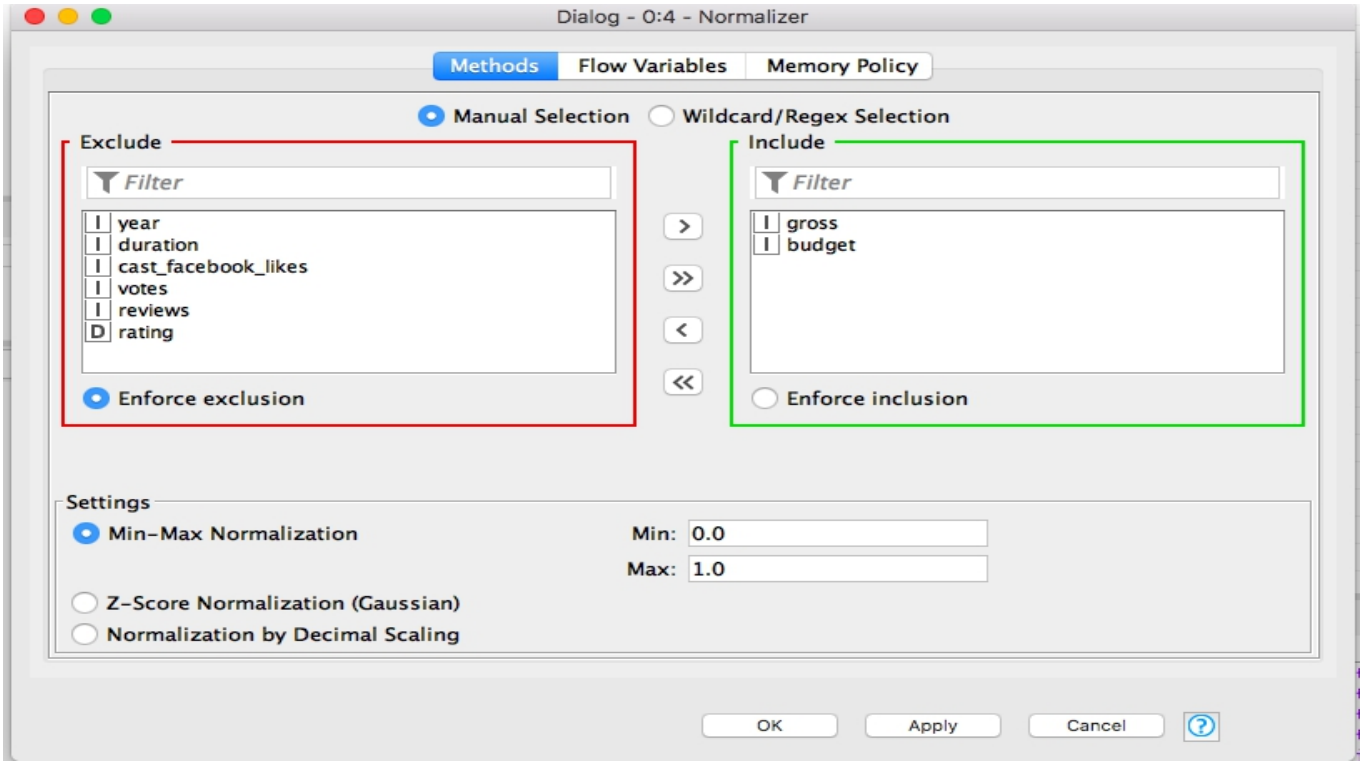


Row ID	title	genre	director	year	duration	gross	budget	cast_f...	votes	reviews	rating
Row1638	Date Movie	Comedy	Aaron Seltzer	2006	85	48546578	20000000	6539	50415	712	2.7
Row1640	Phat Girlz	Comedy	Nnegest Likke	2006	99	7059537	3000000	2321	8279	138	3
Row1641	Larry the ...	Comedy	Trent Cooper	2006	89	15655665	4000000	2135	9104	149	3.1
Row1643	Littleman	Comedy	Keenen Ivor...	2006	98	58255287	64000000	6334	39471	265	4.3
Row1645	The Shagg...	Comedy	Brian Robbins	2006	98	61112916	50000000	24664	14888	159	4.4
Row1648	Big Momm...	Comedy	John Whitesell	2006	99	70163652	40000000	19334	31968	147	4.6
Row1650	Pulse	Drama	Jim Sonzero	2006	90	20259297	20000000	19952	24969	406	4.7
Row1652	Madea's F...	Comedy	Tyler Perry	2006	107	63231524	6000000	5264	8962	183	5
Row1657	My Super ...	Comedy	Ivan Reitman	2006	95	22526144	30000000	2737	53884	350	5.1
Row1658	Scary Movi...	Comedy	David Zucker	2006	89	90703745	45000000	5855	93748	561	5.1
Row1661	Aquamarine	Comedy	Elizabeth Al...	2006	104	18595716	12000000	3963	30462	216	5.3
Row1662	Just My Luck	Comedy	Donald Petrie	2006	103	17324744	28000000	3211	44103	247	5.3
Row1663	Employee ...	Comedy	Greg Coolid...	2006	103	28435406	12000000	4441	37681	236	5.5
Row1664	American ...	Comedy	Paul Weitz	2006	107	7156725	19000000	5992	22639	370	5.5
Row1668	The Bench...	Comedy	Dennis Dugan	2006	75	57651794	33000000	13125	40651	299	5.6
Row1669	Failure to ...	Comedy	Tom Dey	2006	95	88658172	50000000	37967	58412	385	5.6
Row1670	You, Me a...	Comedy	Anthony Ru...	2006	110	75604320	54000000	847	68417	331	5.6
Row1671	Lady in th...	Drama	M. Night Sh...	2006	110	42272747	70000000	5609	78635	1324	5.6
Row1672	Eye of the...	Comedy	Michael D. ...	2006	100	71904	2500000	1491	806	36	5.7
Row1673	The Break...	Comedy	Peyton Reed	2006	106	118683135	52000000	8315	102167	666	5.8
Row1676	Friends wi...	Comedy	Nicole Holof...	2006	88	13367101	6500000	1140	19715	277	5.9
Row1677	School for ...	Comedy	Todd Phillips	2006	108	17803796	20000000	4374	26100	207	5.9
Row1681	World Tra...	Drama	Oliver Stone	2006	129	70236496	63000000	14421	67395	806	6
Row1685	The Good...	Drama	Steven Sode...	2006	105	1304837	32000000	2355	21481	358	6.1
Row1688	Poultrygei...	Comedy	Lloyd Kauf...	2006	103	23000	500000	1411	5931	146	6.2
Row1689	I Want So...	Comedy	Jeff Garlin	2006	80	194568	1500000	2179	2963	60	6.2
Row1690	Running w...	Comedy	Ryan Murphy	2006	122	6754898	12000000	1291	20000	320	6.2
Row1691	Man of th...	Comedy	Barry Levins...	2006	115	37442180	20000000	52571	28005	347	6.2

❖ **Normalizer :**

Steps :

- Connect normalizer with the row filter.
- Configure normalizer as methods which are to be included for normalization technique and set min and max values.
- Include:
  a) Gross
  b) Budget
  c) Min : 0.0
  d) Max : 0.1
- Execute the normalizer and check for the values in the table where you will find the normalized values of the table.



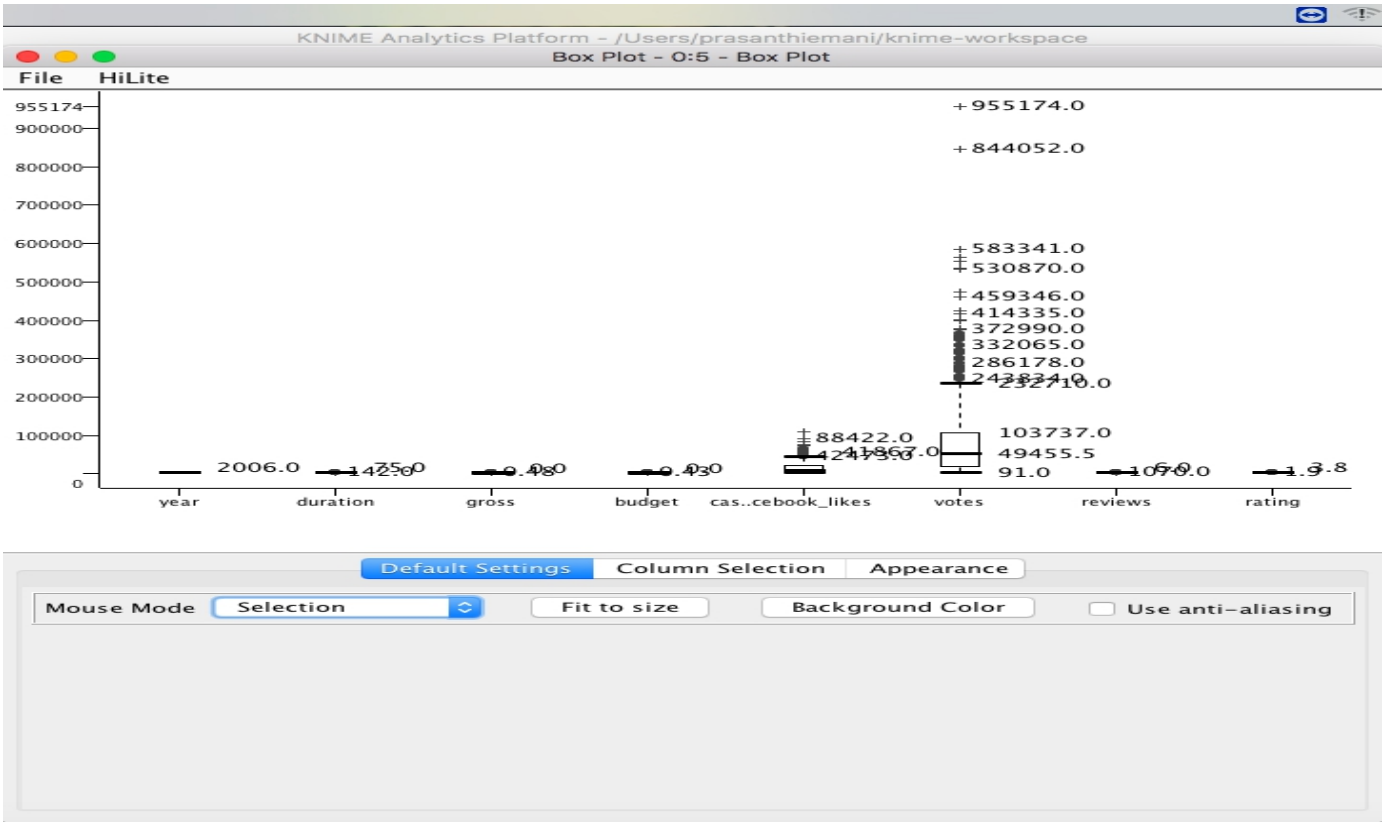
❖ **Boxplot :**

A box plot displays robust statistical parameters: minimum, lower quartile, median, upper quartile, and maximum. A box plot for one numerical attribute is constructed in the following way: The box itself goes from lower quartile (Q1) to upper quartile (Q3). Median is drawn as horizontal bar inside box. Distance between Q1 and Q3 is called interquartile range (IQR). Above and below box are so-called whiskers. They are drawn at minimum and maximum value as horizontal bars and are connected with the box by a dotted line.

Steps :

- Make connection between normalizer and boxplot.
- View for the boxplot directly.
- We can select the specific columns for the individual boxplot through column selection.

➢ **Boxplot for the whole of dataset :**

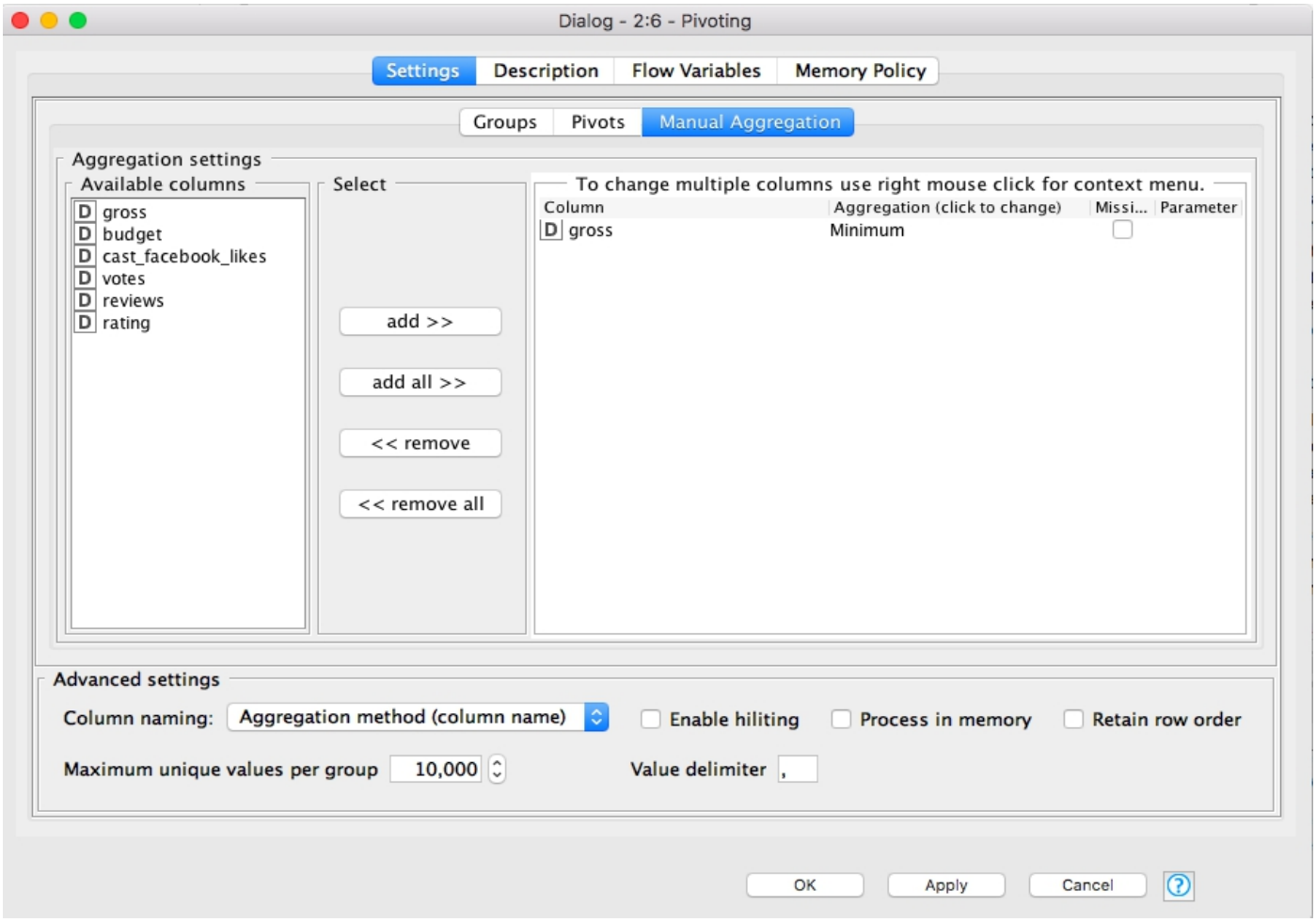


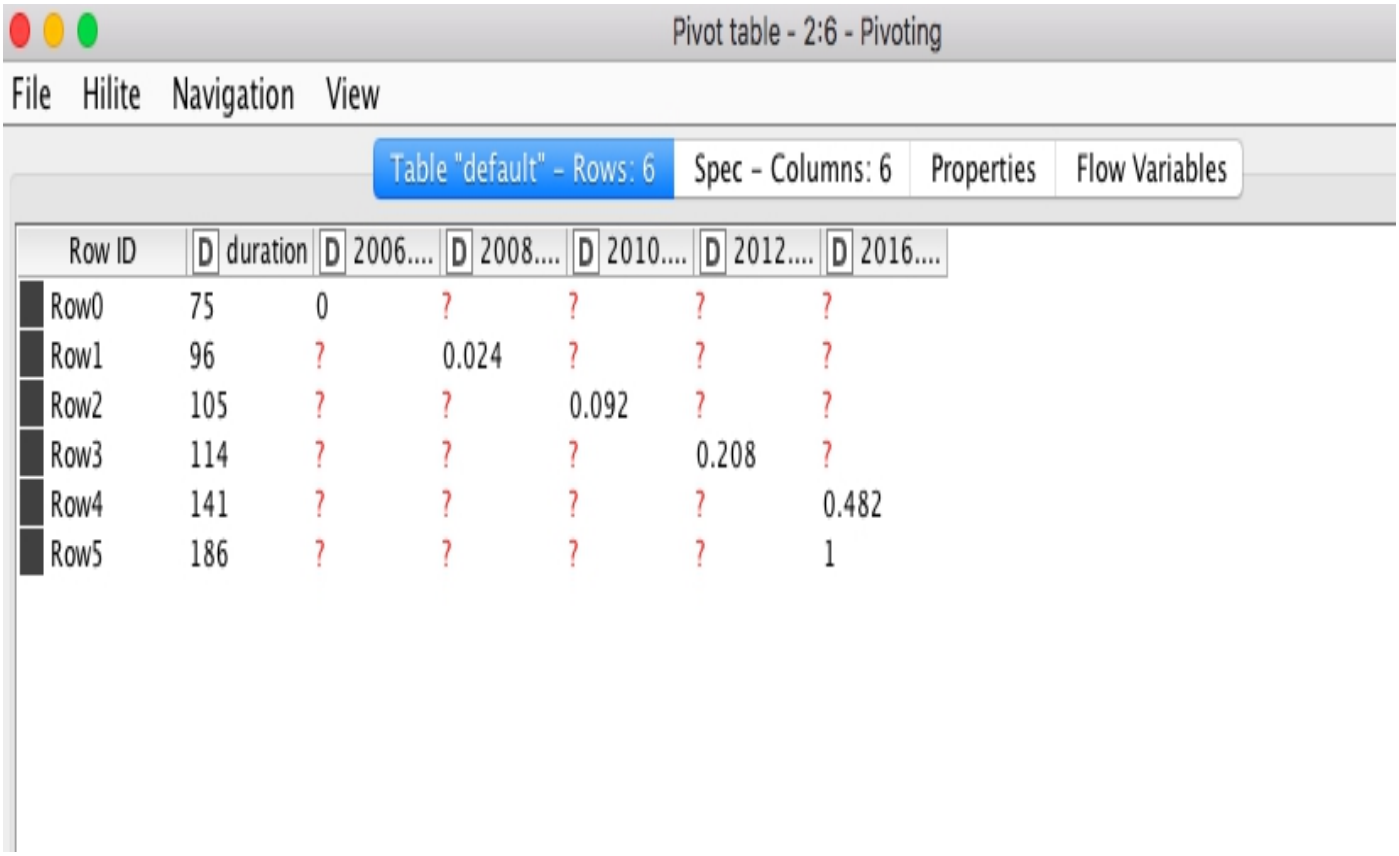
❖ **Pivoting :**

Performs a pivoting on the given input table using a selected number of columns for grouping and pivoting. The group columns will result into unique rows, whereby the pivot values turned into columns for each set of column combinations together with each aggregation method. In addition, the node returns the total aggregation (a) based on only the group columns and (b) based on only the pivoted columns resulting in a single row; optionally, with the total aggregation without pivoting.

Steps :

- Connect pivot with boxplot and have the connection between them.
- Configure pivoting with 3 different columns for same data type for :
  Groups – duration
  Pivots - year
  Manual Aggregation – gross
- Execute pivoting and checkout for the changes in the table.





Row ID	D duration	D 2006....	D 2008....	D 2010....	D 2012....	D 2016....
Row0	75	0	?	?	?	?
Row1	96	?	0.024	?	?	?
Row2	105	?	?	0.092	?	?
Row3	114	?	?	?	0.208	?
Row4	141	?	?	?	?	0.482
Row5	186	?	?	?	?	1

❖ **Joiner :**

A Joiner node joins two tables together on one or more common key values. Possible join modes: inner join, left outer join, right outer join, full outer join. Two tabs: "Joiner Settings" and "Column Selection". "Joiner Settings" defines the parameters for the join operation: join mode and column keys. "Column Selection" sets which columns to keep and/or drop and strategies to deal with duplicate columns.

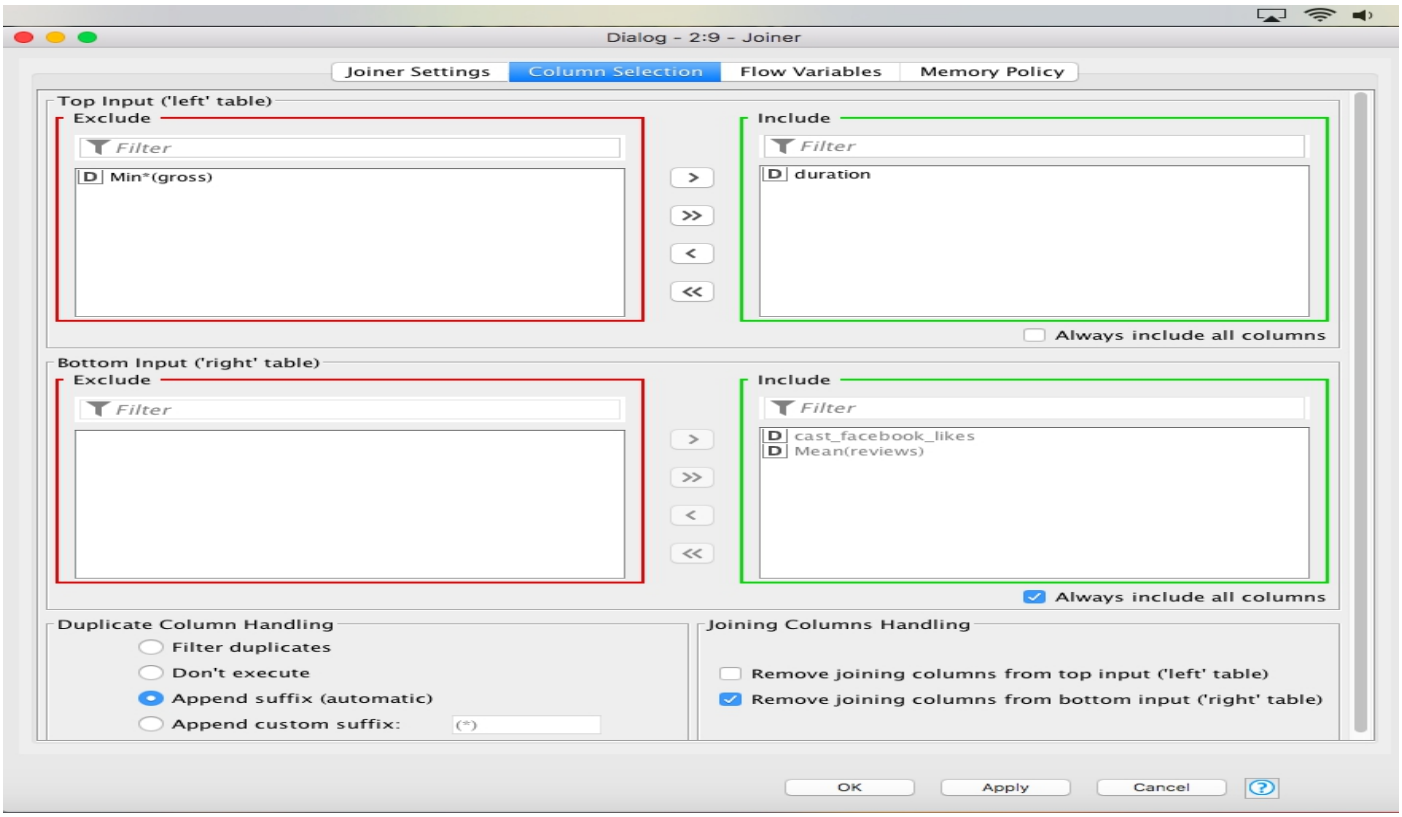
Steps :

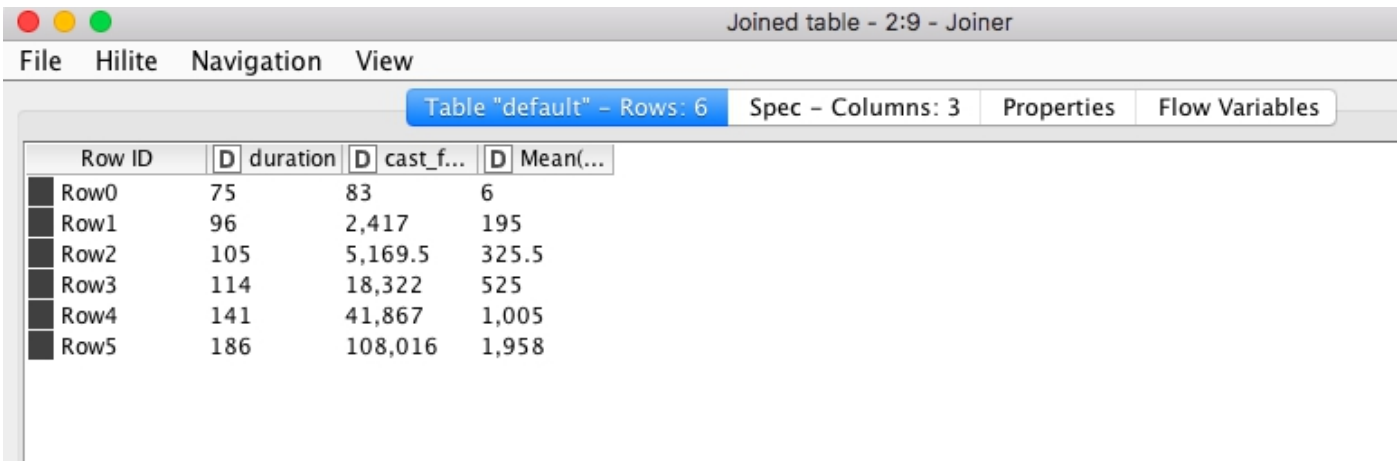
- Have the connection between joiner and pivot so that it is easy to analyse.
- Configure joiner with columns if necessary as :

➢ Top Input (left table):
Include : duration

➢ Bottom Input(right table):
Include :
→ cast_facebook_likes
→ Mean

- Execute the joiner and check for the final table.





Row ID	duration	cast_f...	Mean(...
Row0	75	83	6
Row1	96	2,417	195
Row2	105	5,169.5	325.5
Row3	114	18,322	525
Row4	141	41,867	1,005
Row5	186	108,016	1,958

❖ **Group By :**

Groups the rows of a table by the unique values in the selected group columns. A row is created for each unique set of values of the selected group column. The remaining columns are aggregated based on the specified aggregation settings. The output table contains one row for each unique value combination of the selected group columns.

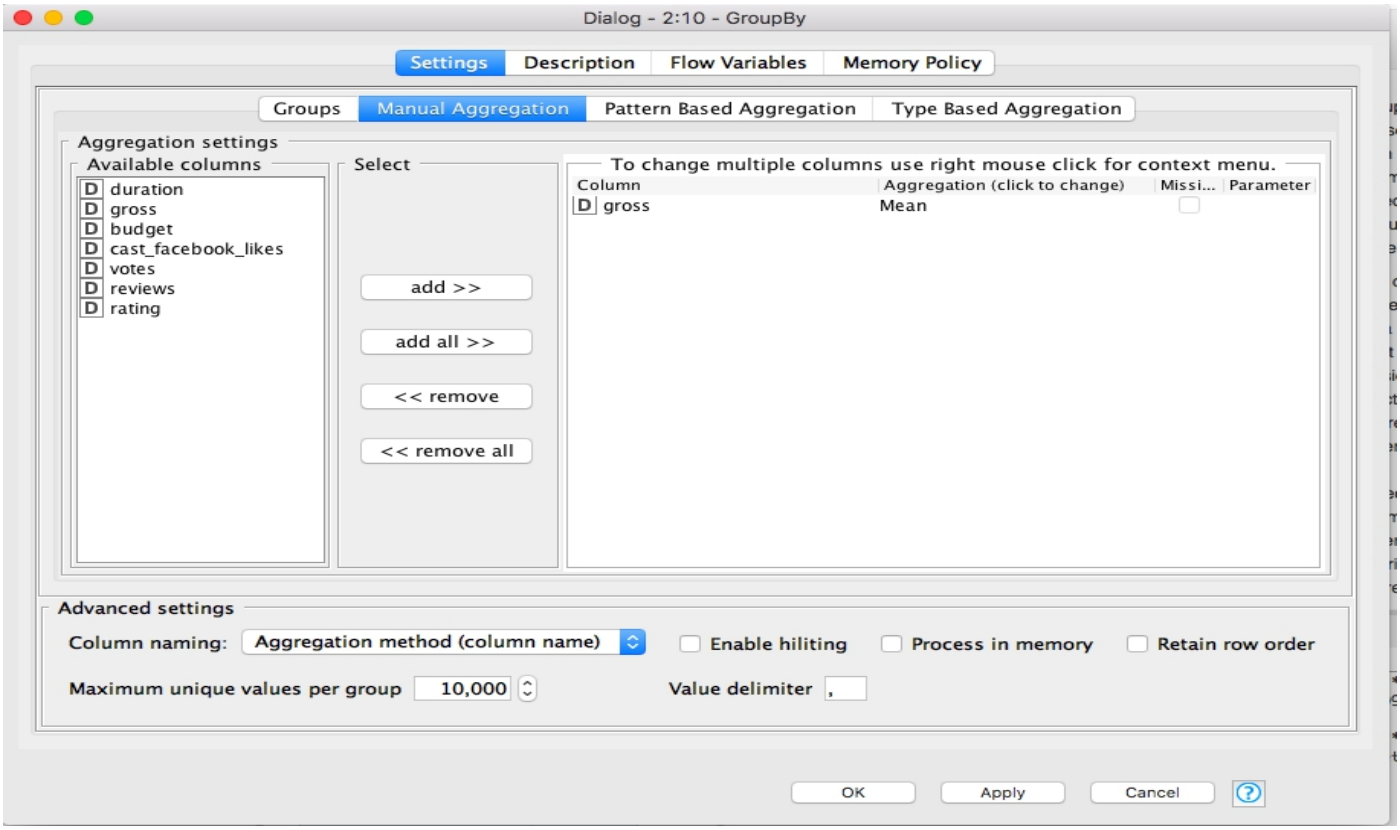
Steps :

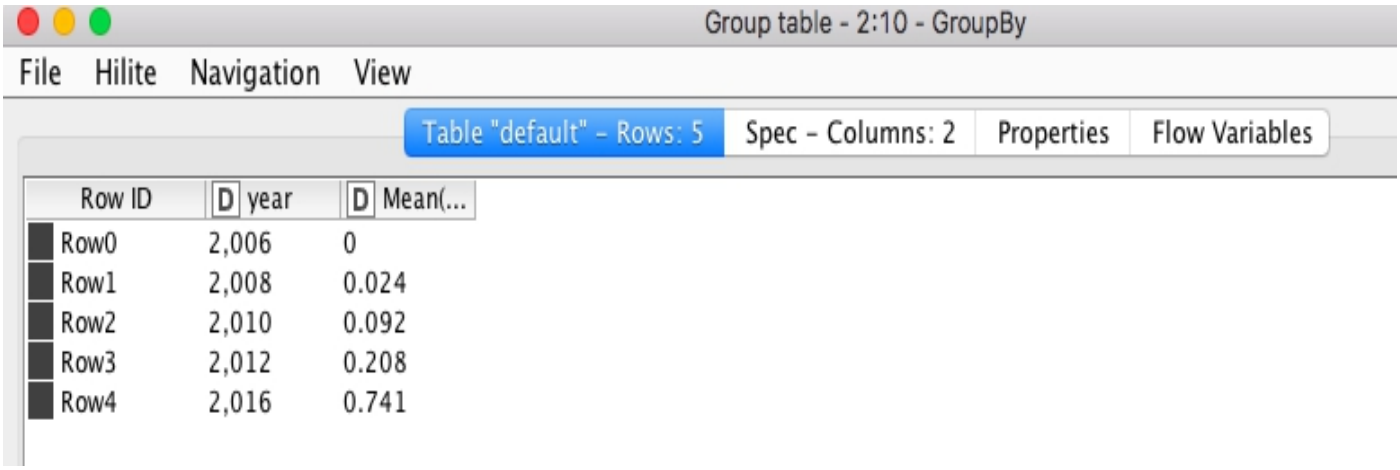
- Make the connection to group by with the boxplot directly.
- Configure group by as the following :

➢ Groups : year

➢ Manual Aggregation : gross

- Execute the group by and check for the analysed table.





RESULT :

Thus, the operations that will be done on the table for the better access of the data will be done in this way where by applying normalization, pivoting and group by techniques.