Day_1

1) The intervals and corresponding frequencies are as follows. age frequency

1-5. 200

5-15 450

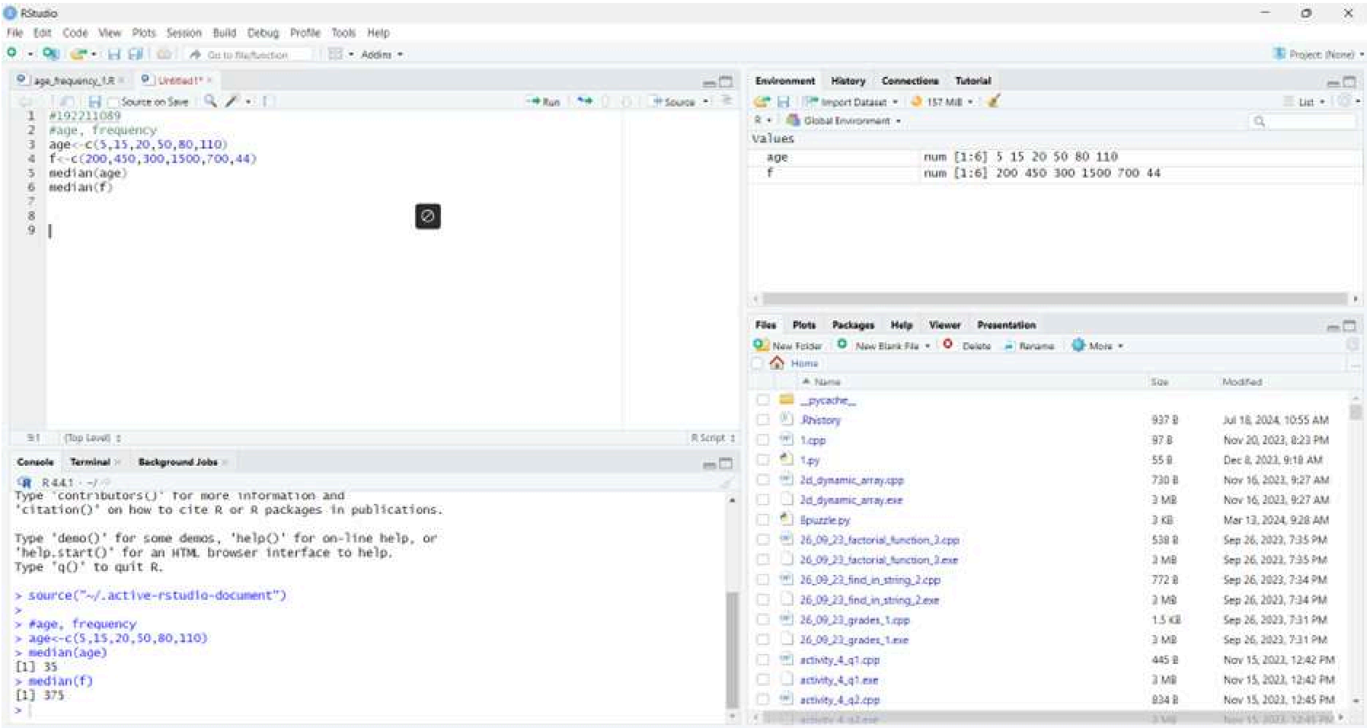
15-20 300

20-50 1500

50-80 700

80-110 44

Compute an approximate median value for the data



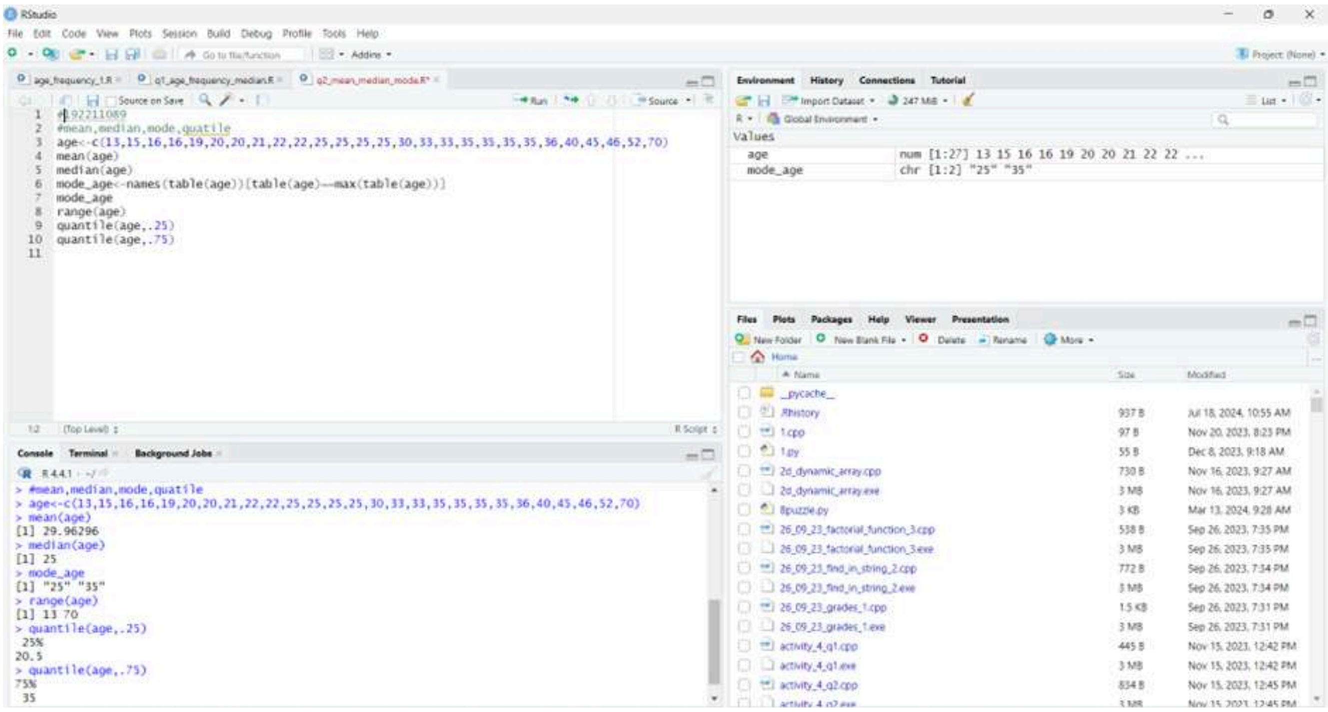
2) **Suppose that the data for analysis includes the attribute age. The age values for the data tuples are (in increasing order) 13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70.**

**(a) What is the mean of the data? What is the median?**

**(b) What is the mode of the data? Comment on the data's modality (i.e., bimodal, trimodal, etc.).**

**(c) What is the midrange of the data?**

**(d) Can you find (roughly) the first quartile (Q1) and the third quartile (Q3) of the data?**



**3&4 in same code:**

3) **Data Preprocessing: Reduction and Transformation**

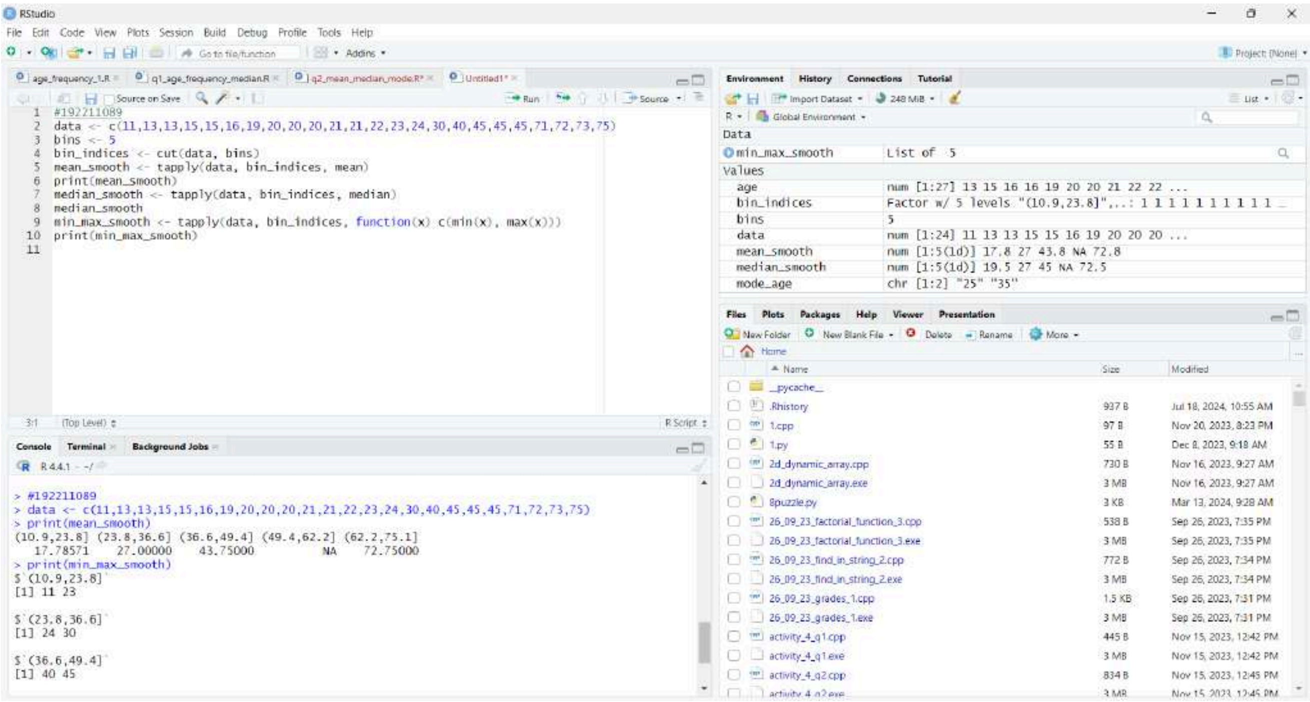
**Use the two methods below to normalize the following group of data: 200, 300, 400, 600, 1000 (a) min-max normalization by setting min = 0 and max = 1 (b) z-score normalization**

**4.Data:11,13,13,15,15,16,19,20,20,20,21,21,22,23,24,30,40,45,45,45,71,72,73,75**

**a) Smoothing by bin mean**

**b) Smoothing by bin median**

**c) Smoothing by bin boundaries**

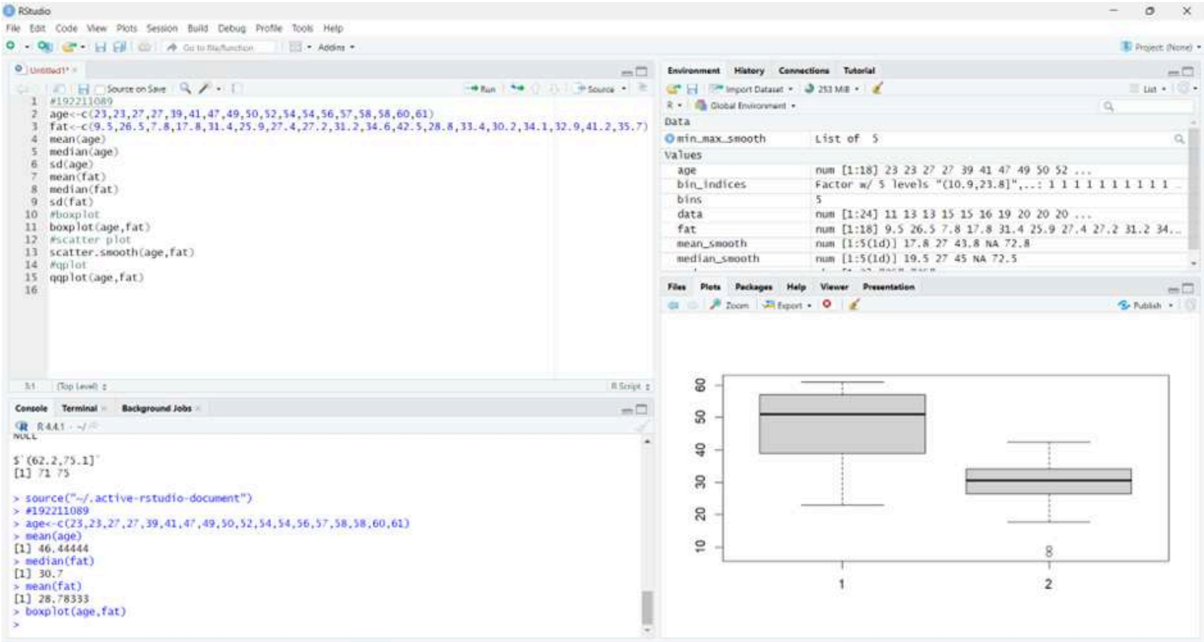



**5) Suppose that a hospital tested the age and body fat data for 18 randomly selected adults with the following results:**

**(a) Calculate the mean, median, and standard deviation of age and %fat.**

**(b) Draw the boxplots for age and %fat.**

**(c) Draw a scatter plot and a q-q plot based on these two variables.**

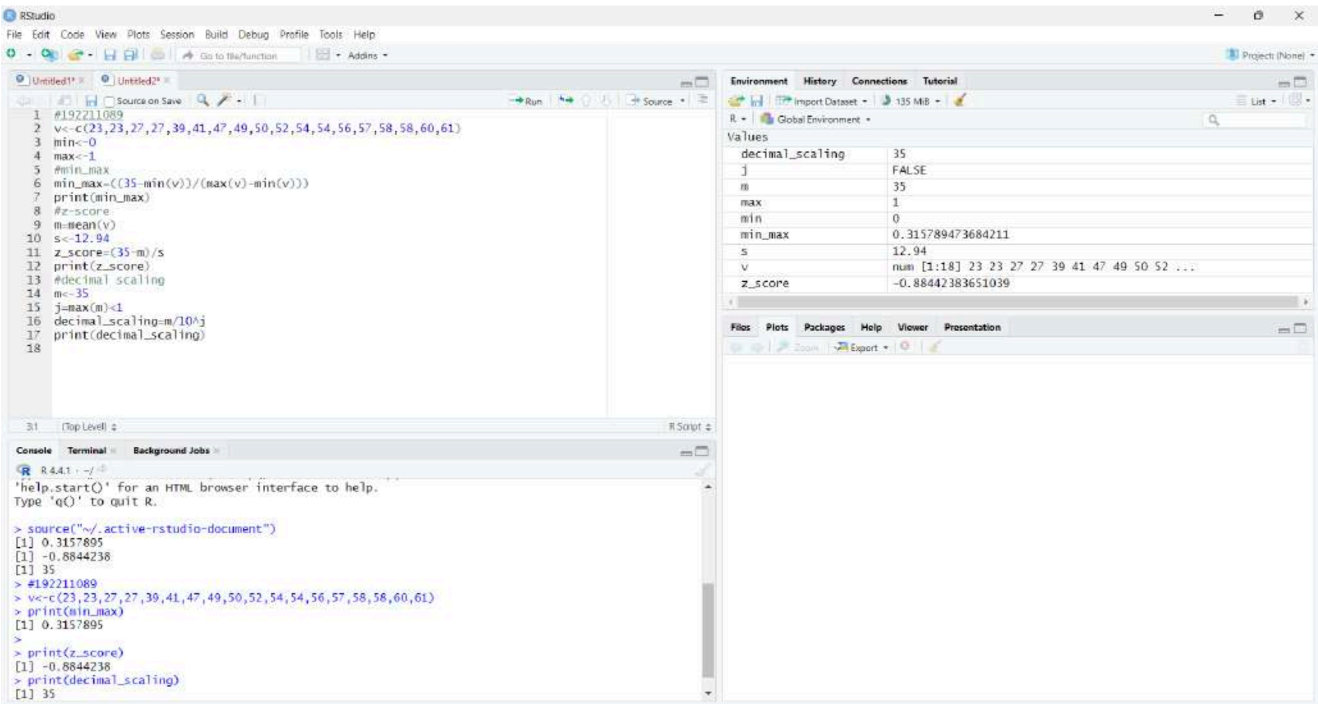


**6) Suppose that a hospital tested the age and body fat data for 18 randomly selected adults with the following results:**

**(i) Use min-max normalization to transform the value 35 for age onto the range [0.0, 1.0].**

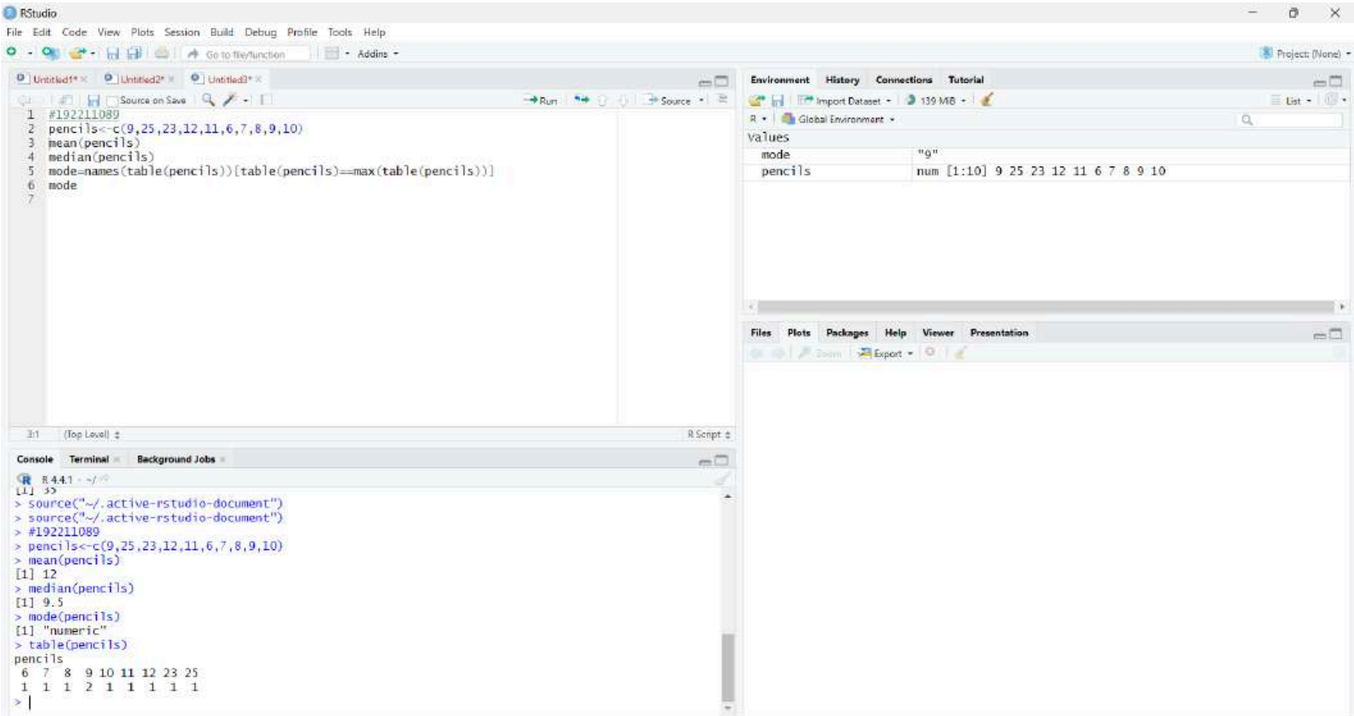
**(ii) Use z-score normalization to transform the value 35 for age, where the standard deviation of age is 12.94 years.**

**(iii) Use normalization by decimal scaling to transform the value 35 for age. Perform the above functions using R – tool**



**7) The following values are the number of pencils available in the different boxes. Create a vector and find out the mean, median and mode values of set of pencils in the given data.**

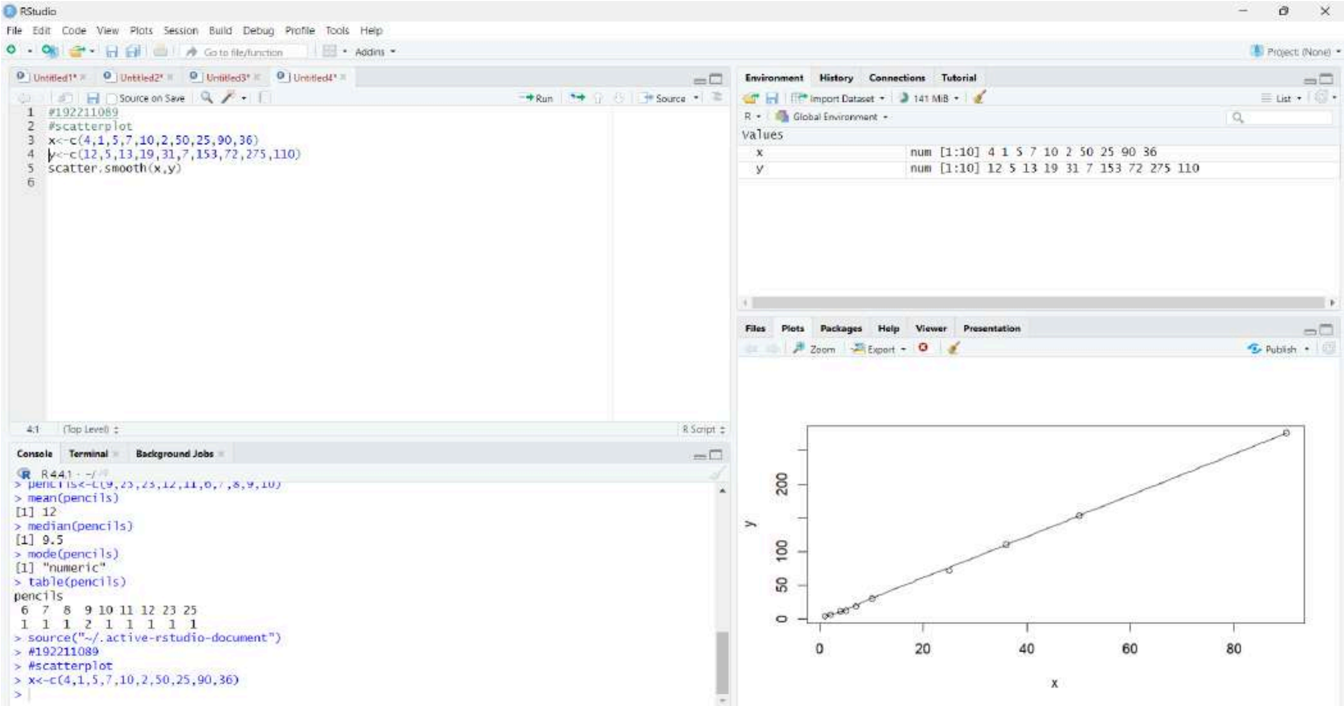
| Box1 | Box2 | Box3 | Box4 | Box5 | Box6 | Box7 | Box8 | Box9 | Box 10 |
|---|---|---|---|---|---|---|---|---|---|
| 9 | 25 | 23 | 12 | 11 | 6 | 7 | 8 | 9 | 10 |



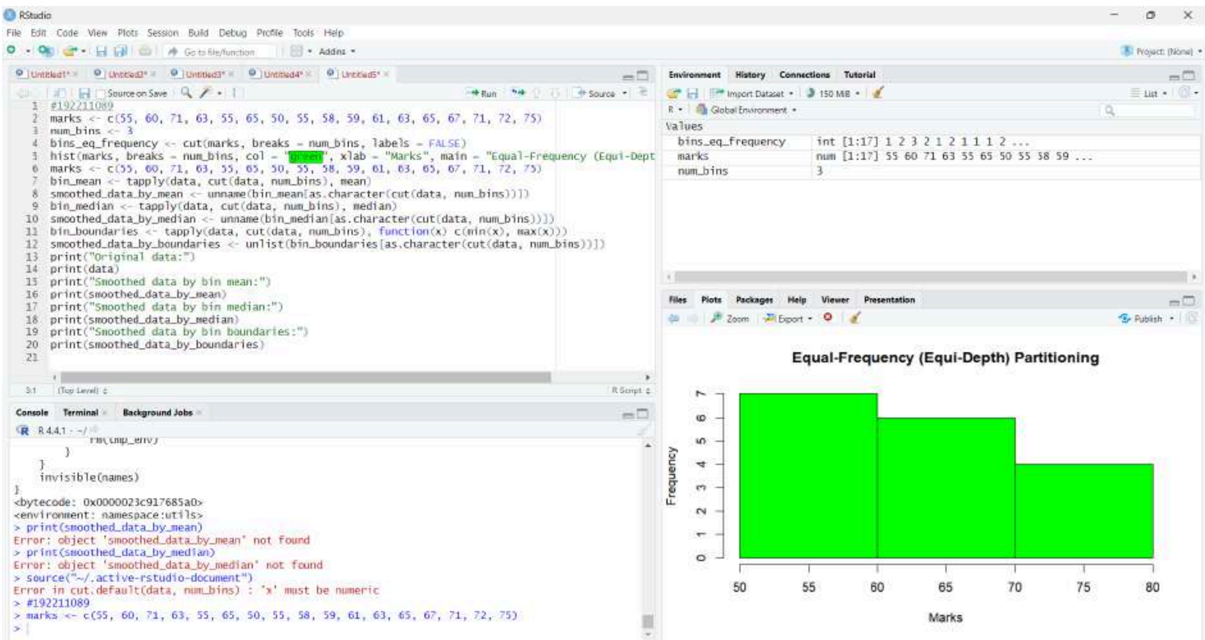
**8) . The following table would be plotted as (x,y) points, with the first column being the x values as several mobile phones sold and the second column being the y values as money. To use the scatter plot for how many mobile phones are sold.**

**x:4 1 5 7 10 2 50 25 90 36**

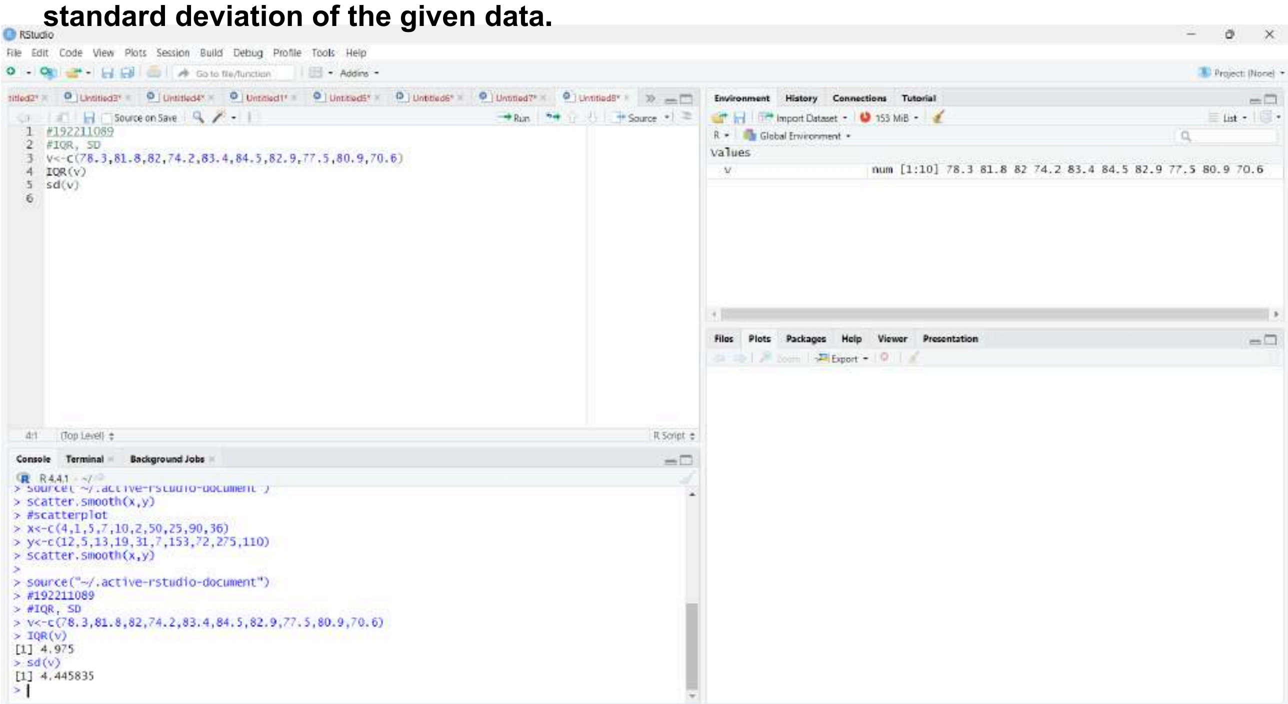
**y:12 5 13 19 31 7 153 72 275 110**



**9) Implementing the R script using marks scored by a student in his model exam has been sorted as follows: 55, 60, 71, 63, 55, 65, 50, 55,58,59,61,63,65,67,71,72,75. They are partitioned into three bins using each of the following methods. Plot the data points using the histogram.a) equal-frequency (equi-depth) partitioning (b) equal-width partitioning**

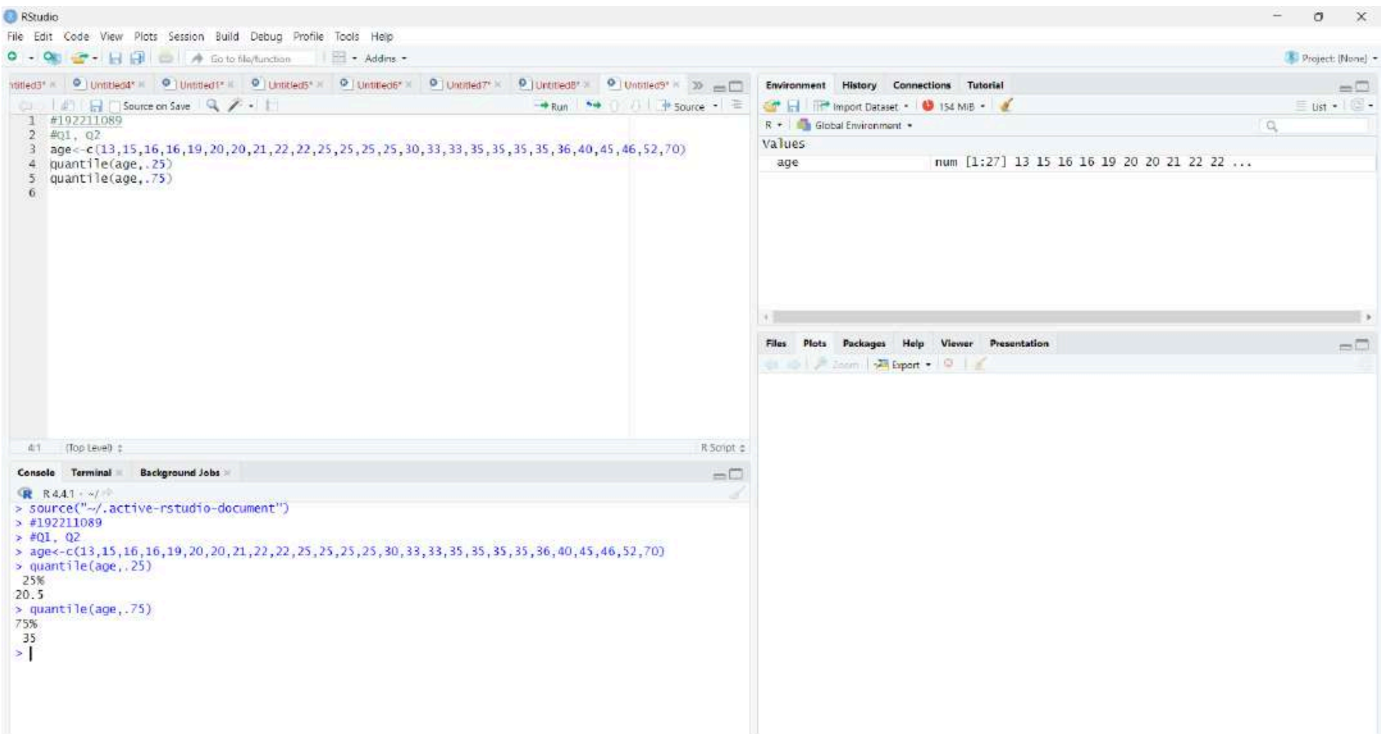


10) **Suppose that the speed car is mentioned in different driving style.Regular 78.3 81.8 82 74.2 83.4 84.5 82.9 77.5 80.9 70.6 SpeedCalculate the Inter quantile and standard deviation of the given data.**



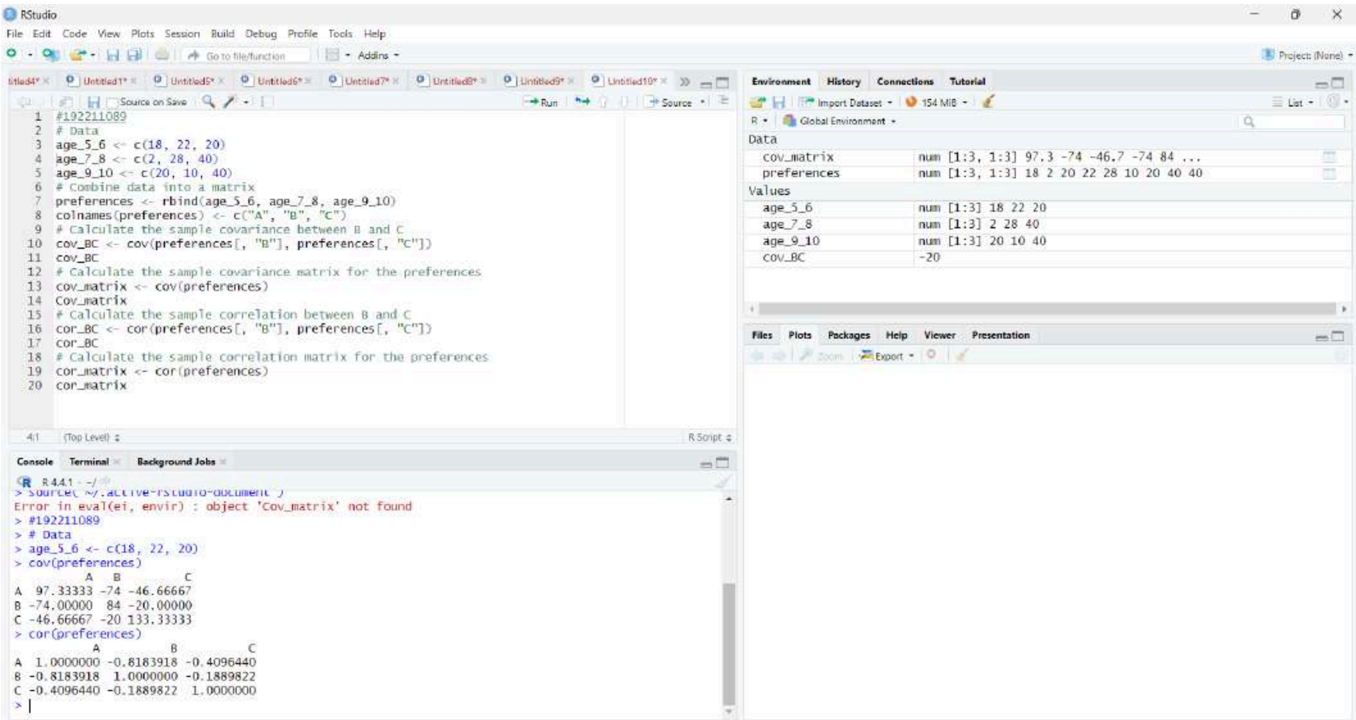

11) Suppose that the data for analysis includes the attribute age. The age values for the data tuples are (in increasing order) 13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70.

Can you find (roughly) the first quartile (Q1) and the third quartile (Q3) of the data?



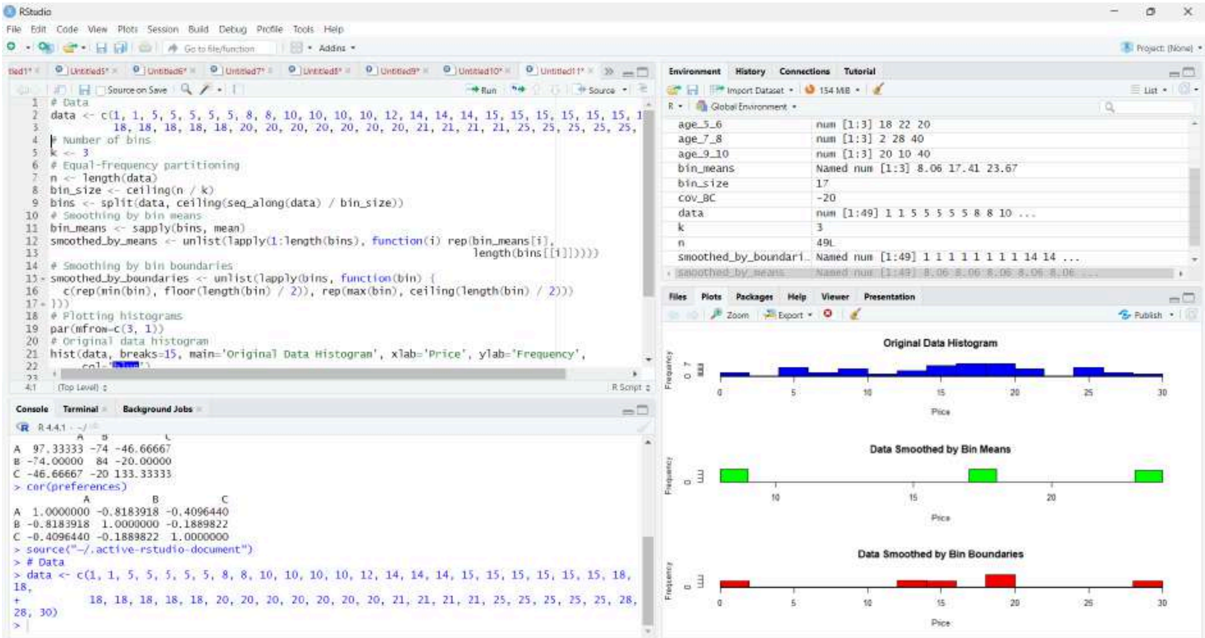
DAY_02

**1) 1.Covariance and correlation Children of three ages are asked to indicate their preference**
**for three photographs of adults. Do the data suggest that there is a significant relationship**
**between age and photograph preference? What is wrong with this study?**
**Photograph: Age**
**of child A B C 5-6 years: 18 22 20 7-8 years: 2 28 40 9-10 years: 20 10 40**
**1.Use cov() to calculate the sample covariance between B and C.**
**2. Use another call to cov() to calculate the sample covariance matrix for the preferences.**
**3. Use cor() to calculate the sample correlation between B and C.**
**4. Use another call to cor() to calculate the sample correlation matrix for the**

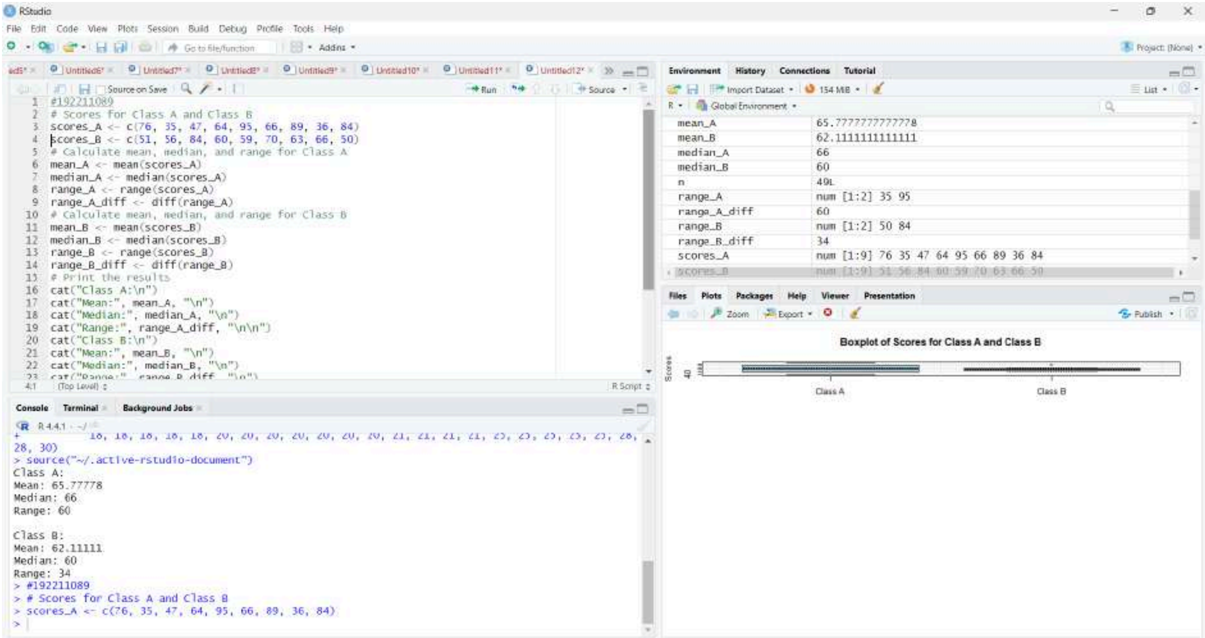



**2) Imagine that you have selected data from the All Electronics data warehouse for analysis.**
**The data set will be huge! The following data are a list of All Electronics prices for commonly**
**sold items (rounded to the nearest dollar). The numbers have been sorted: 1, 1, 5, 5, 5, 5, 5,**
**8, 8, 10, 10, 10, 10, 12, 14, 14, 14, 15, 15, 15, 15, 15, 15, 18, 18, 18, 18, 18, 8, 18, 20, 20,**
**20, 20, 20, 20, 20, 21, 21, 21, 21, 25, 25, 25, 25, 25, 28, 28, 30 the dataset using an**

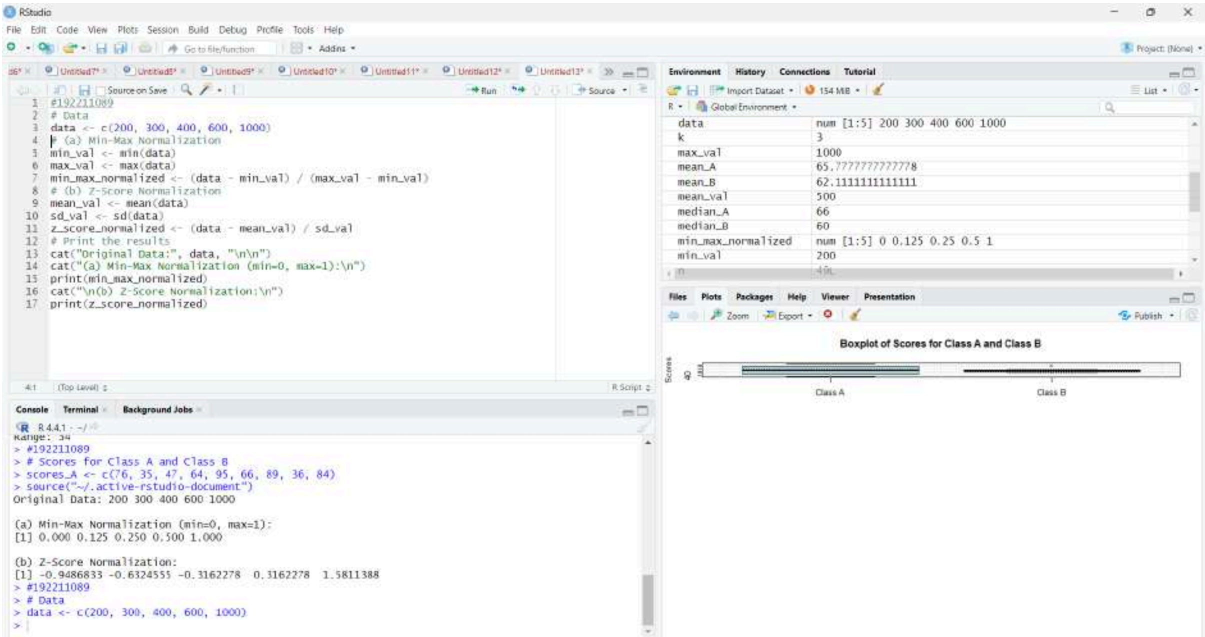
**equal-frequency partitioning method with bin equal to 3 (ii) apply data smoothing using bin**
**means and bin boundary. (iii) Plot Histogram for the above frequency division**



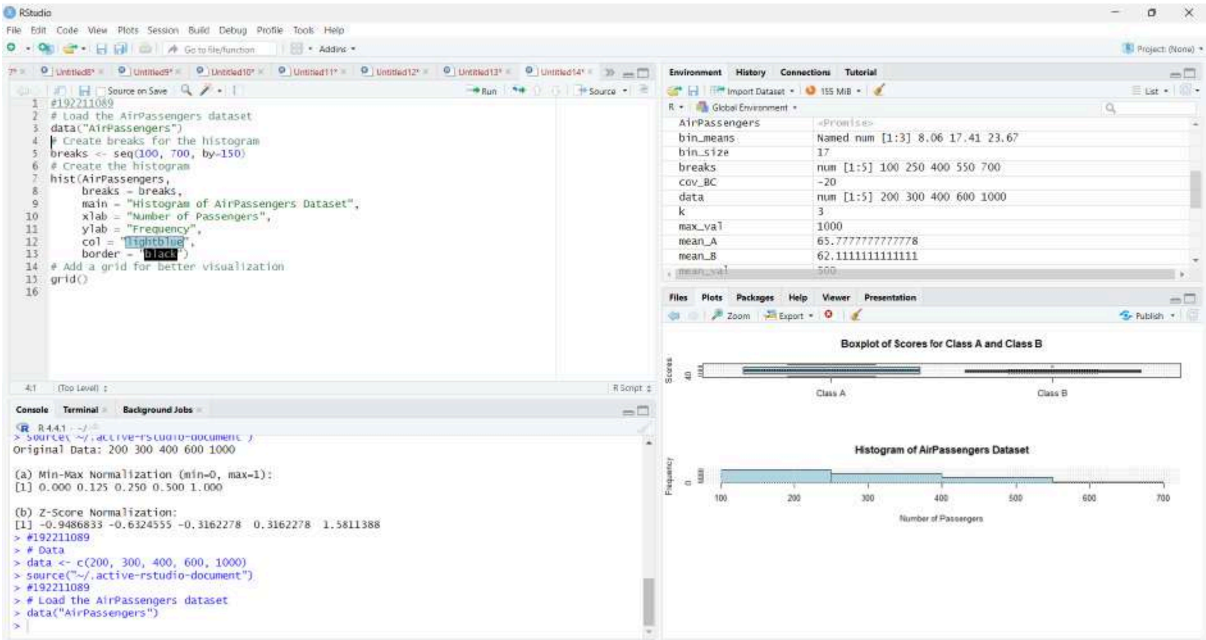

**3) 3)3.Two Maths teachers are comparing how their Year 9 classes performed in the end of**
**year exams. Their results are as follows: Class A: 76, 35, 47, 64, 95, 66, 89, 36, 8476,35,47,64,95,66,89,36,84 Class B: 51, 56, 84, 60, 59, 70, 63, 66, 5051,56,84,60,59,70,63,66,50 (i) Find which class had scored higher mean, median and range. (ii) Plot above in boxplot and give the inferences Class B: 51, 56, 84, 60, 59, 70, 63,**
**66, 5051,56,84,60,59,70,63,66,50**



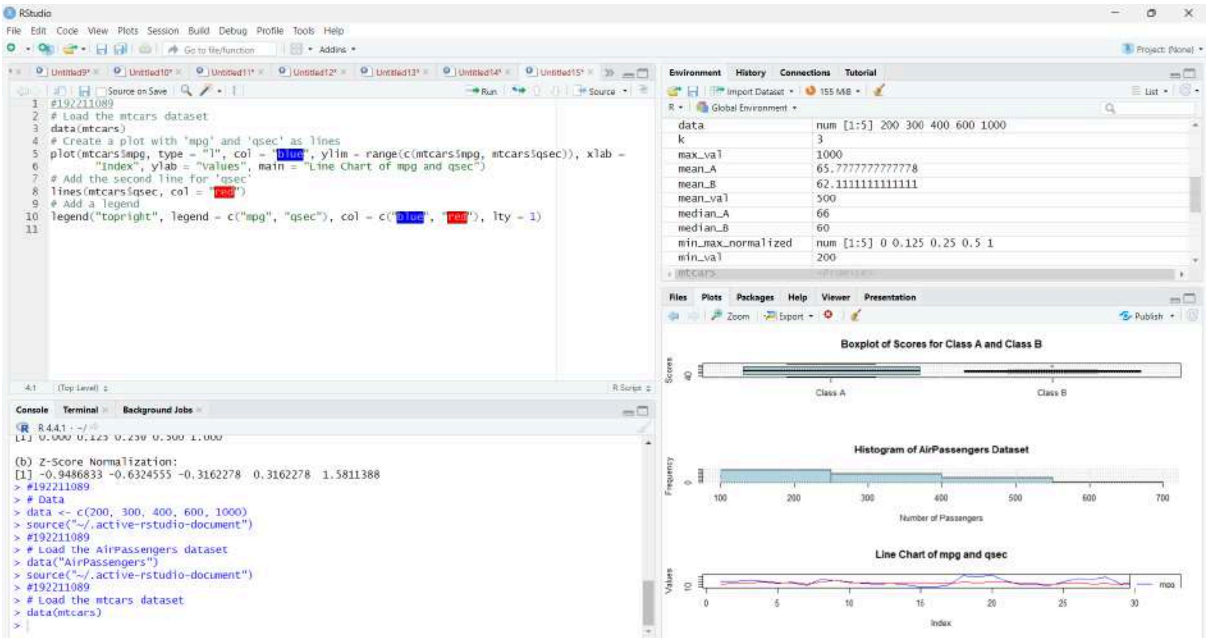

**4) Let us consider one example to make the calculation method clear. Assume that the minimum and maximum values for the feature F are $50,000 and $100,000 correspondingly.**
**It needs to range F from 0 to 1. By min-max normalization, v = $80, b) Use the two methods below to normalize the following group of data: 200, 300, 400, 600, 1000**
**(a) min-max normalization by setting min = 0 and max = 1 (b) z-score normalization**



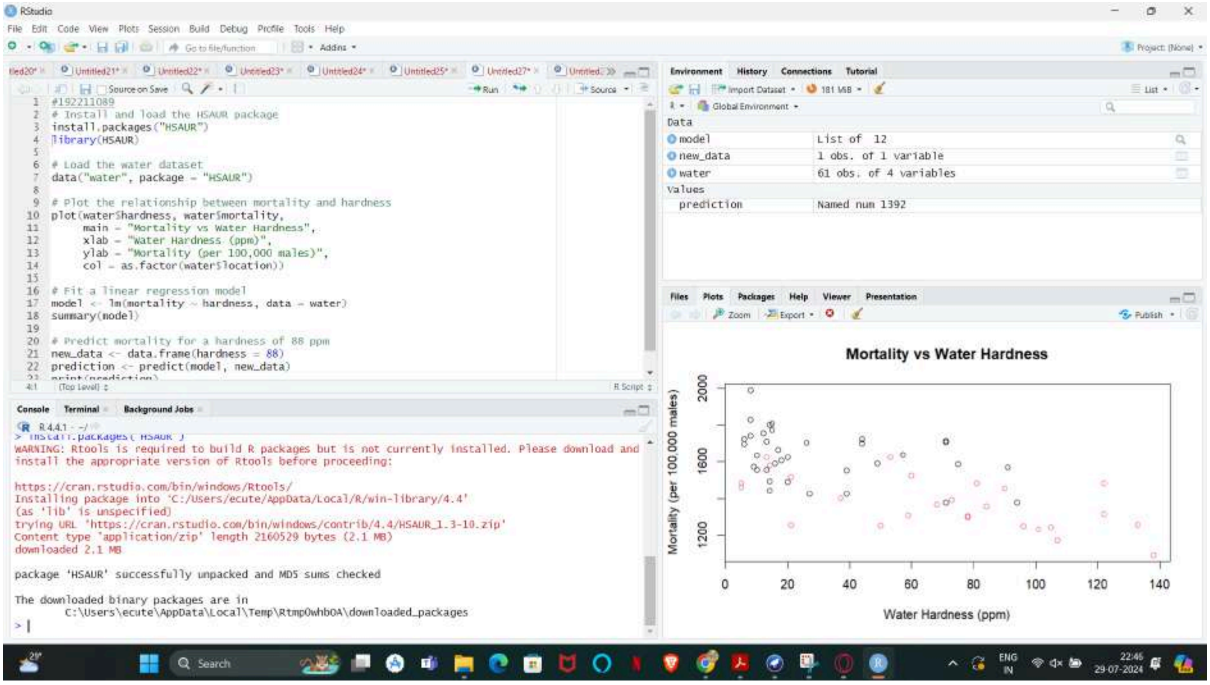
**5) Make a histogram for the “AirPassengers “dataset, starting at 100 on the x-axis, and from**
**values 200 to 700, make the bins 150 wide**



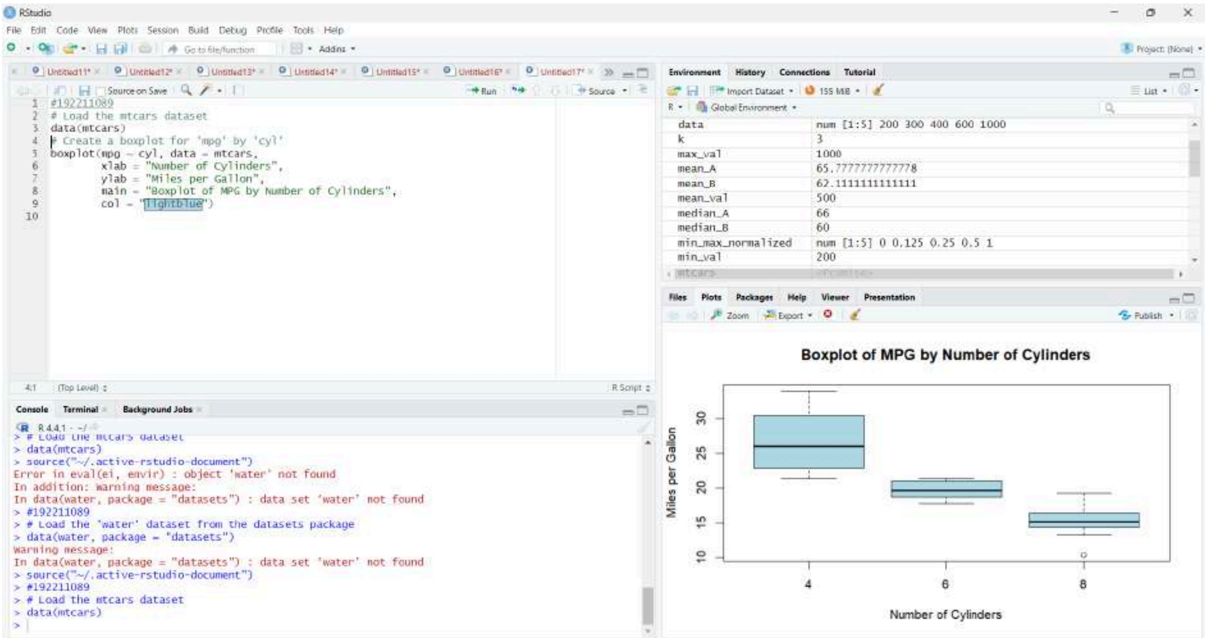

**6) Obtain Multiple Lines in a Line Chart using a single Plot Function in Use attributes“mpg” and“sec” of the dataset “mtcars”**
**Code:**



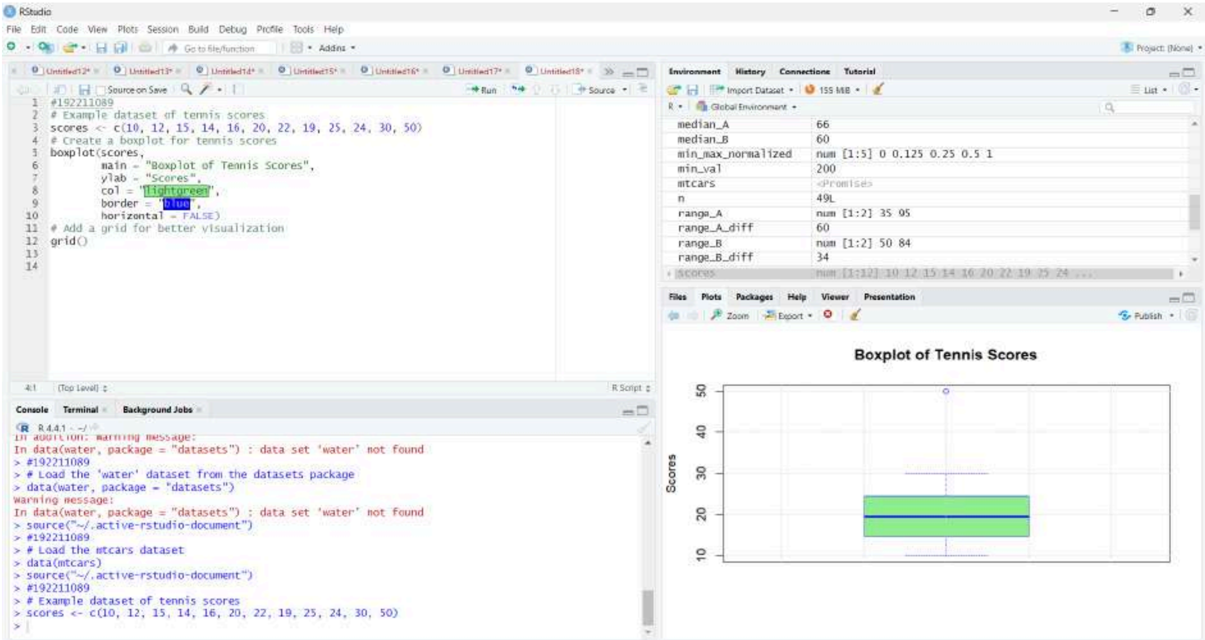
**7) Download the Dataset "water" From the R dataset Link. Find out whether there is a linear**
**relation between attributes"mortality" and"hardness" by plot function.Fit**



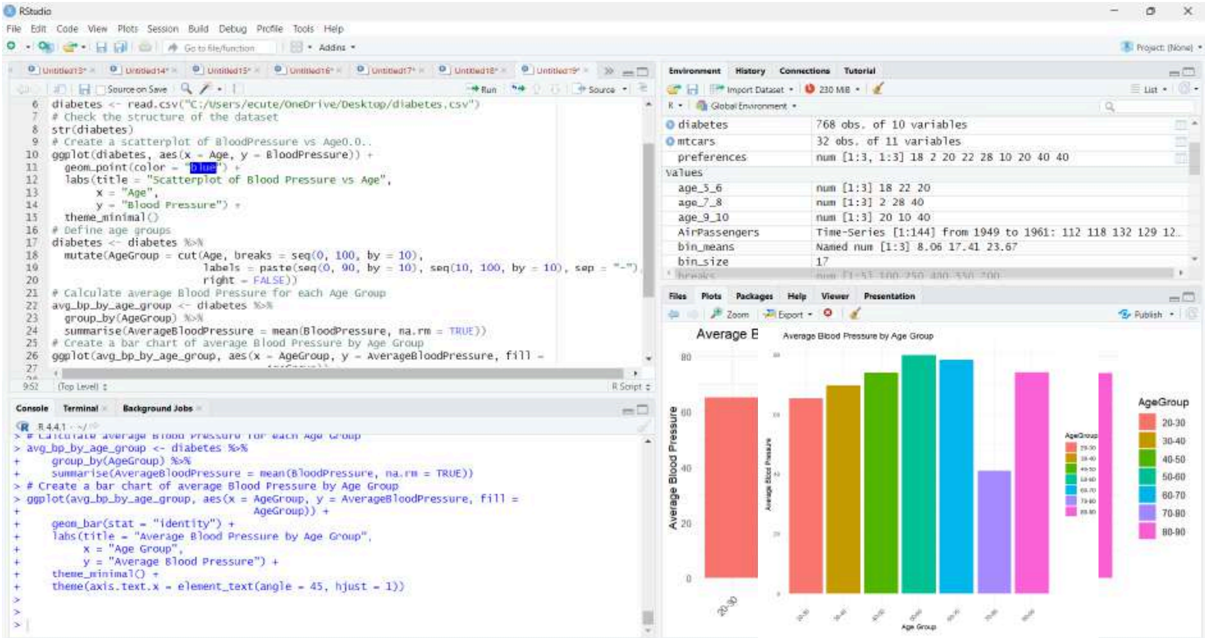

**8) 8.Create a Boxplot graph for the relation between "mpg"(miles per gallon) and "cyl"(number**
**of Cylinders) for the dataset "mtcars" available in R Environment**
**Code:**



**9) Assume the Tennis coach wants to determine if any of his team players are scoring outliers. To visualize the distribution of points scored by his players, then how can he decide**
**to develop the box plot? Give a suitable example using the Boxplot visualization technique.**
**Code:**

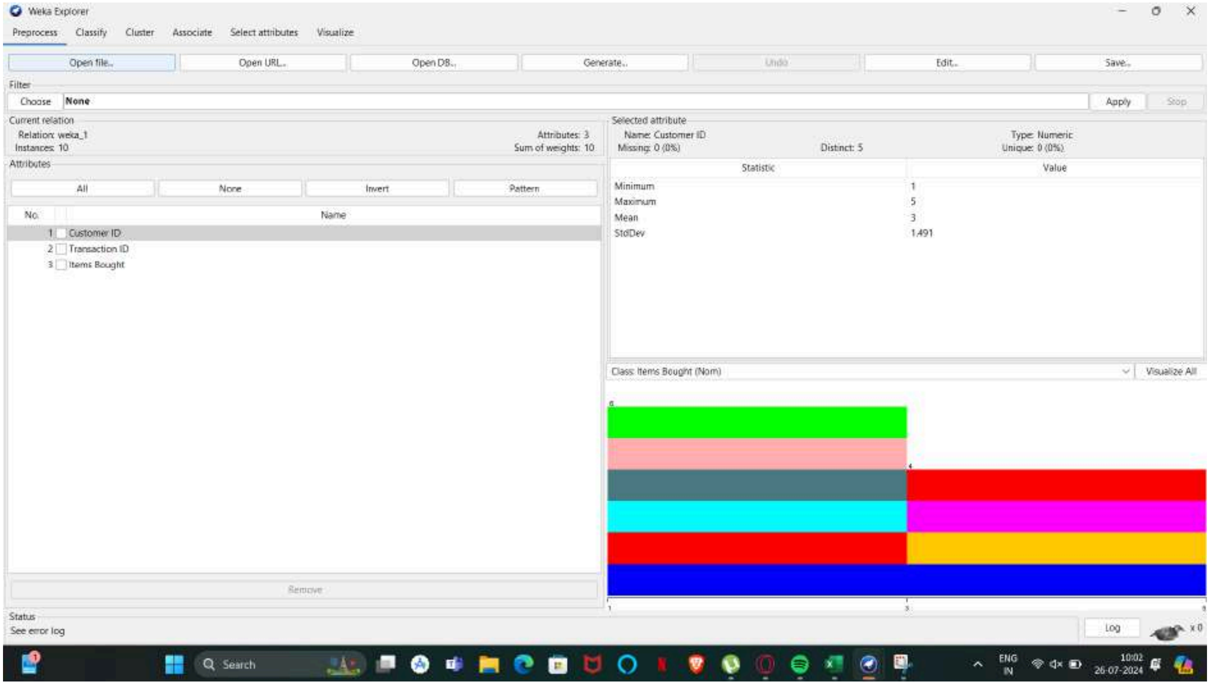


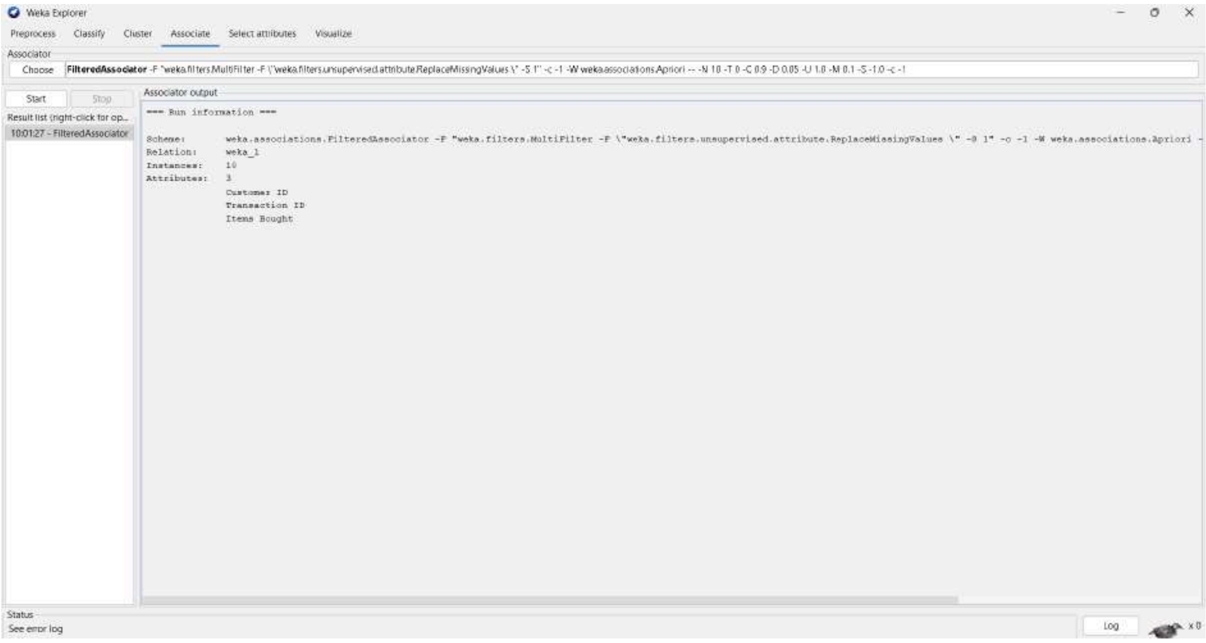

**10) Implement using R language in which age group of people are affected byblood pressure**
**based on the diabetes dataset show it using scatterplot and bar chart (that is BloodPressure**
**vs Age using dataset “diabetes.csv”)**
**Code:**




## Day_3

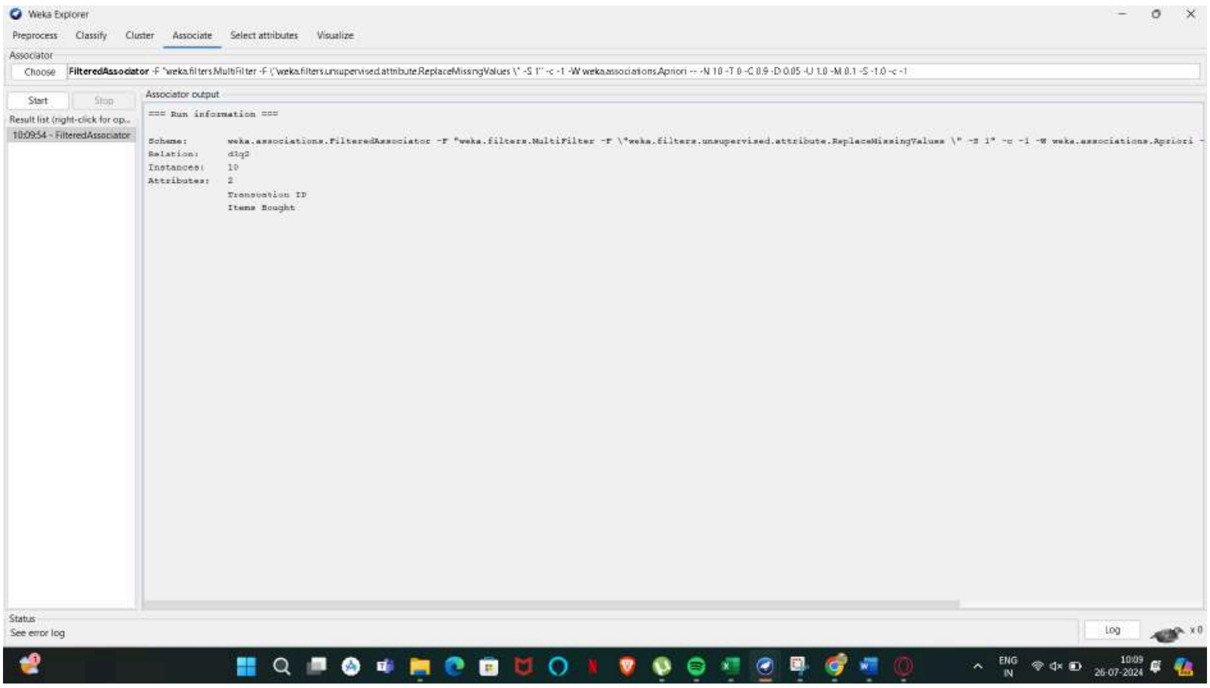
1) **Consider the data set and perform the Apriori Algorithm and FP algorithm support:3 and confidence=50%**

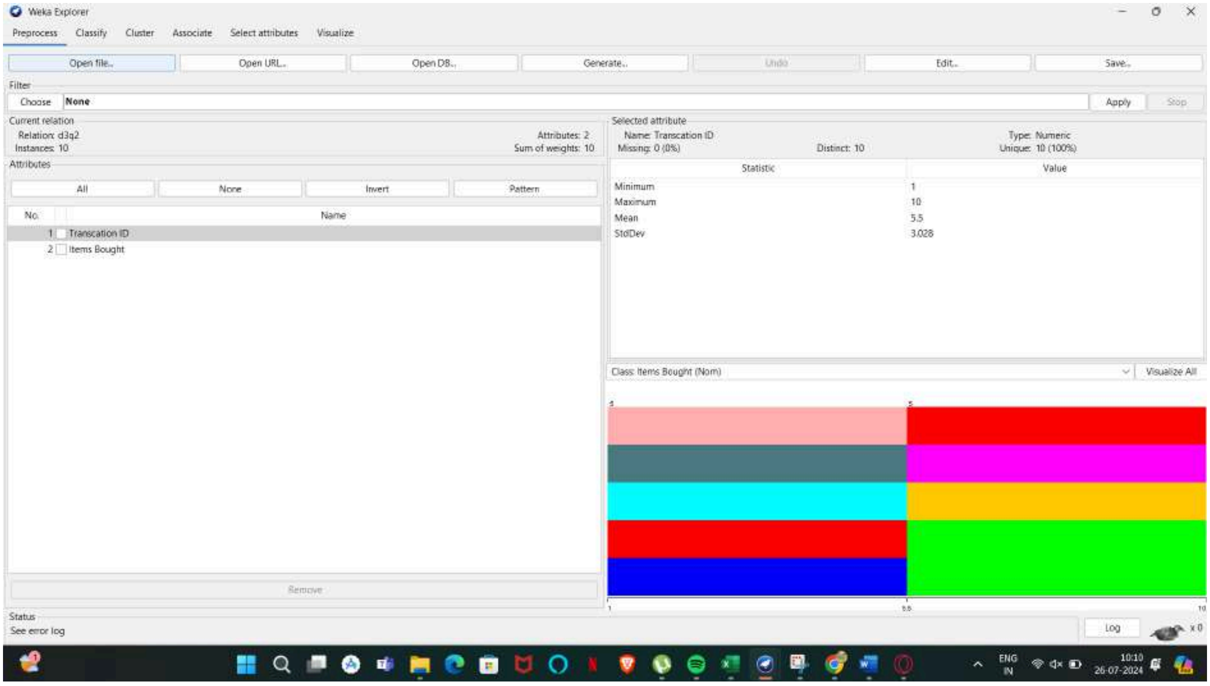

Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Open file...
Open URL...
Open DB...
Generate...
Undo
Edit...
Save...
Filter
Choose
None
Apply
Stop
Current relation
Relation: weka_1
Instances: 10
Attributes: 3
Sum of weights: 10
Attributes
All
None
Invert
Pattern
No.
Name
1 Customer ID
2 Transaction ID
3 Items Bought
Remove
Selected attribute
Name: Customer ID
Missing: 0 (0%)
Distinct: 5
Type: Numeric
Unique: 0 (0%)
Statistic
Value
Minimum 1
Maximum 5
Mean 3
StdDev 1.491
Class: Items Bought (Nom)
Visualize All
Status
See error log
Log
x 0


Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Associator
Choose
FilteredAssociator -F "weka.filters.MultiFilter -F \"weka.filters.unsupervised.attribute.ReplaceMissingValues \" -S 1" -c -1 -W weka.associations.Apriori -- -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-click for op...
10:01:27 - FilteredAssociator
Associator output
=== Run information ===
Scheme: weka.associations.FilteredAssociator -F "weka.filters.MultiFilter -F \"weka.filters.unsupervised.attribute.ReplaceMissingValues \" -S 1" -c -1 -W weka.associations.Apriori -
Relation: weka_1
Instances: 10
Attributes: 3
Customer ID
Transaction ID
Items Bought
Status
See error log
Log
x 0

**2)**

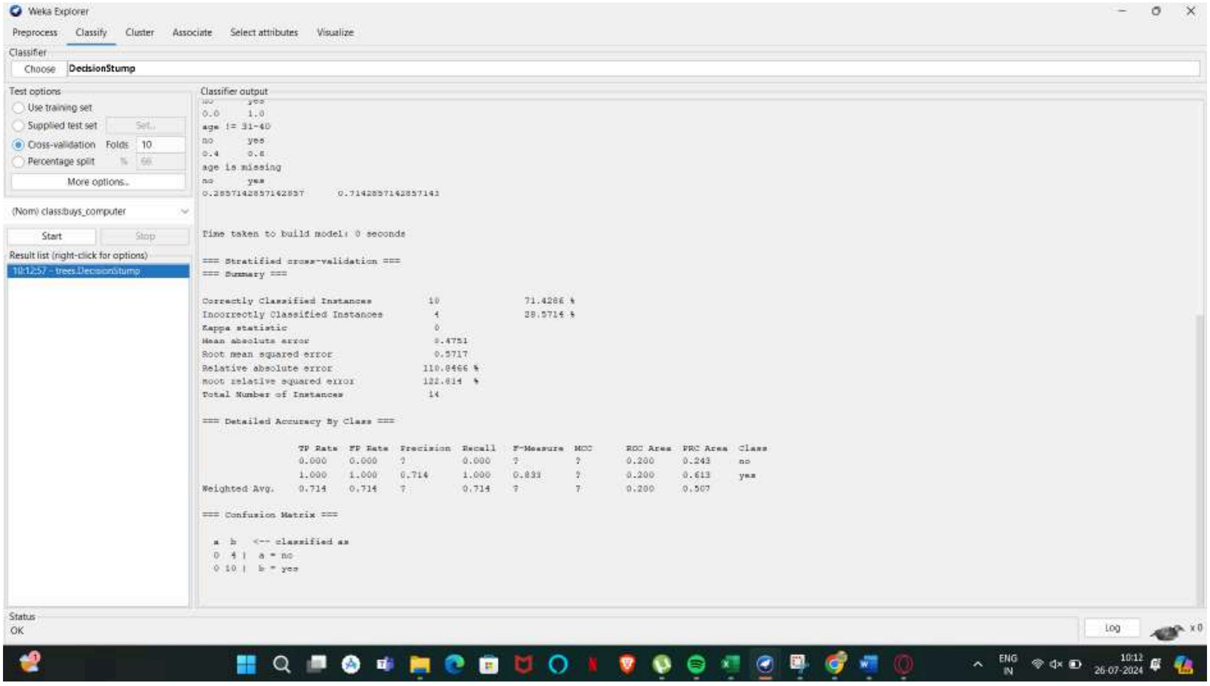
| Transaction ID | Items Bought |
|---|---|
| 1 | {Milk, Beer, Diapers} |
| 2 | {Bread, Butter, Milk} |
| 3 | {Milk, Diapers, Cookies} |
| 4 | {Bread, Butter, Cookies} |
| 5 | {Beer, Cookies, Diapers} |
| 6 | {Milk, Diapers, Bread, Butter} |
| 7 | {Bread, Butter, Diapers} |
| 8 | {Beer, Diapers} |
| 9 | {Milk, Diapers, Bread, Butter} |
| 10 | {Beer, Cookies} |

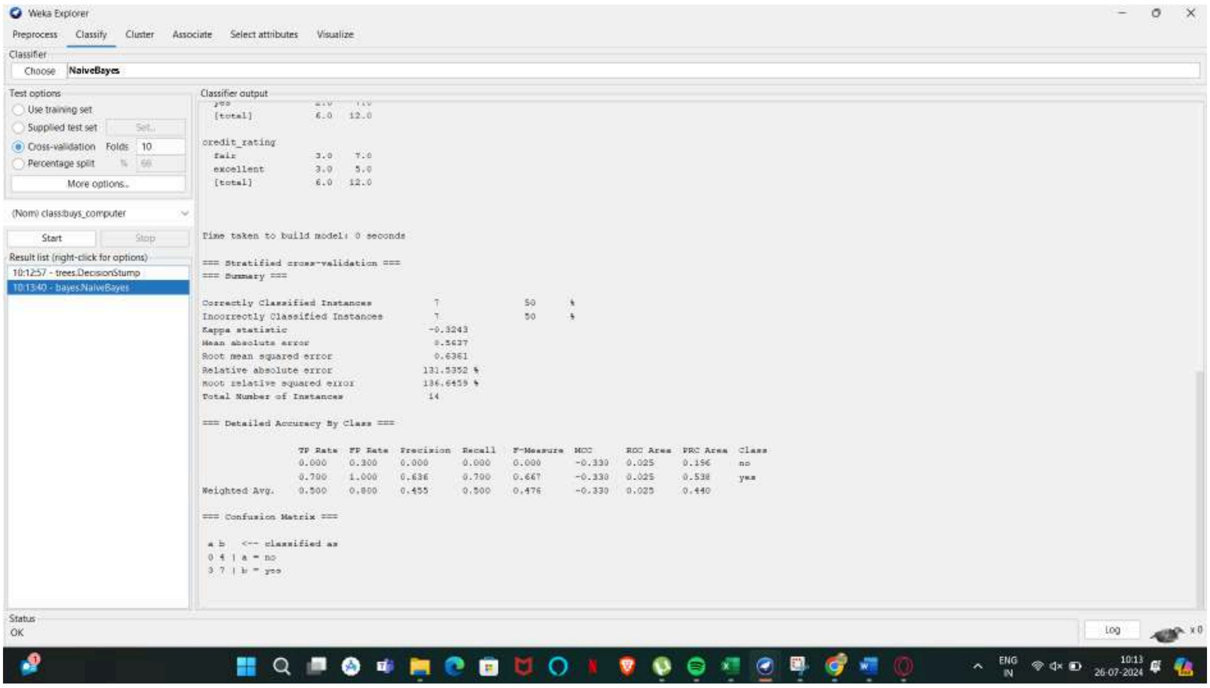




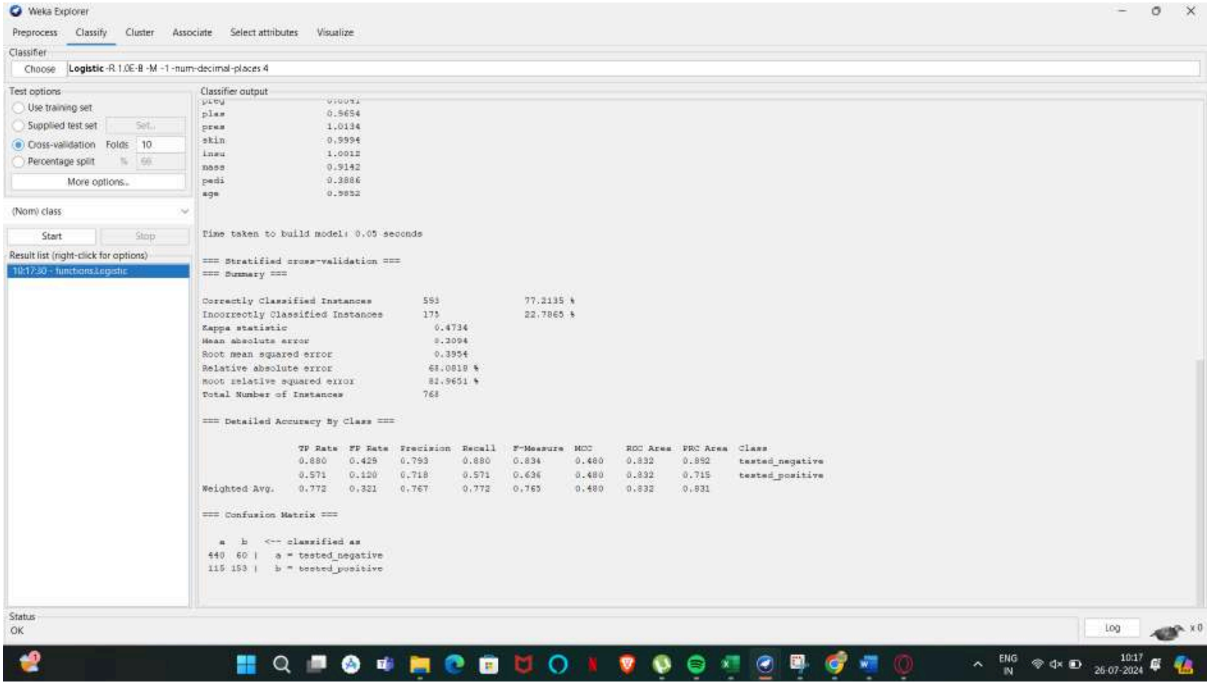

**3)**

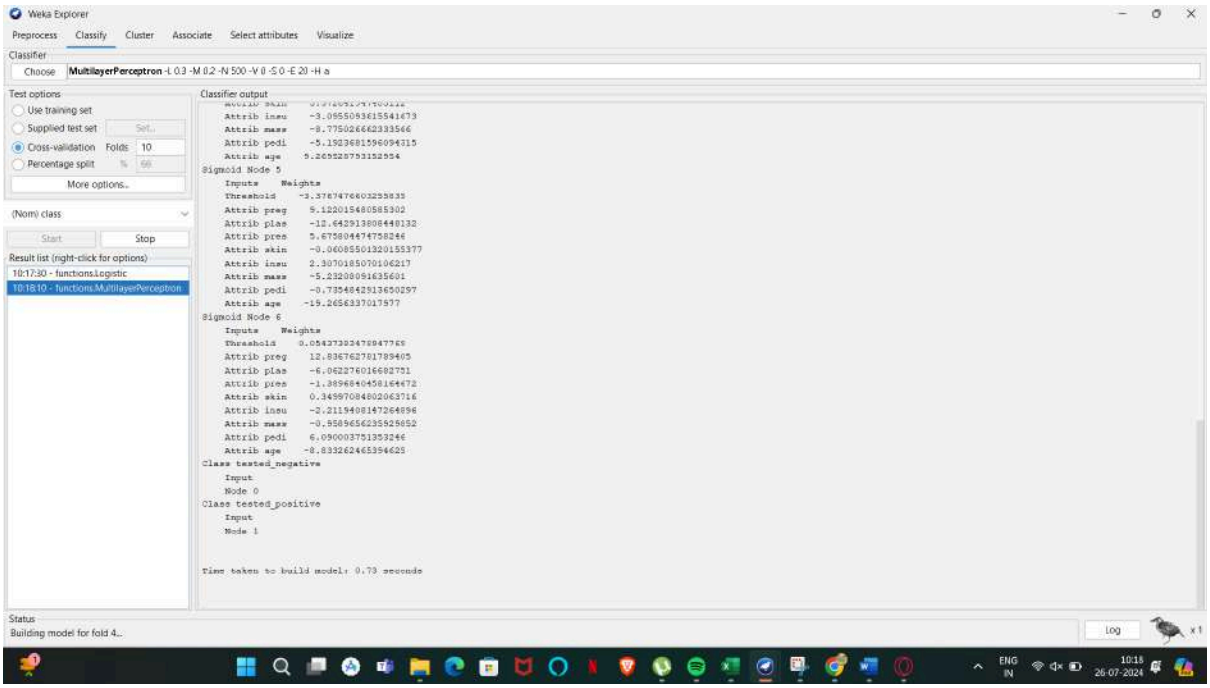
| *RID* | *age* | *income* | *student* | *credit_rating* | *Class: buys_computer* |
|---|---|---|---|---|---|
| 1 | <=30 | high | no | fair | no |
| 2 | <=30 | high | no | excellent | no |
| 3 | 31 . . . 40 | high | no | fair | yes |
| 4 | >40 | medium | no | fair | yes |
| 5 | >40 | low | yes | fair | yes |
| 6 | >40 | low | yes | excellent | no |
| 7 | 31 . . . 40 | low | yes | excellent | yes |
| 8 | <=30 | medium | no | fair | no |
| 9 | <=30 | low | yes | fair | yes |
| 10 | >40 | medium | yes | fair | yes |
| 11 | <=30 | medium | yes | excellent | yes |
| 12 | 31 . . . 40 | medium | no | excellent | yes |
| 13 | 31 . . . 40 | high | yes | fair | yes |
| 14 | >40 | medium | no | excellent | no |


Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
DecisionStump
Test options
Use training set
Supplied test set
Set...
Cross-validation
Folds
10
Percentage split
%
66
More options...
(Nom) class:buys_computer
Start
Stop
Result list (right-click for options)
10:12:57 - trees.DecisionStump
Classifier output
0.0 1.0
age != 31-40
no yes
0.4 0.6
age is missing
no yes
0.2857142857142857 0.7142857142857143
Time taken to build model: 0 seconds
=== Stratified cross-validation ===
=== Summary ===
Correctly Classified Instances 10 71.4286 %
Incorrectly Classified Instances 4 28.5714 %
Kappa statistic 0
Mean absolute error 0.4751
Root mean squared error 0.5717
Relative absolute error 110.8466 %
Root relative squared error 122.814 %
Total Number of Instances 14
=== Detailed Accuracy By Class ===
TP Rate FP Rate Precision Recall F-Measure MCC ROC Area PRC Area Class
0.000 0.000 ? 0.000 ? ? 0.200 0.243 no
1.000 1.000 0.714 1.000 0.833 ? 0.200 0.613 yes
Weighted Avg. 0.714 0.714 ? 0.714 ? ? 0.200 0.507
=== Confusion Matrix ===
a b <-- classified as
0 4 | a = no
0 10 | b = yes
Status
OK
Log
x 0
ENG
IN
10:12
26-07-2024


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) class:buys_computer
Start
Stop
Result list (right-click for options)
10:12:57 - trees.DecisionStump
10:13:40 - bayes.NaiveBayes
Classifier output
[total] 6.0 12.0
credit_rating
fair 3.0 7.0
excellent 3.0 5.0
[total] 6.0 12.0
Time taken to build model: 0 seconds
=== Stratified cross-validation ===
=== Summary ===
Correctly Classified Instances 7 50 %
Incorrectly Classified Instances 7 50 %
Kappa statistic -0.3243
Mean absolute error 0.5627
Root mean squared error 0.6361
Relative absolute error 131.5352 %
Root relative squared error 136.6459 %
Total Number of Instances 14
=== Detailed Accuracy By Class ===
TP Rate FP Rate Precision Recall F-Measure MCC ROC Area PRC Area Class
0.000 0.300 0.000 0.000 0.000 -0.330 0.025 0.156 no
0.700 1.000 0.636 0.700 0.667 -0.330 0.025 0.538 yes
Weighted Avg. 0.500 0.800 0.455 0.500 0.476 -0.330 0.025 0.440
=== Confusion Matrix ===
a b <-- classified as
0 4 | a = no
3 7 | b = yes
Status
OK
Log
x 0
ENG
IN
10:13
26-07-2024

**4) Analysis the dataset “diabetes. csv” how the diabetes trend is for different age people, using linear regression and multiple regression.**



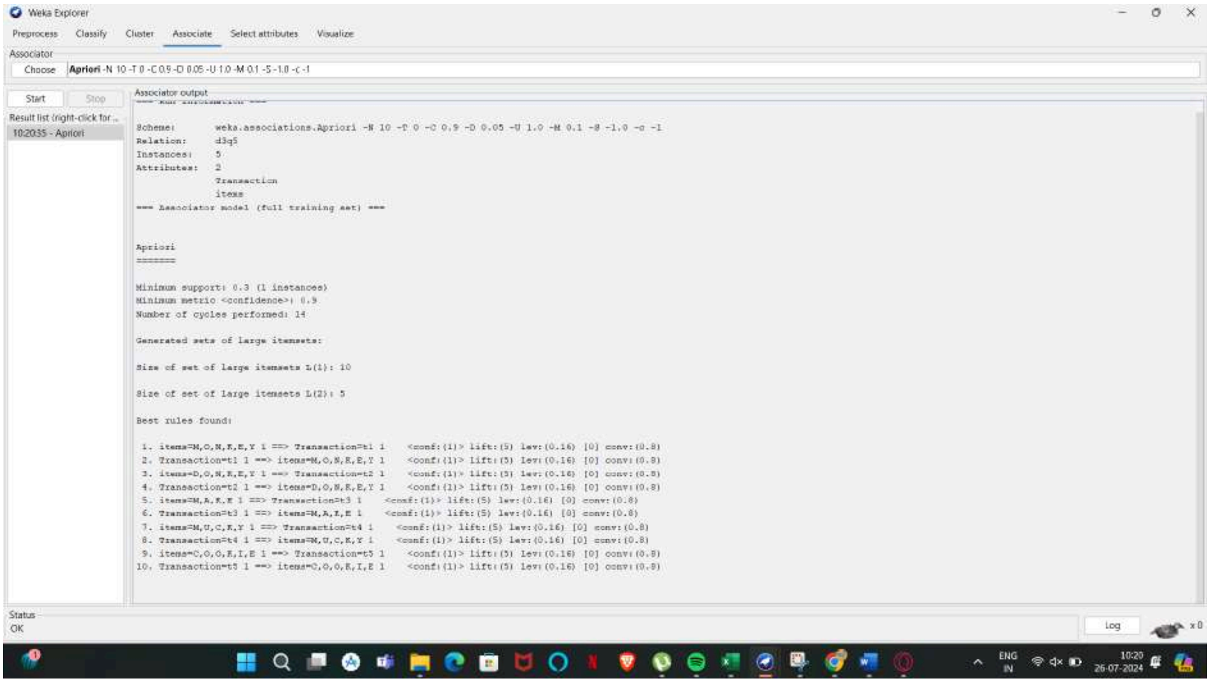


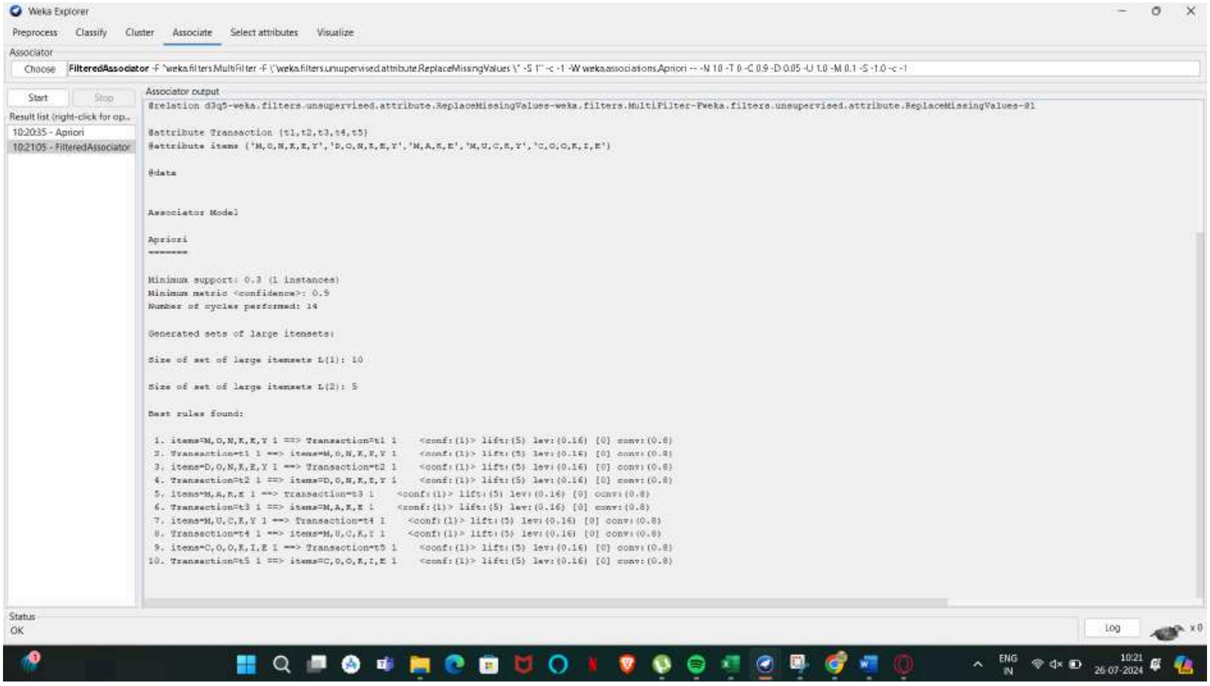
**5)**

**Implement using WEKA for the given Suppose a database has five transactions. Let min sup= 50%(2) and min con f = 80%.**

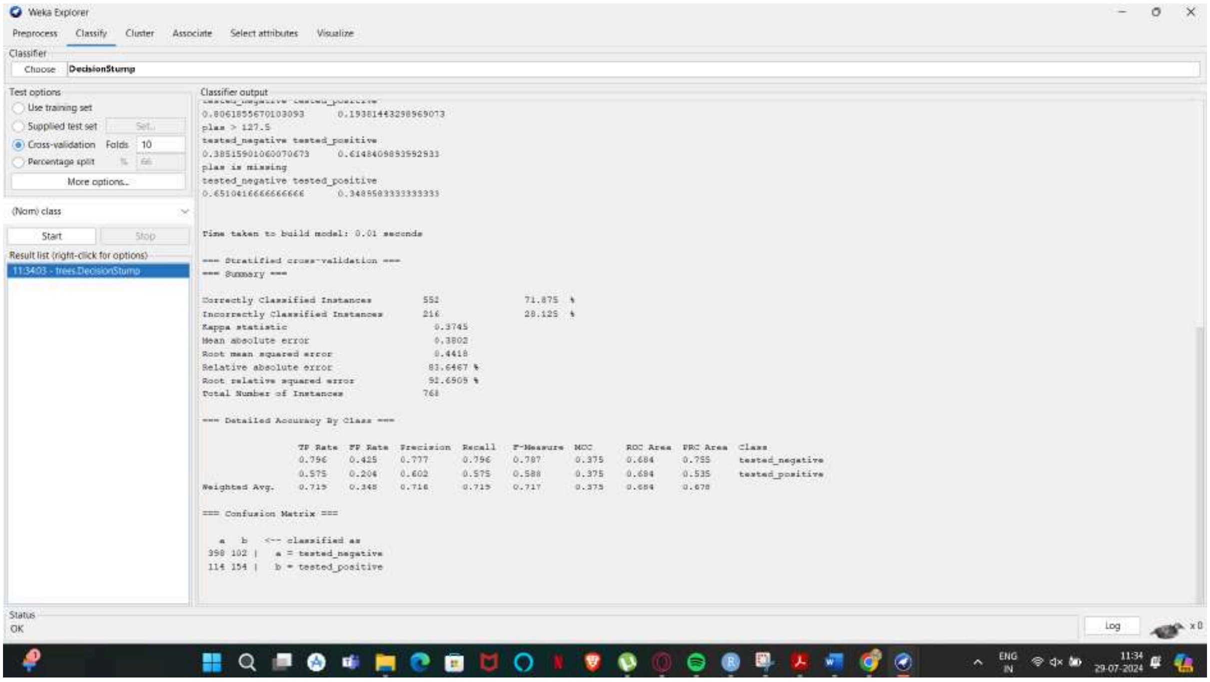
| Transactions | Items |
|---|---|
| **T1** | **(M, O, N, K, E, Y)** |
| **T2** | **(D, O, N, K, E, Y)** |
| **T3** | **(M, A, K, E)** |
| **T4** | **(M, U, C, K, Y)** |
| **T5** | **(C,O, O, K, I ,E)** |

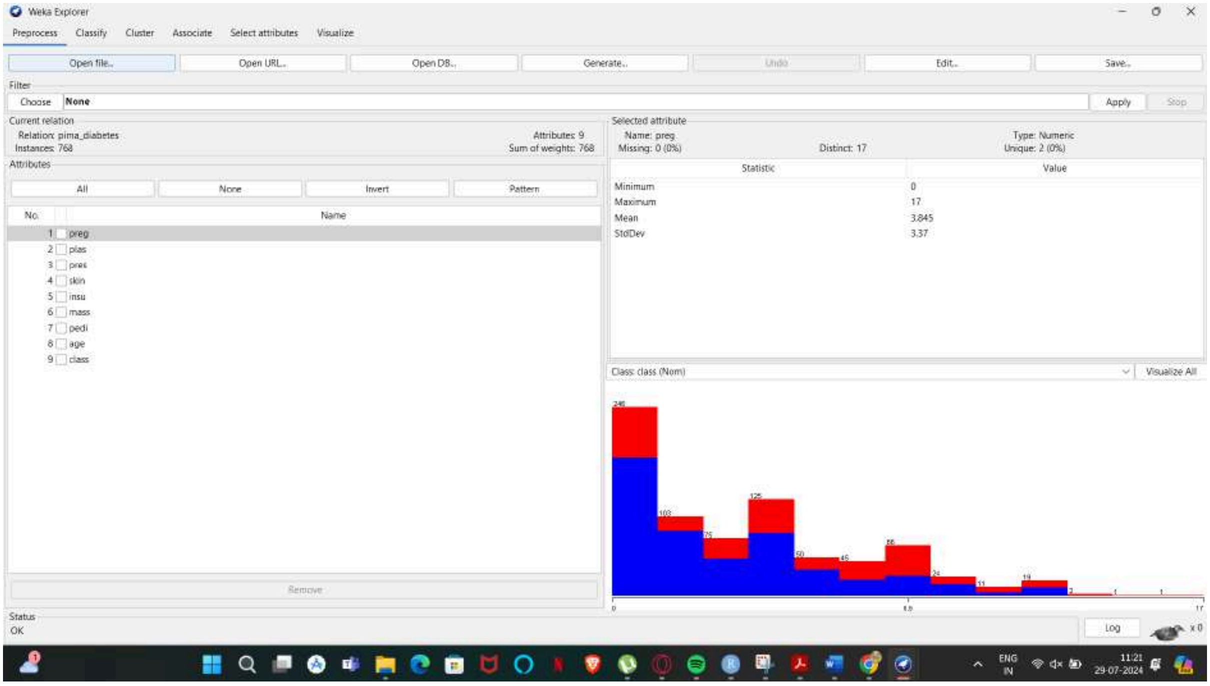
- **Find all frequent item sets using Apriori algorithm**
- **Also draw FP-Growth Tree**

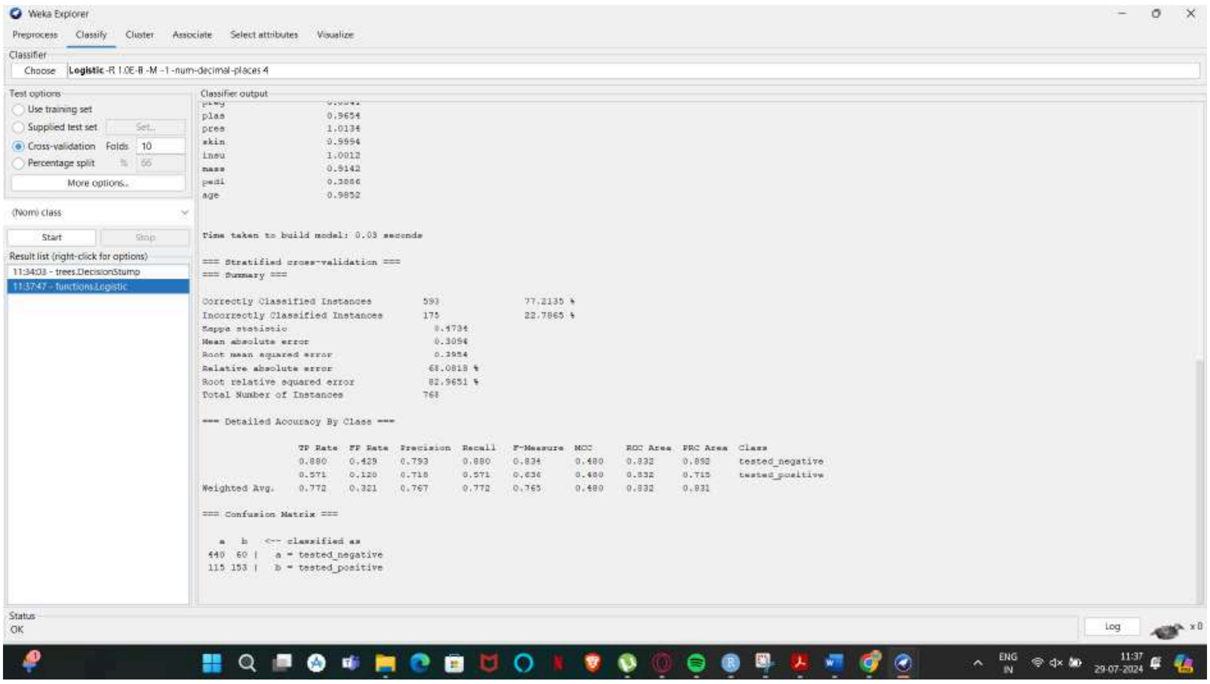





**6) Prediction of Categorical Data using Decision Tree Algorithm through WEKA using any datasets. a) Tree b) Preprocess c) Logistic**



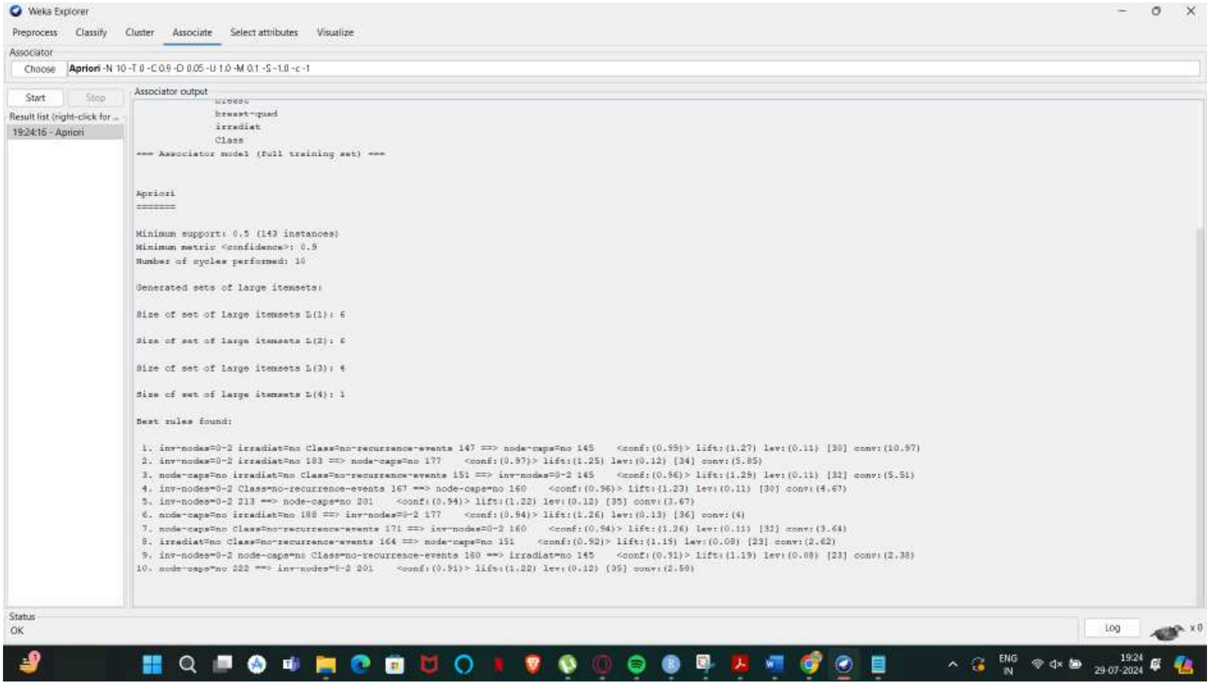


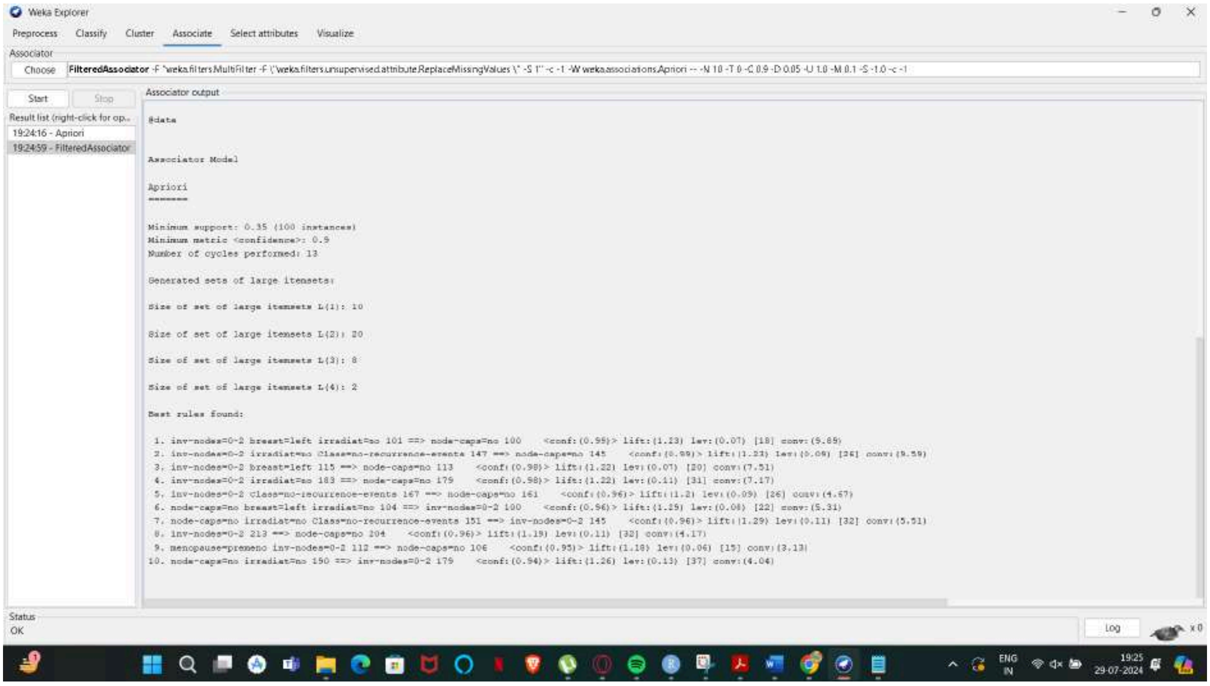




**7) Create the dataset using ARFF file format:**

**a.Find the frequent itemsets and generate association rules on this. Assume that minimum support threshold (s = 33.33%) and minimum confident threshold (c = 60%).**

**b.List the various rule generated by apriori and FP tree algorthim ,mention wheather accepted or rejcted.**

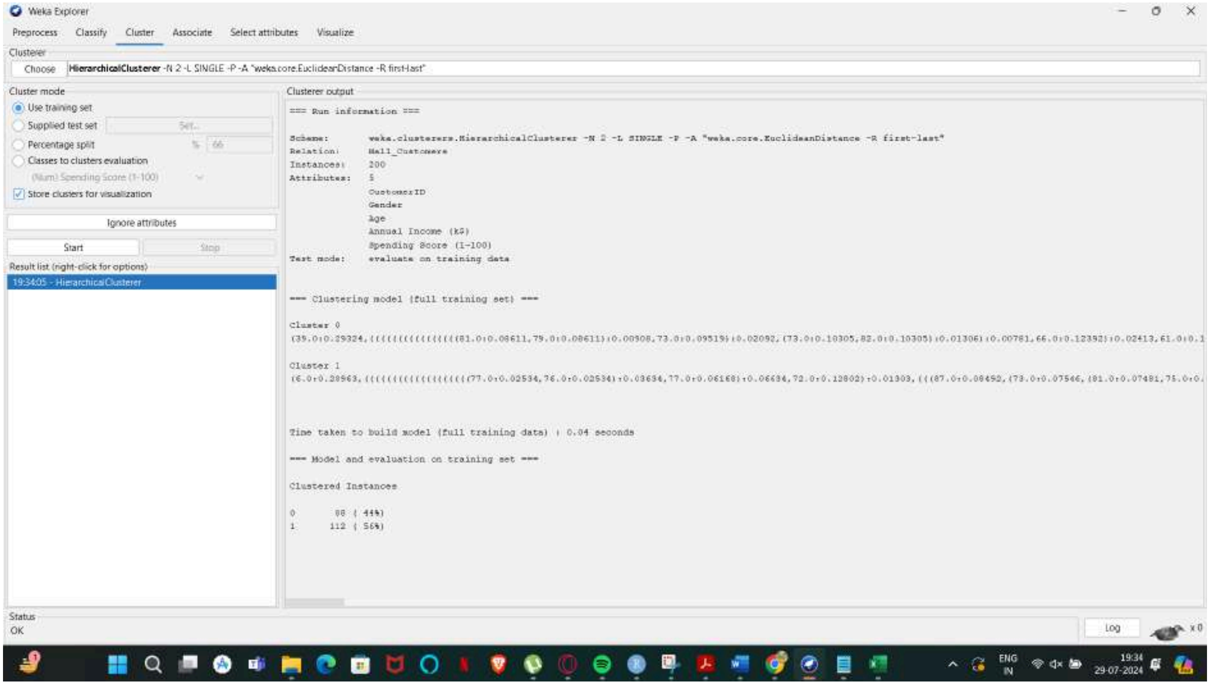


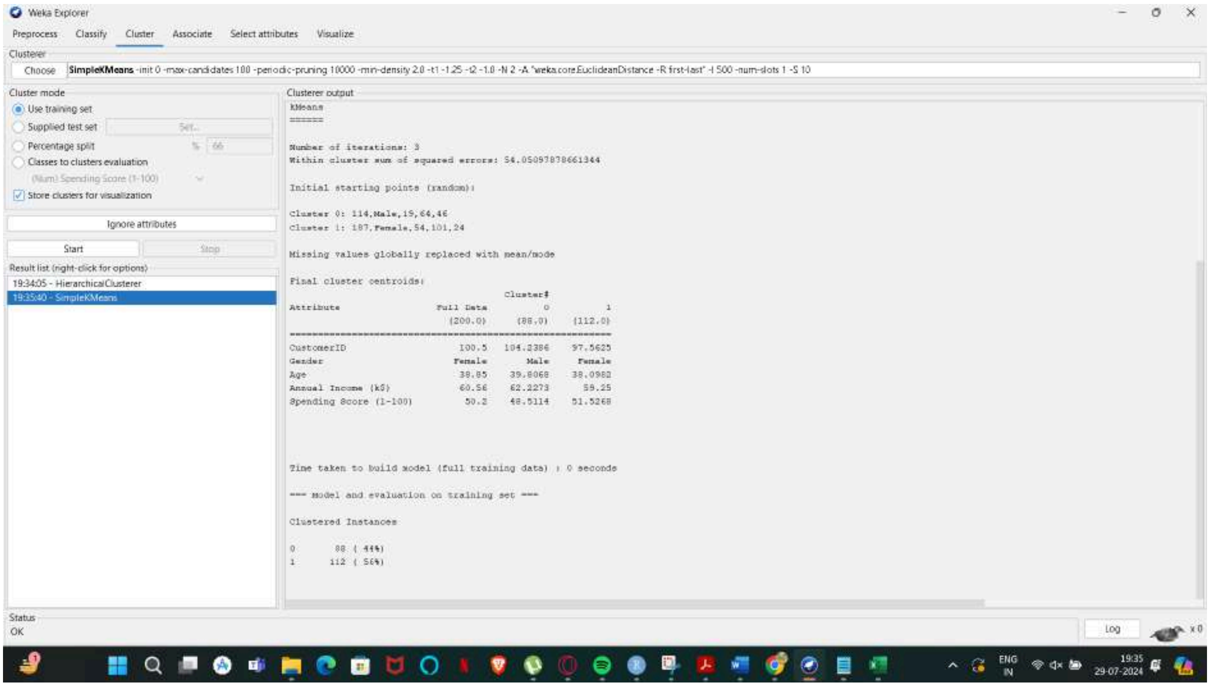


**Day_04**

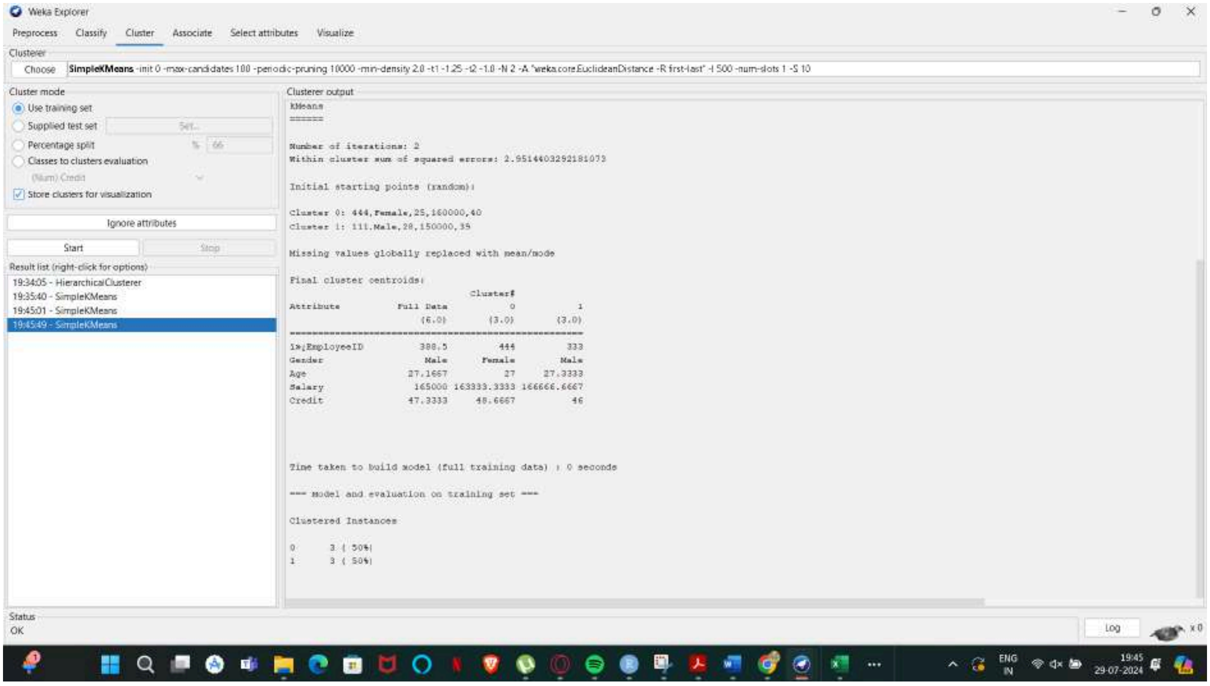
1) **Consider that you are owning a supermarket mall and through membership cards, you have some basic data about your customers like Customer ID, age, gender, annual income and spending score. For the above scenario, the Problem**

**Statement was You want to understand the customers who can easily converge [Target Customers] so that the data can be given to the marketing team and plan the strategy accordingly. For the above scenario prepare a dataset and perform Clustering Analysis to segment the customers in the Mall. There are clearly Five segments of Customers based on their Annual Income and Spending Score namely *Usual Customers, Priority Customers, Senior Citizen Target Customers, and Young Target Customers.***

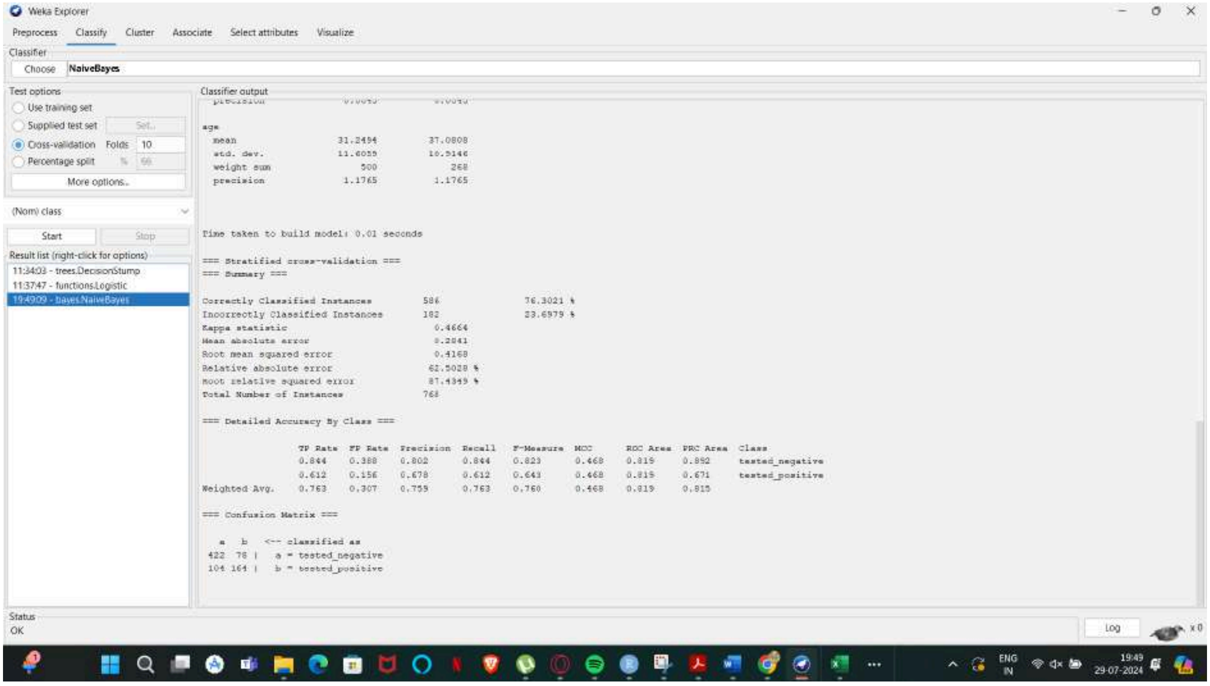


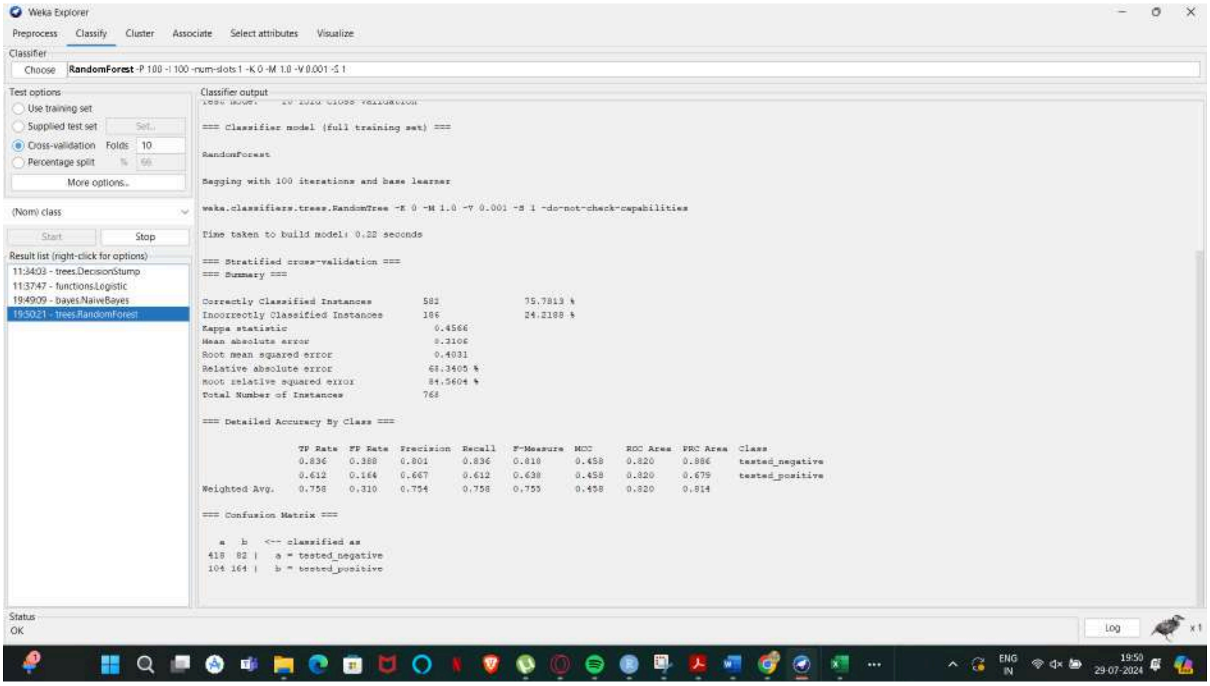



2) **Create the following dataset using CSV file format. To perform cluster analysis using K- Means in WEKA. To change the cluster size and plot the graph and illustrate the visualization of cluster.**

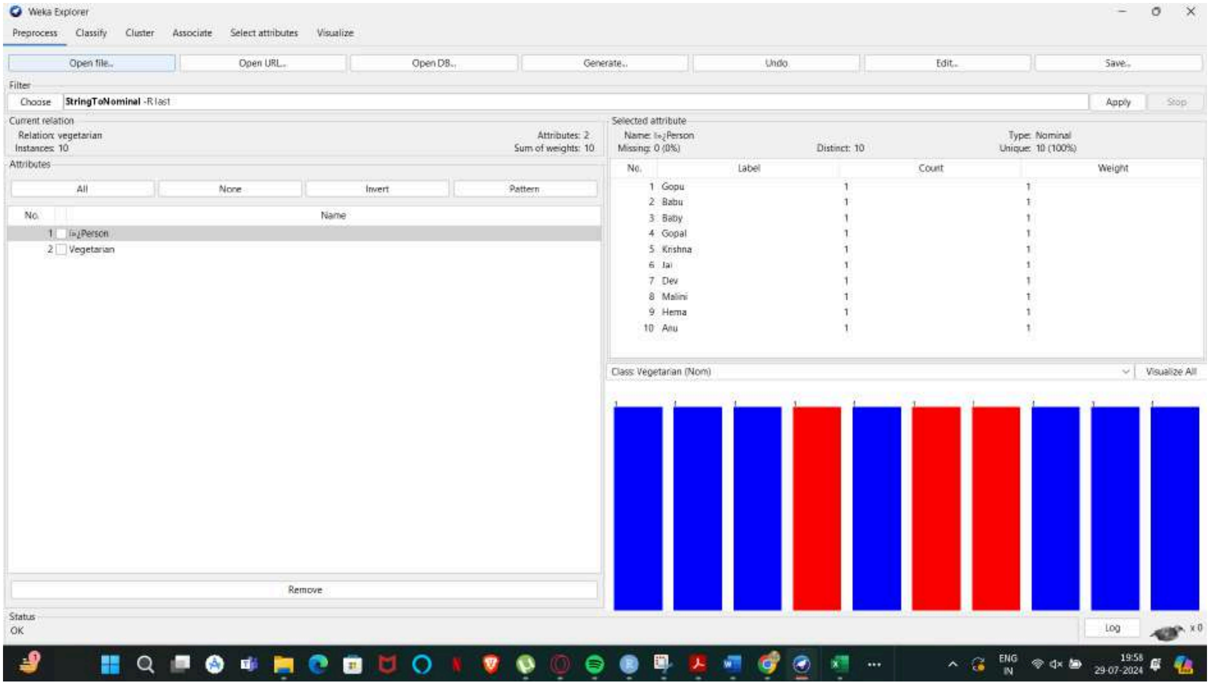


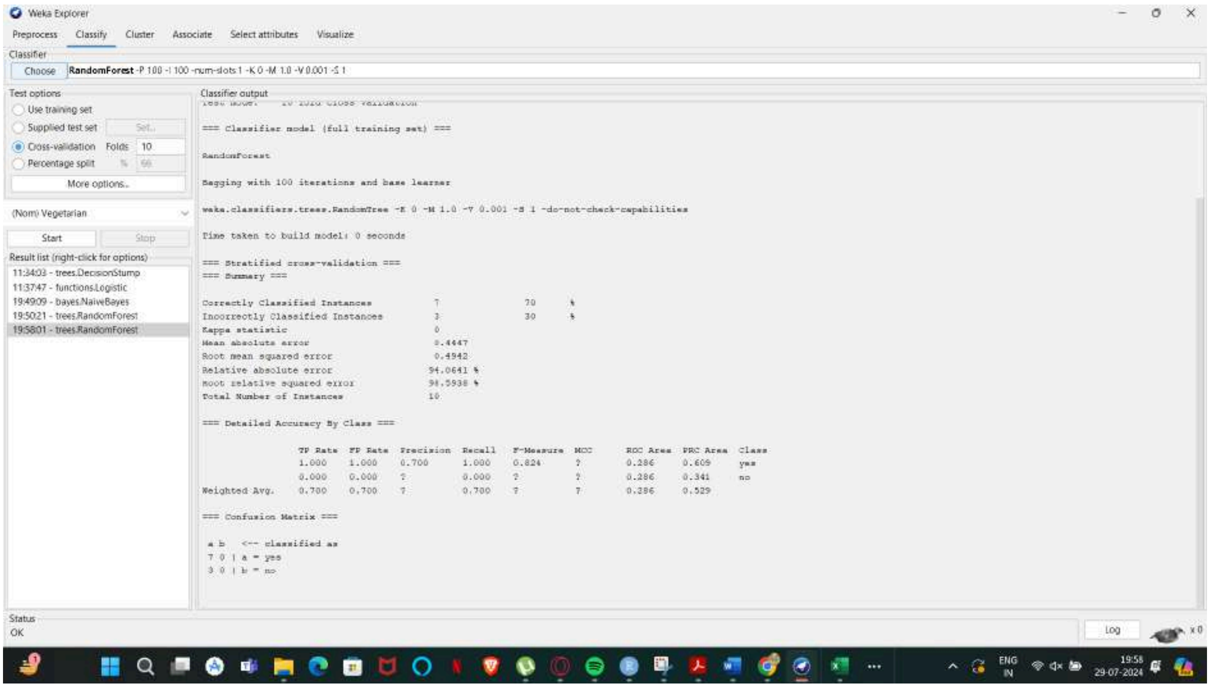

3) **Prediction of categorical data using Naïve Bayes classification through WEKA using any datasets. Compare the Naïve Bayes algorithm with SVM using the summary of results given by the classifiers and plot the graph.**



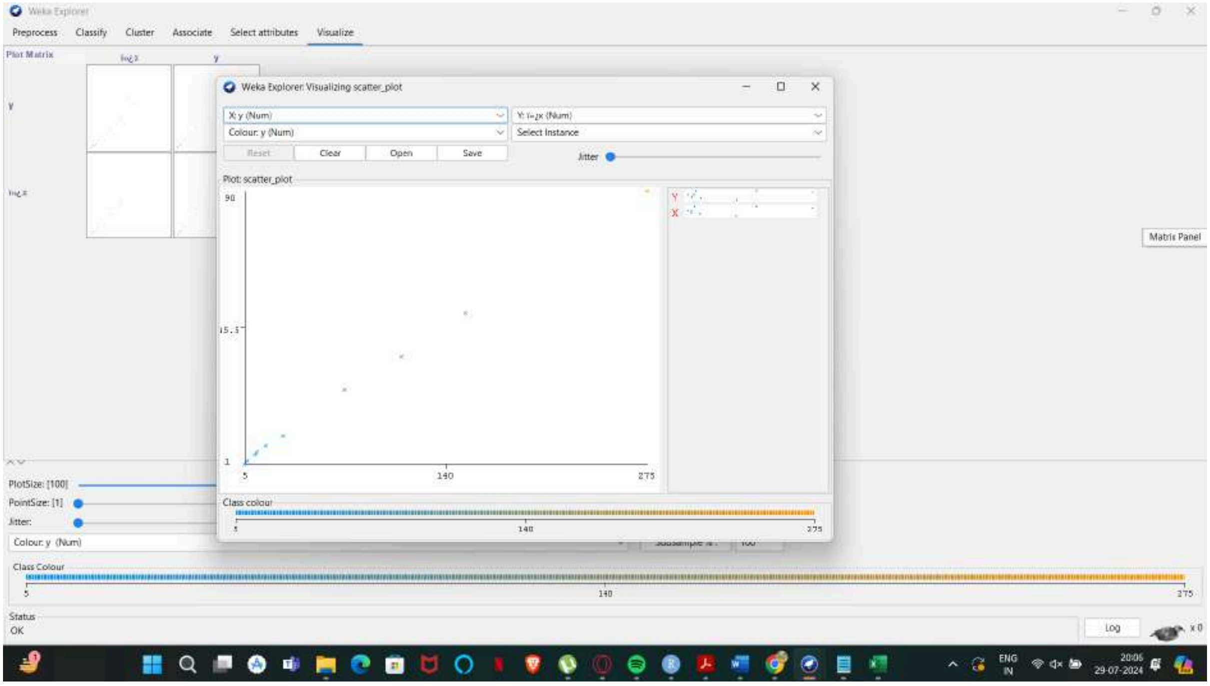




4) **The following list of persons with vegetarian or not details is given in the table. How will you find out how many of them are vegetarian and how many of them are non-vegetarian? Which type of the person's total count has greater value?**

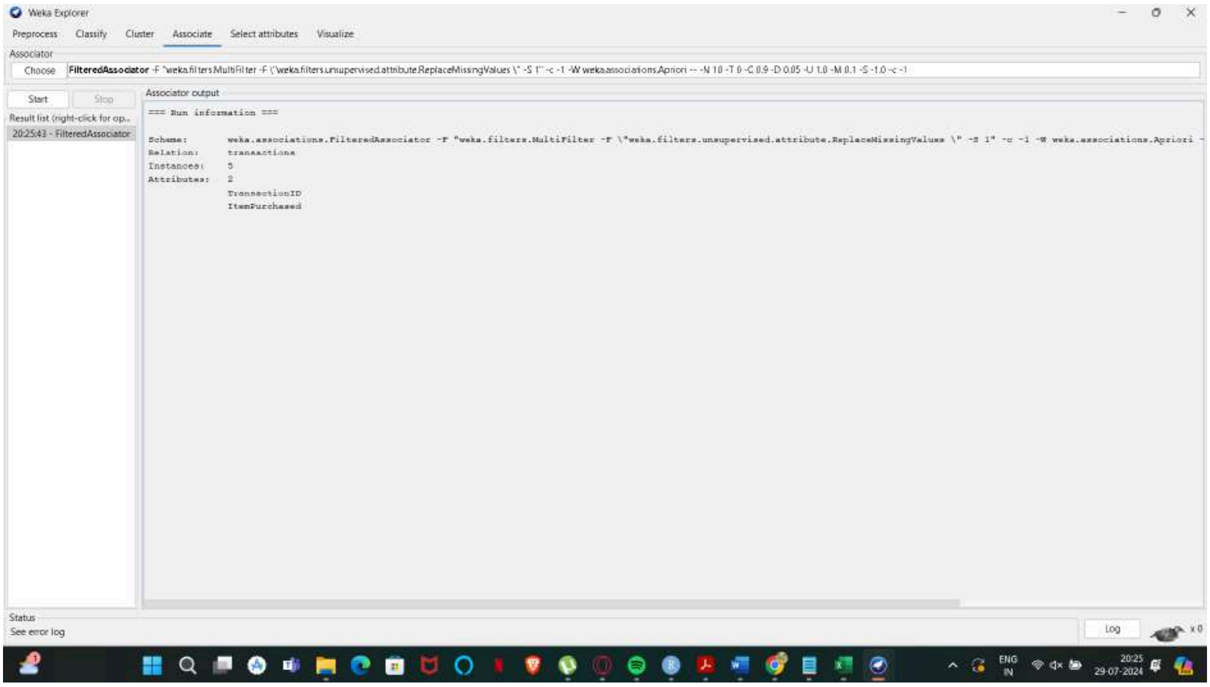


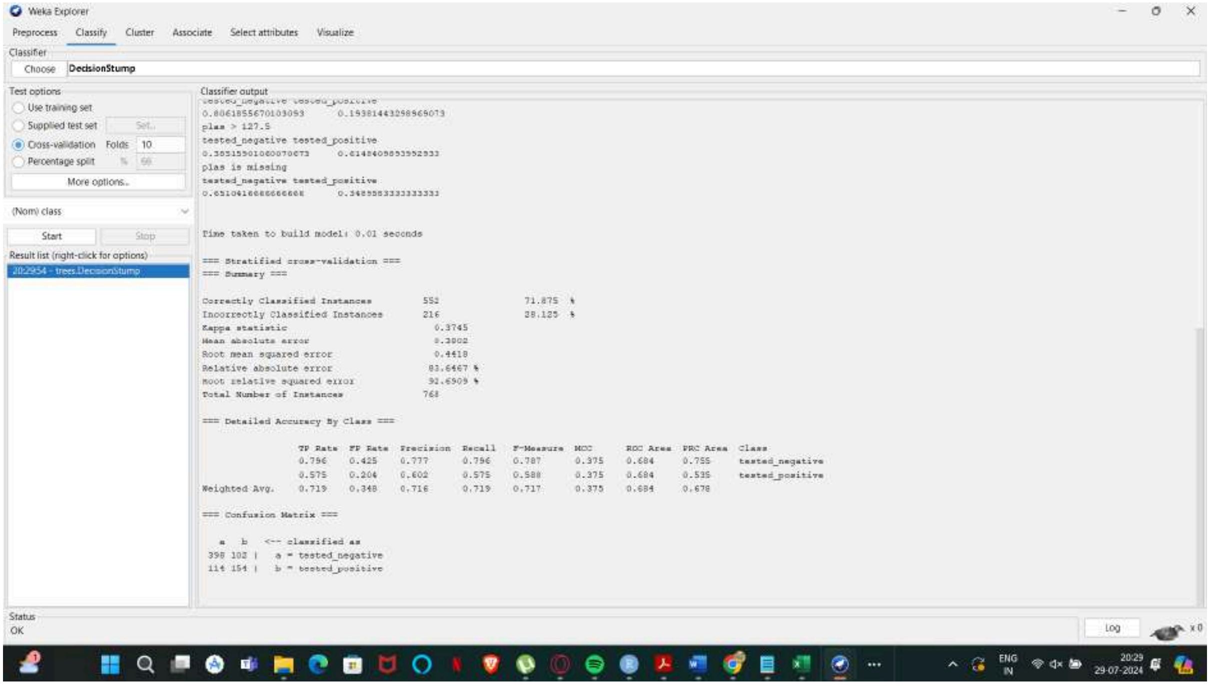

5) **The following table would be plotted as (x,y) points, with the first column being the x values as number of mobile phones sold and the second column being the y values as money. To use the scatter plot for how many mobile phones sold.**

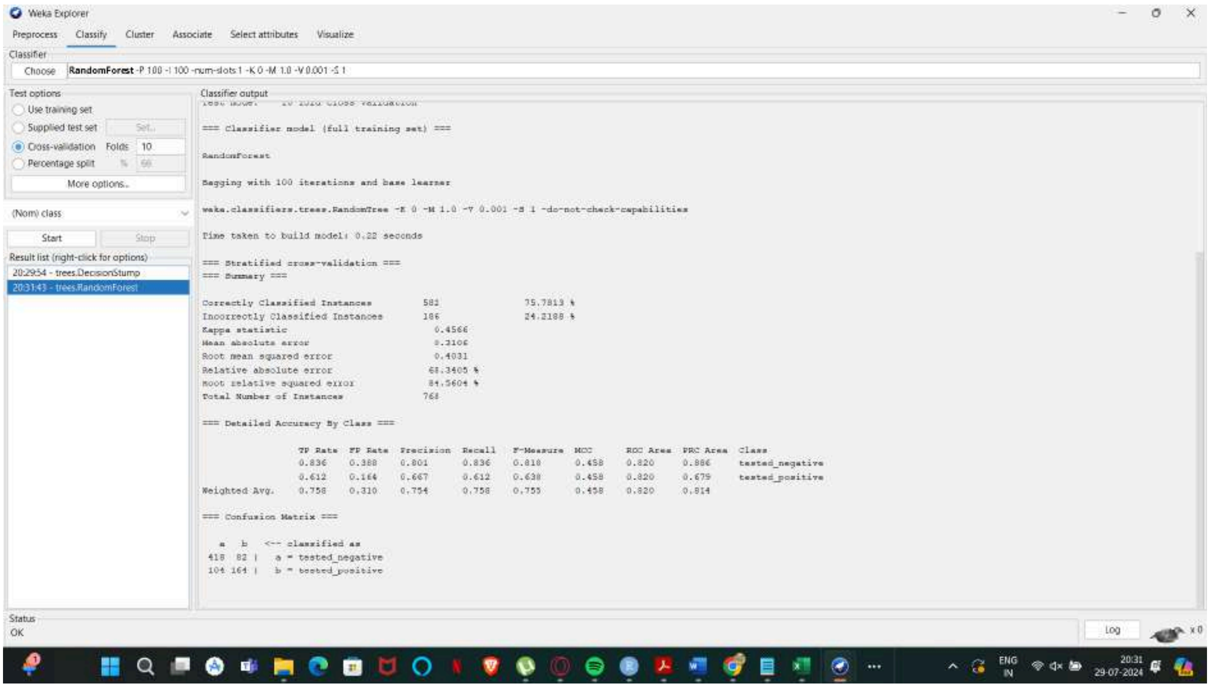


6) **Generate rules using FP growth algorithm using the given dataset which has the following transactions with items purchased: Consider the values as support=50% and confidence=75%.**



7) **Prediction of Diabetes Data using Decision tree classifier in WEKA. Compare it with Support Vector Machine classifier. Show the result accuracy and F1 measure calculation .Plot the graph and explain the summary of results.**



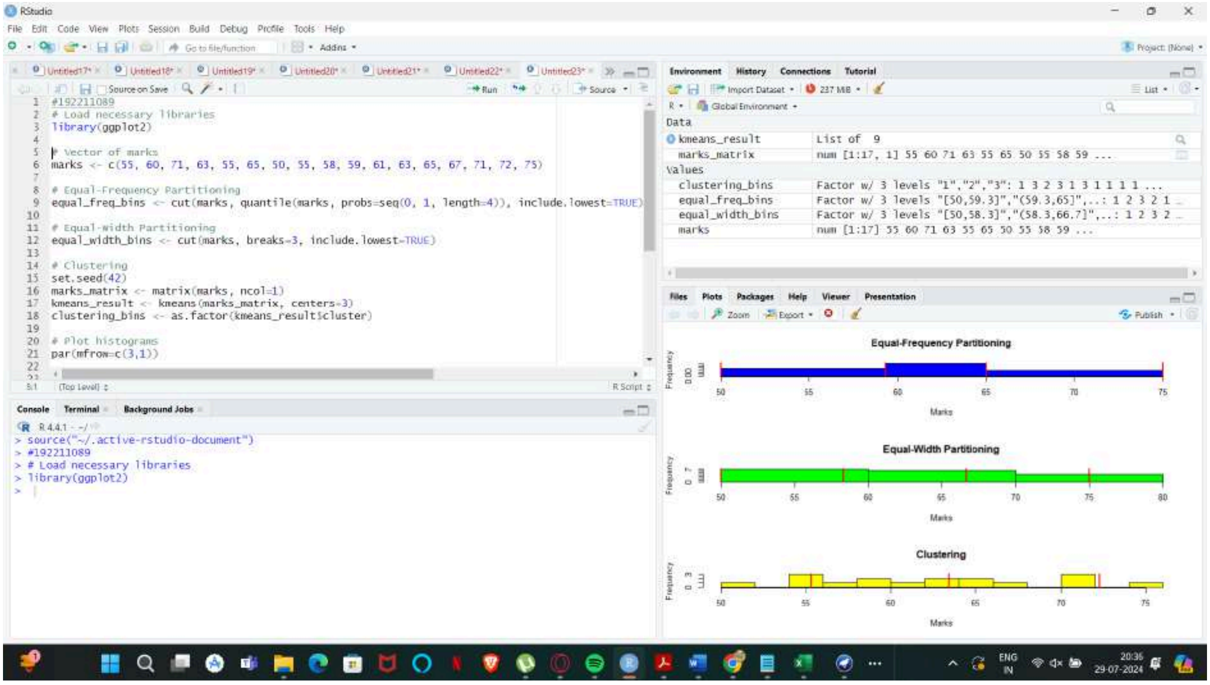



8) .Implement of the R script using marks scored by a student in his model exam has been sorted as follows: 55, 60, 71, 63, 55, 65, 50, 55,58,59,61,63,65,67,71,72,75. Partition them into three bins by each of the following methods. Plot the data points using histogram.

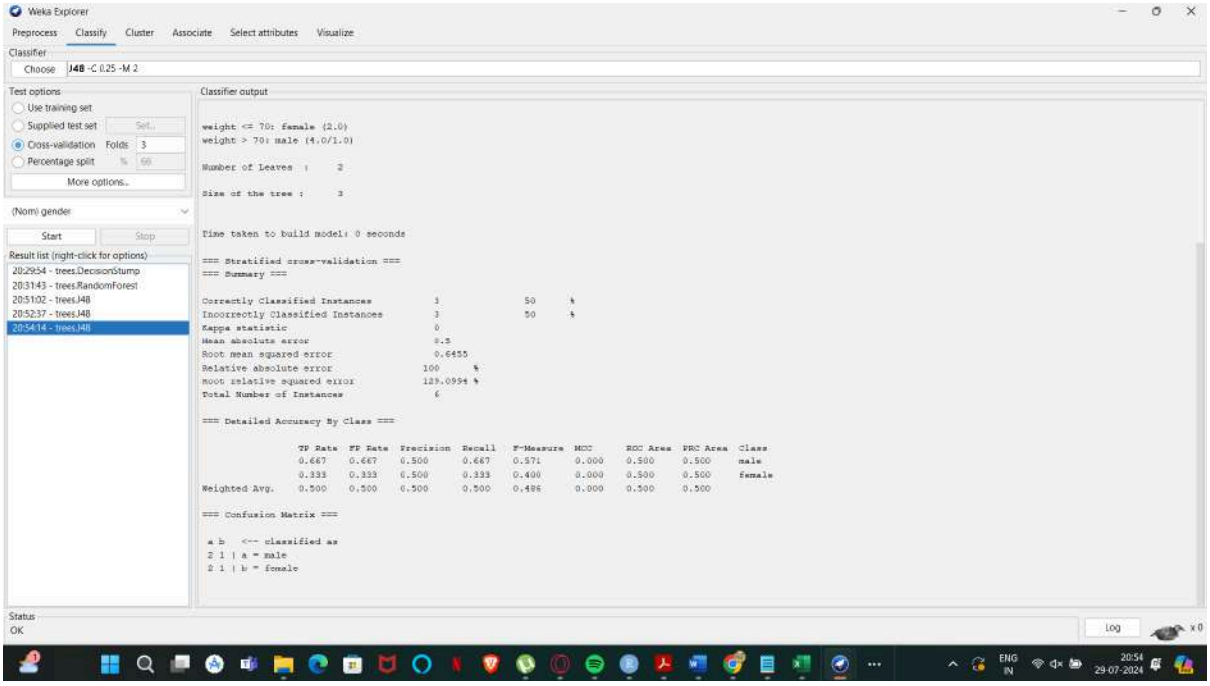
(a) equal-frequency (equi-depth) partitioning

(b) equal-width partitioning

(c) clustering

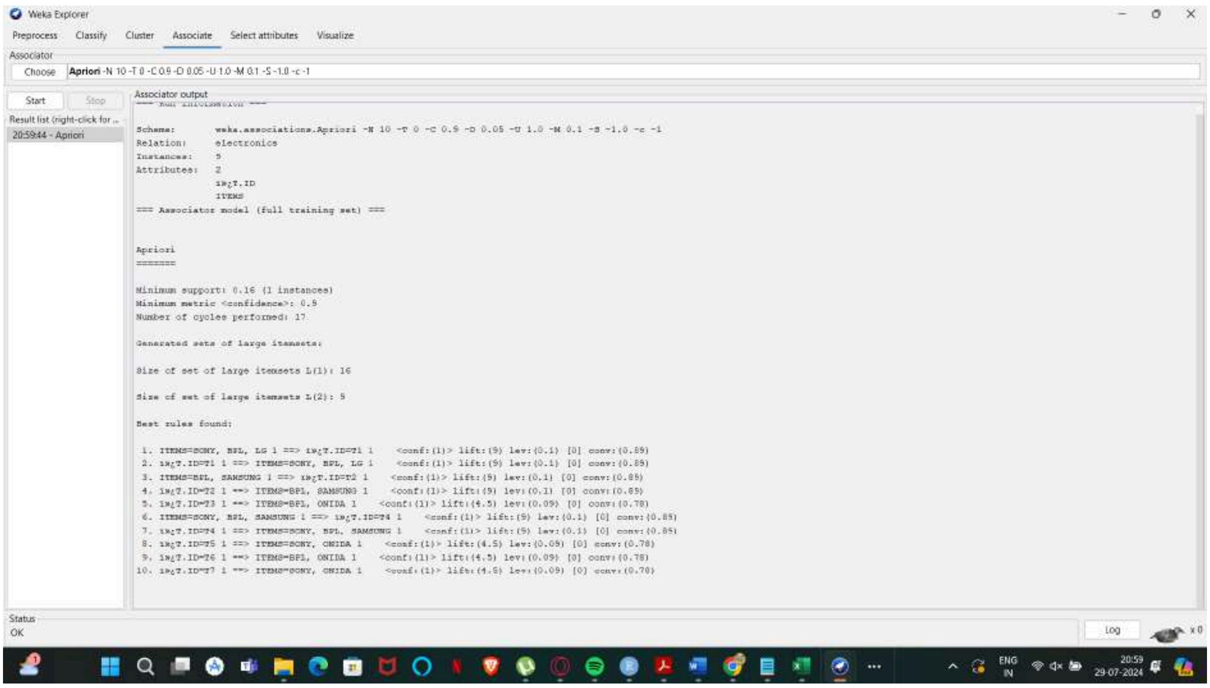


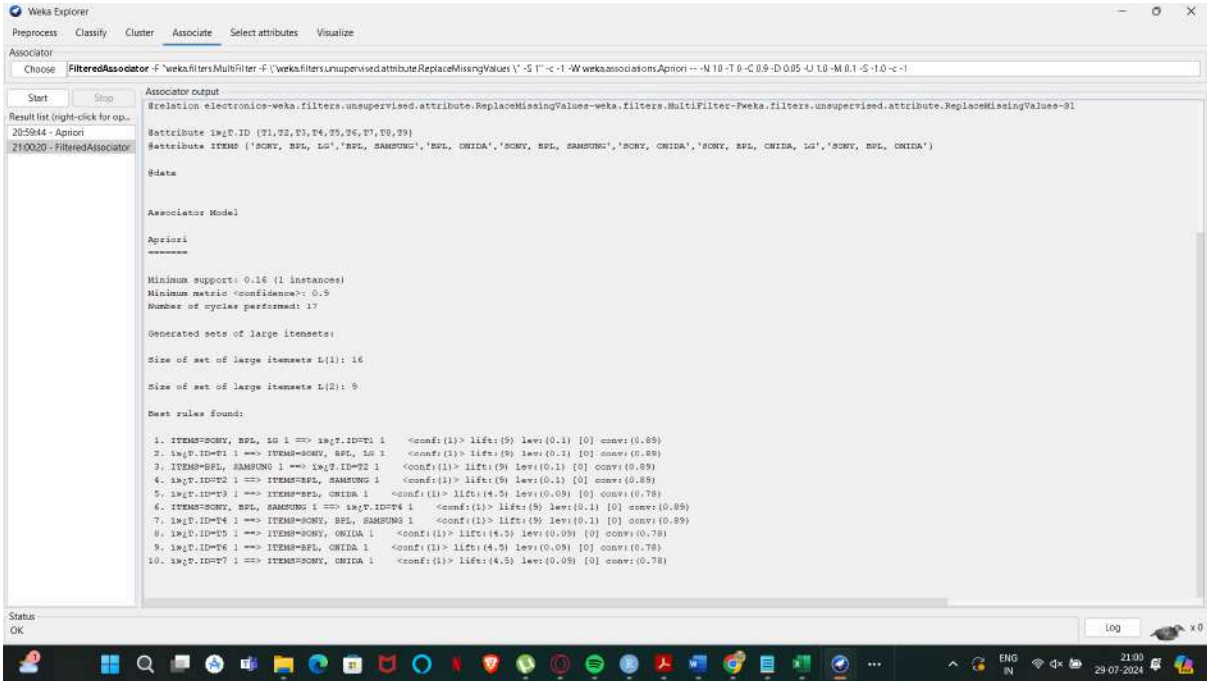

9) Consider this Decision tree :
a)create the data set for the below tree using ARFF format and calculate accuracy and decision for the same
b) Using this decision tree generate the rules based on rule based induction.
c) Compare both the algorithms and plot the confusion matrix.




10) .Create an ARFF file for the table below and implement for the Apriori Algorithm and FP growth algorithm and compare the rules generated by both the algorithms. Identify the unique rules generated by the above algorithms.

NOTE: Assume Min_sup=2 and confidence= 50%





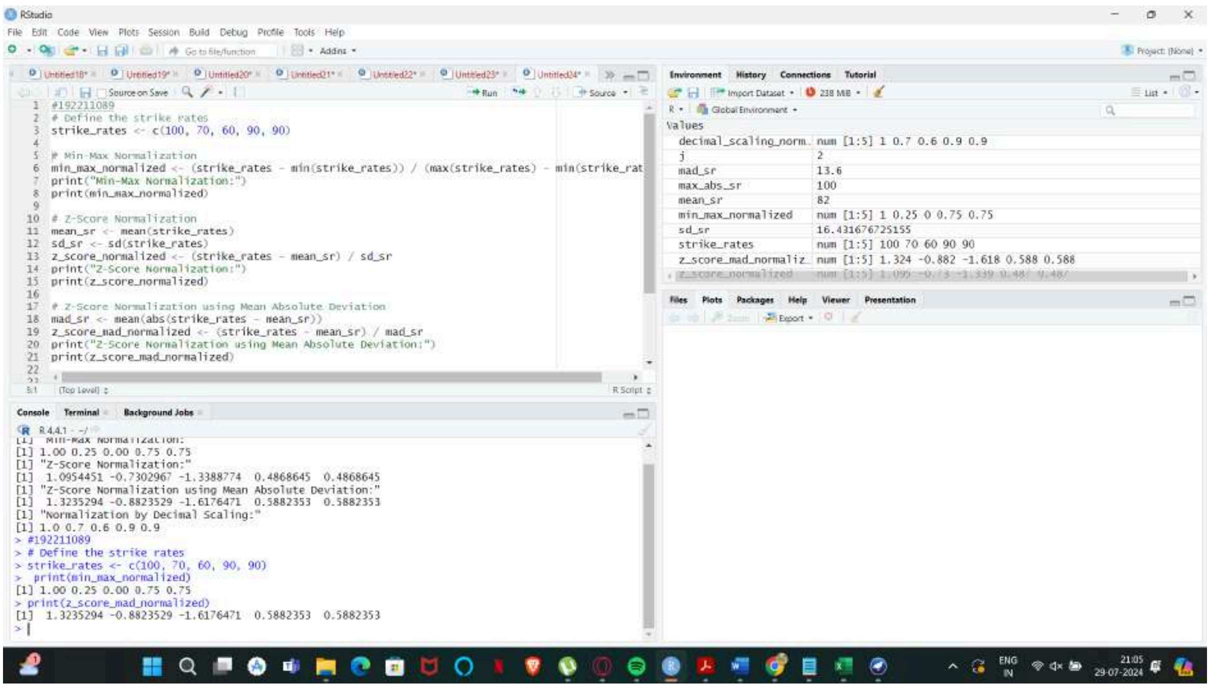
11) The given are the strike-rates scored by a batsman in season 1 in different tournaments. 100, 70, 60, 90, 90

(a) min-max normalization by setting min = 0 and max = 1

(b) z-score normalization

(c) z-score normalization using the mean absolute deviation instead of standard deviation

(d) normalization by decimal scaling

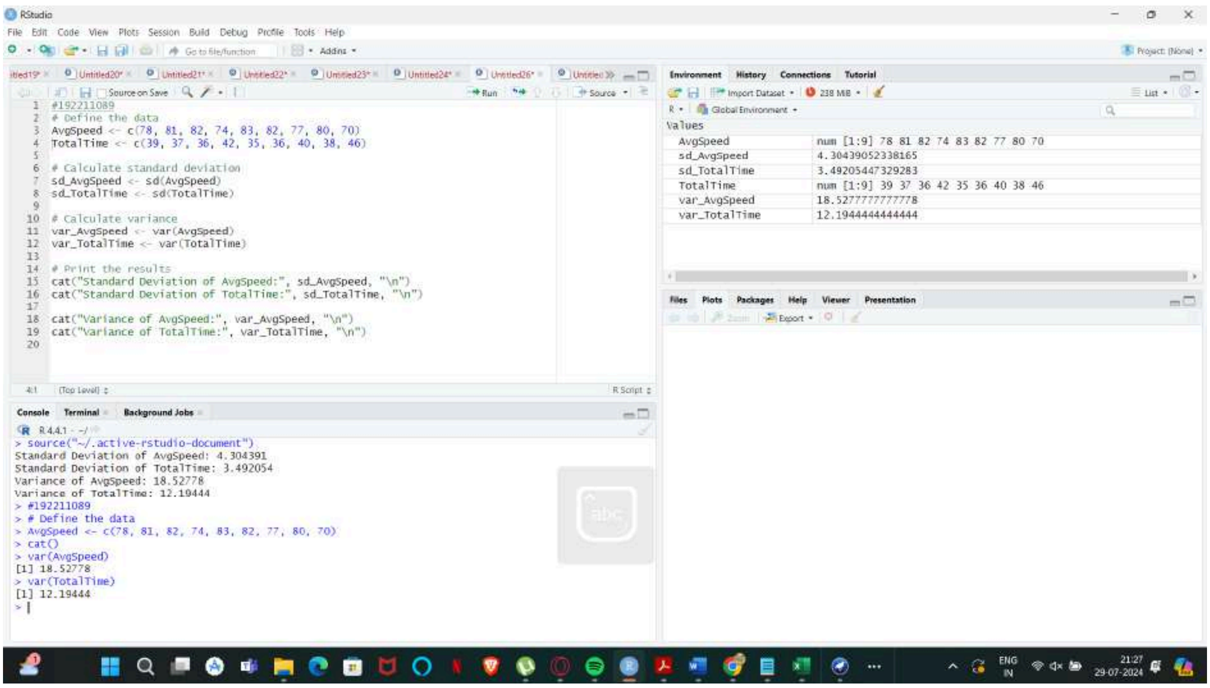


**12)** **Suppose some car is tested for the AvgSpeed and TotalTime data for 9 randomly selected car with the following result**

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **AvgSpeed (in kph)** | **78** | **81** | **82** | **74** | **83** | **82** | **77** | **80** | **70** |
| **TotalTime (in mins)** | **39** | **37** | **36** | **42** | **35** | **36** | **40** | **38** | **46** |

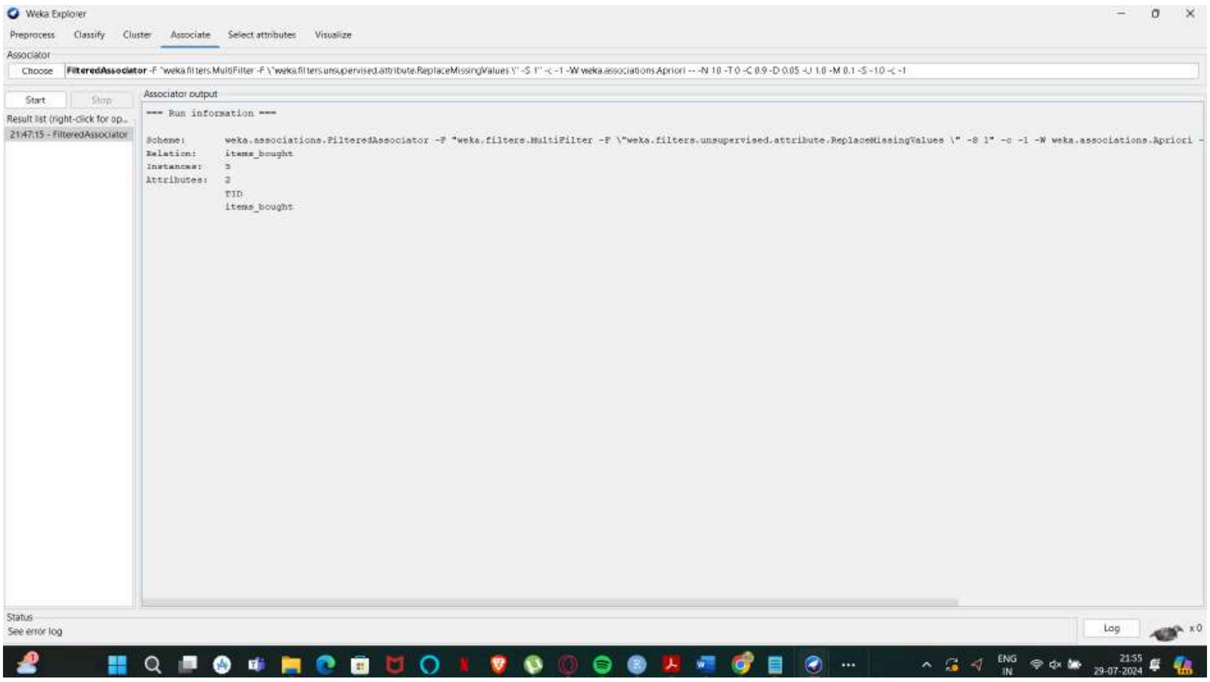
**a) Calculate the standard deviation of AvgSpeed and TotalTime.**

**b) Calculate the Variance of AvgSpeed and TotalTime for the above dataset.**



13)Consider this table

a) **TID items bought**
b) **T100 {M, O, N, K, E, Y}**
c) **T200 {D, O, N, K, E, Y }**
d) **T300 {M, A, K, E}**
e) **T400 {M, U, C, K, Y}**
f) **T500 {C, O, O, K, I ,E}**
g) **(a) Find all frequent item set using Apriori and FP-growth, respectively. Compare the efficiency of the two mining processes.**
h) **(b) List all of the strong association rules (with support s and confidence c) matching the following metarule, where X is a variable representing customers, and itemi denotes variables representing items (e.g., “A”, “B”, etc.):**
i) $\forall$ **x** $\in$ **transaction, buys(X, item1)** $\wedge$ **buys(X, item2)** $\Rightarrow$ **buys(X, item3)**



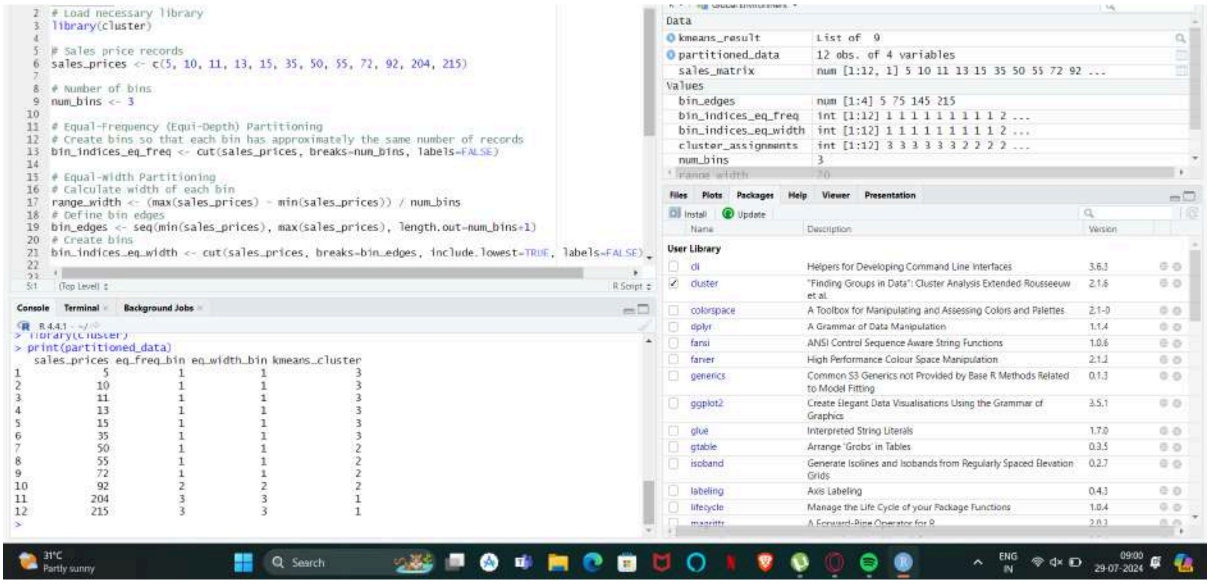
QUESTION_BANK :

**1) Implement of the R script using a group of 12 sales price records has been sorted as follows: 5, 10, 11, 13, 15, 35, 50, 55, 72, 92, 204, 215. Partition them into three bins by each of the following methods.**

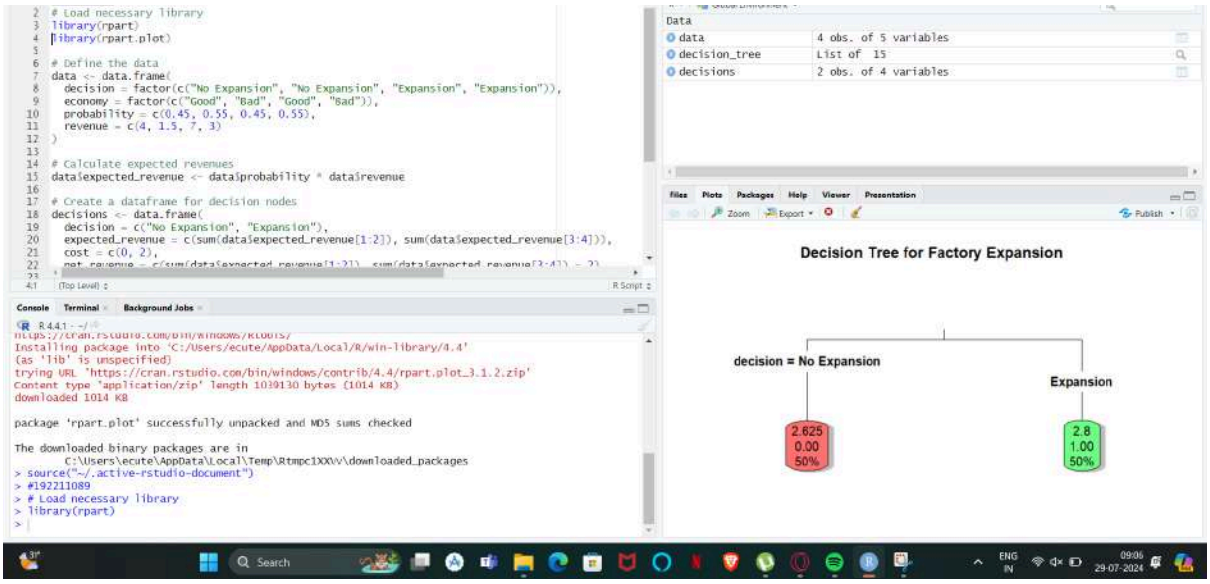
**(a) equal-frequency (equi depth) partitioning**

**(b) equal-width partitioning**

**(c) clustering**




**2) A gadget factory has been quite successful for the past 10 years and Ms.Marry, the manager of the company wondering whether to expand the factory this year or not. The cost to expand factory is $2M. With no expansion, expected revenue is $4M if the economy stays good; while only $1.5M if the economy is bad. If manager expands the factory, expected to receive $7M. if economy is good and $3M if economy is bad. Assume that there is a 45% chance of a good economy and a 55% chance of a bad economy. Draw a Decision Tree showing these choices.**




**3) Implement using WEKA for the given Apply Apriori Algorithm for given database below Assume Min_sup=2 TID Items**

**1**

**Bread, Peanuts, Milk, Fruit, Jam**

**2**

**Bread, Jam, Soda, Chips, Milk, Fruit**

**3**

**Steak, Jam, Soda, Chips, Bread**

**4**

**Jam, Soda, Peanuts, Milk, Fruit**

**5**

**Jam, Soda, Chips, Milk, Bread**

**6**

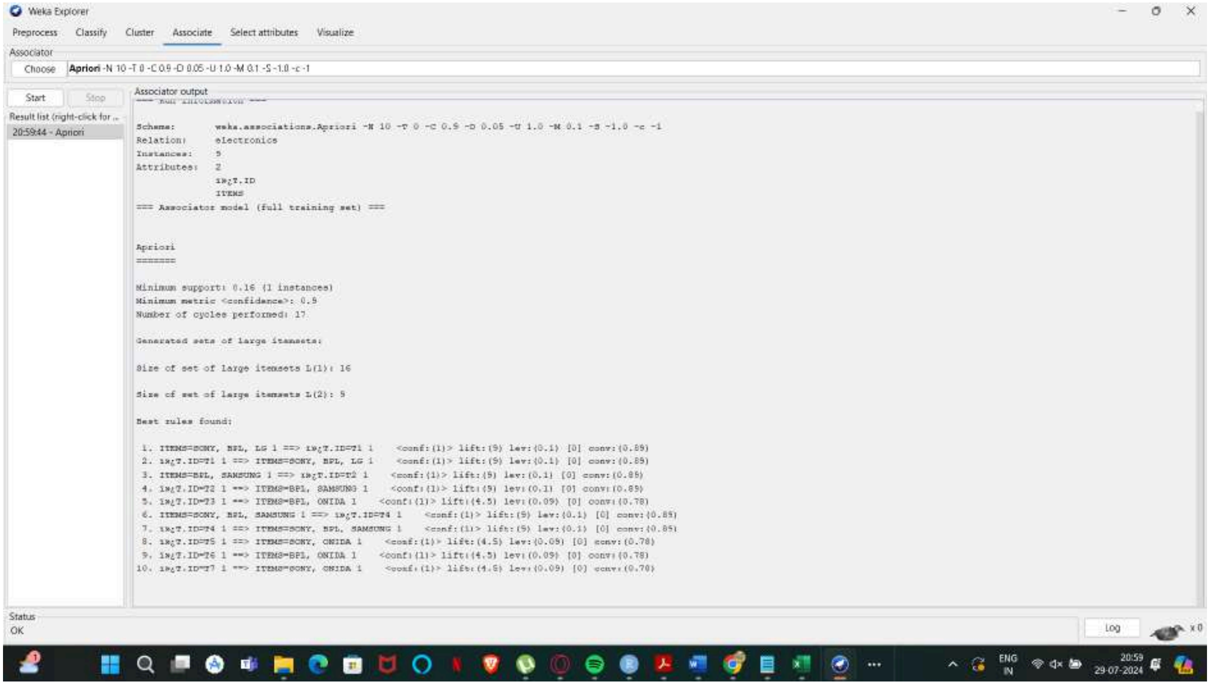
**Fruit, Soda, Chips, Milk**

**7**

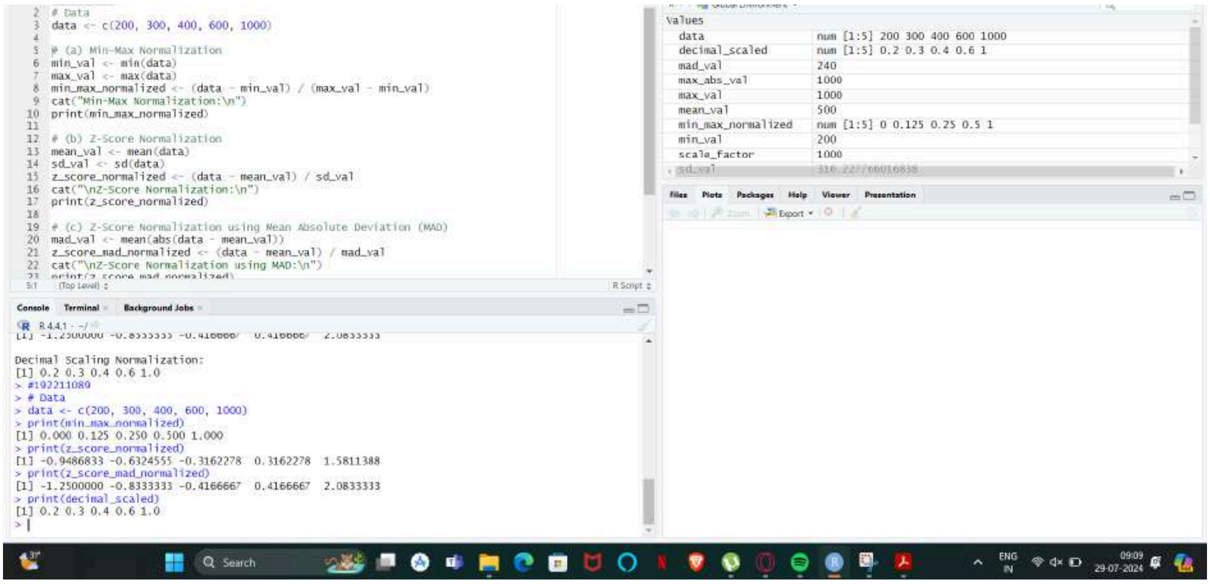
**Fruit, Soda, Peanuts, Milk**

8

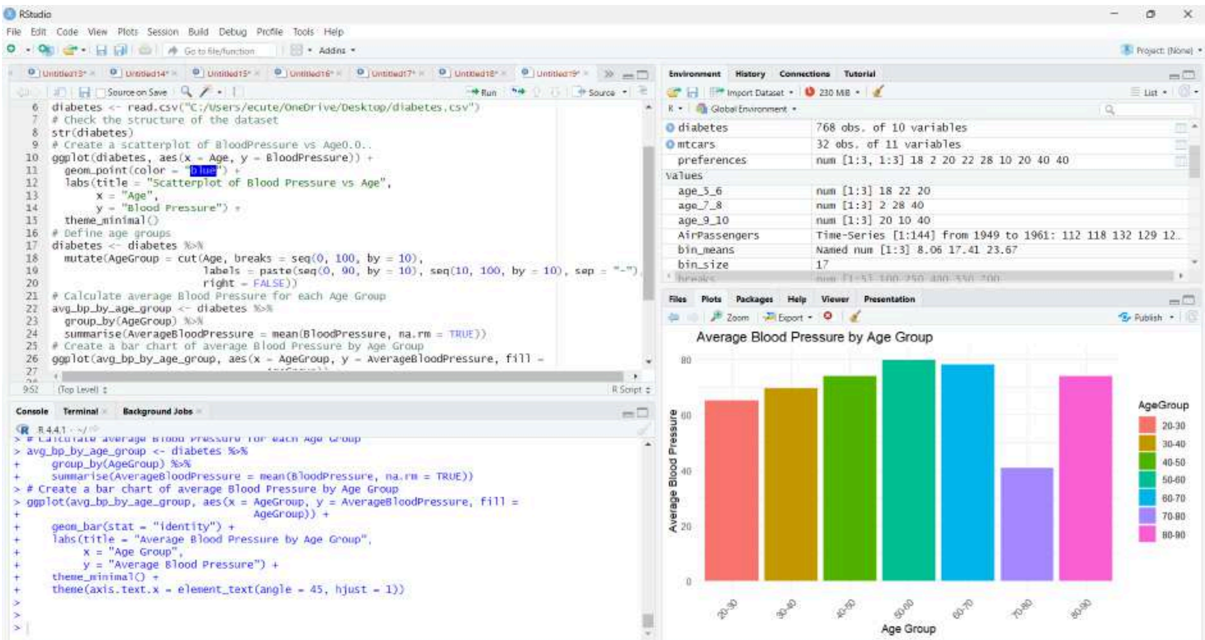
**Fruit, Peanuts, Cheese, Yogurt**



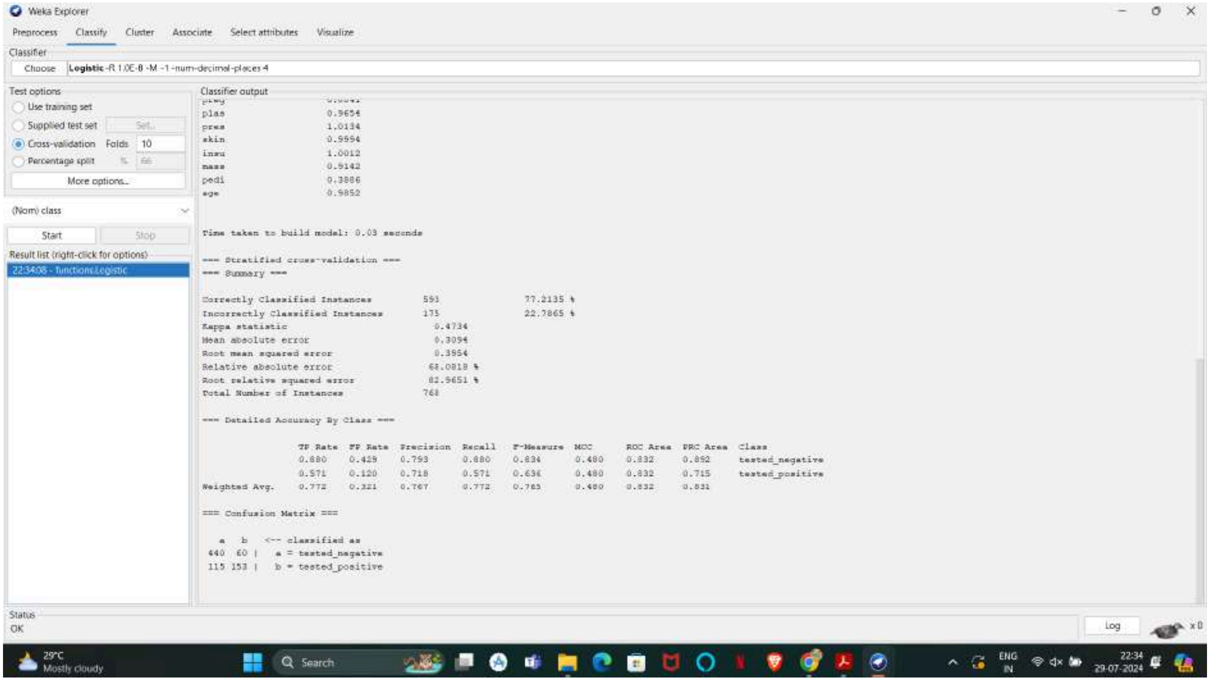

**4) Use following group of data: 200, 300, 400, 600, 1000**
**(a) min-max normalization by setting min = 0 and max = 1 (b)**
**(b) z-score normalization**
**(c) (c) z-score normalization using the mean absolute deviation instead of standard deviation (d) normalization by decimal scaling**

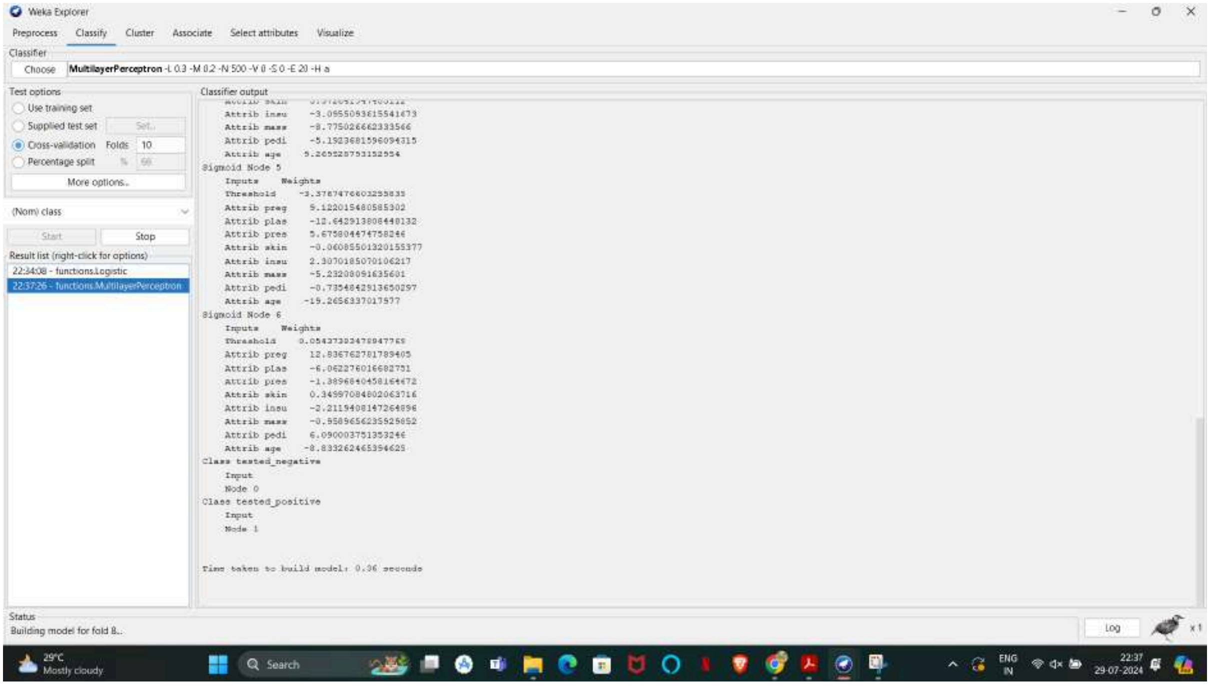


**5) Implement using R language in which age group of people are affected byblood pressure based on the diabetes dataset show it using scatterplot and bar chart (that is BloodPressure vs Age using dataset "diabetes.csv")**



**6) Analysis the dataset "diabetes. csv" how the diabetes trend is for different age people, using linear regression and multiple regression.**




**7) Implement using WEKA for the given Suppose a database has five transactions. Let min sup= 50%(2) and min con f = 80%.**

**Transactions Items**

**T1 (M, O, N, K, E, Y)**

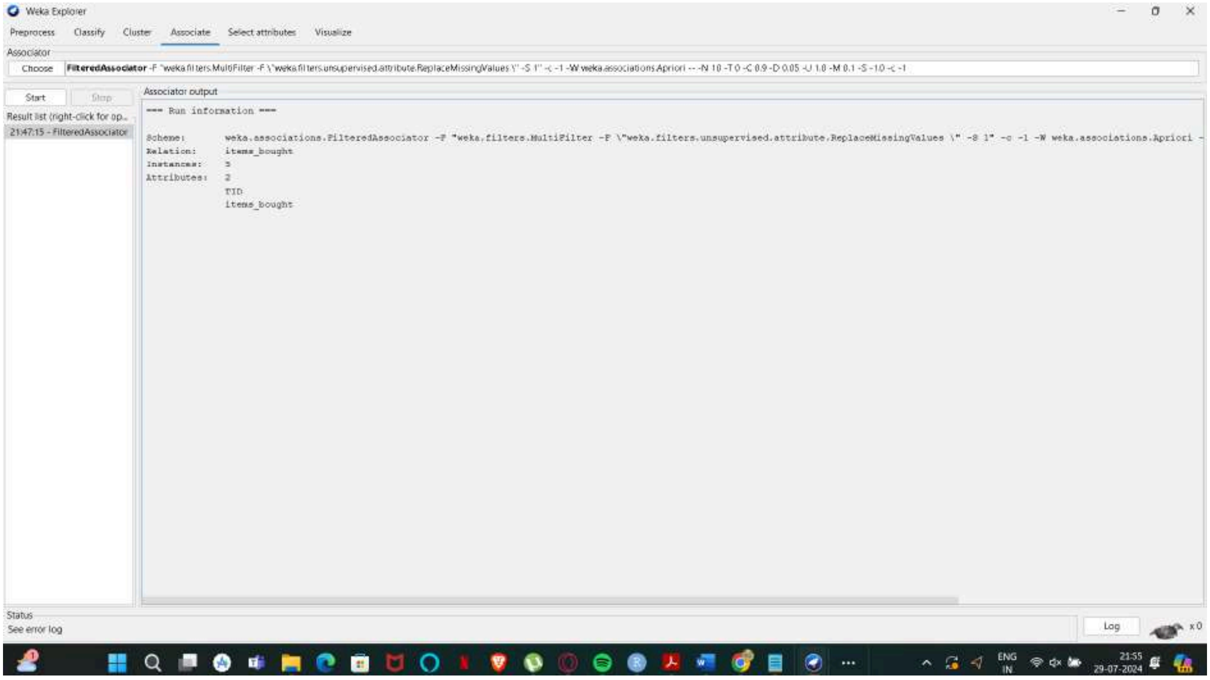
**T2 (D, O, N, K, E, Y)**

**T3 (M, A, K, E)**

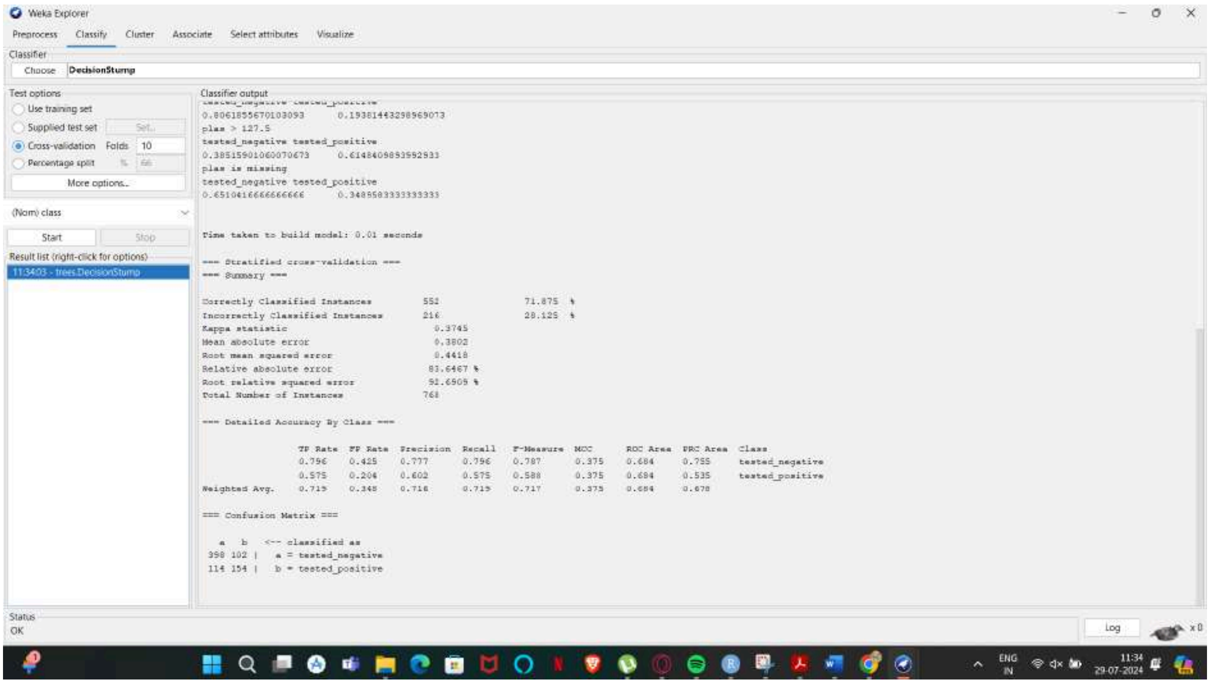
**T4 (M, U, C, K, Y)**

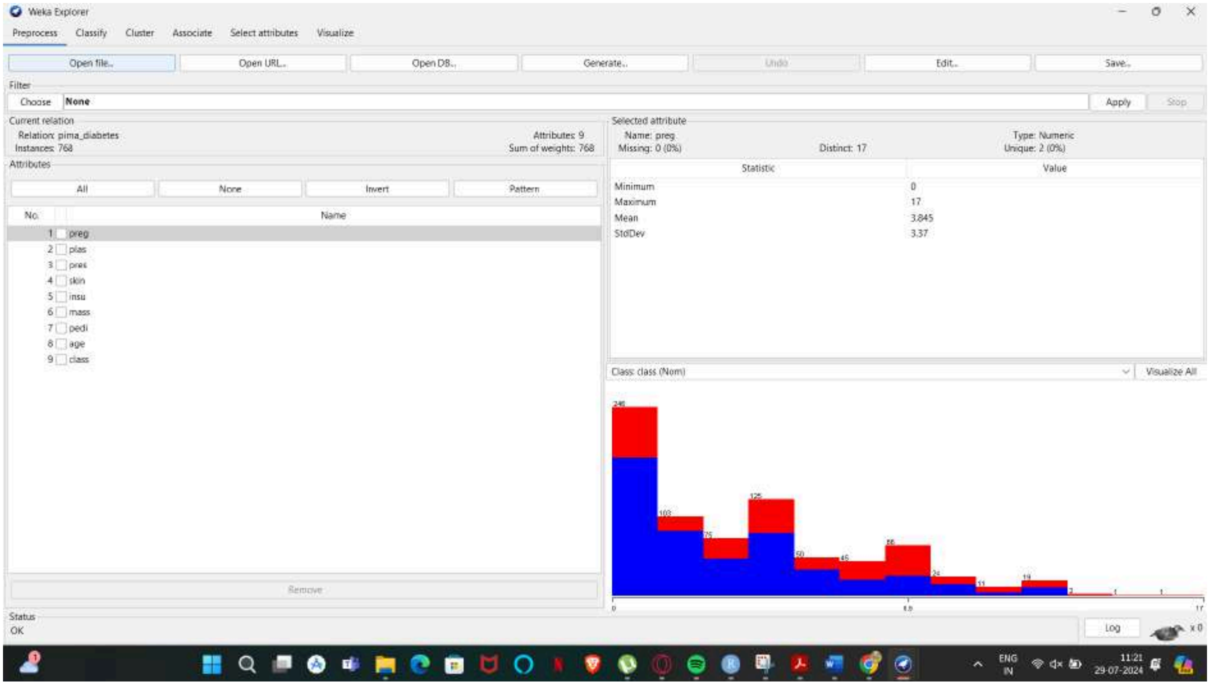
**T5 (C,O, O, K, I ,E)**

- **Find all frequent item sets using Apriori algorithm**
- **Also draw FP-Growth Tree**

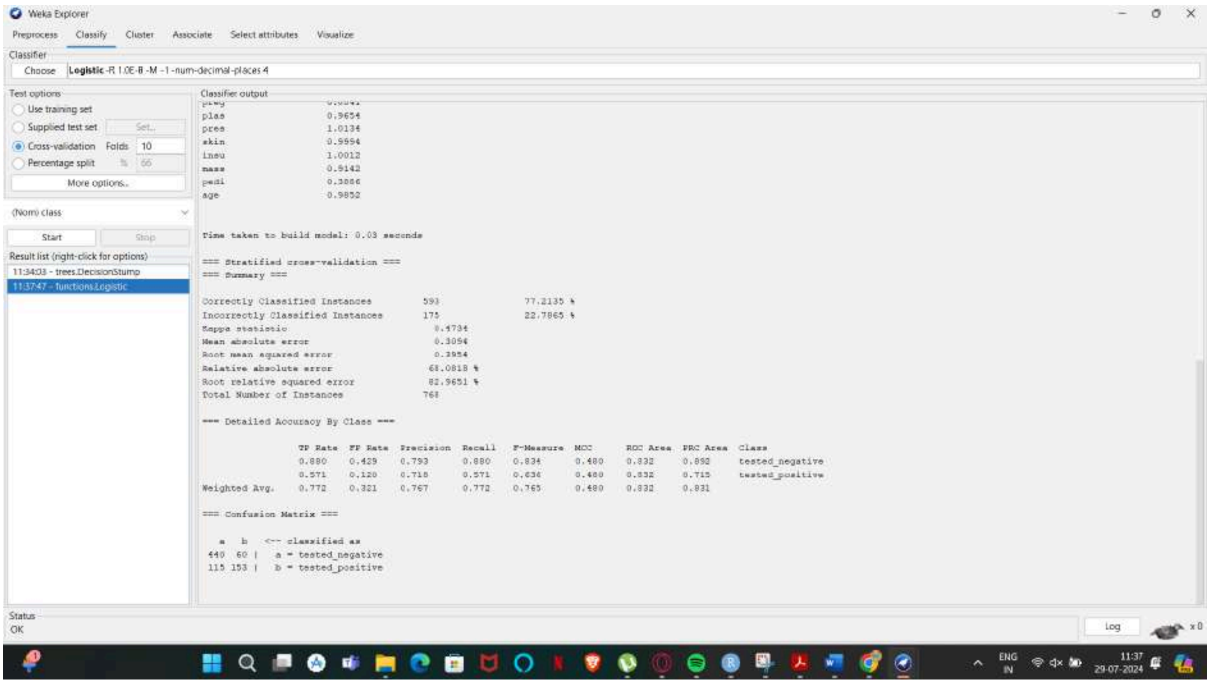



**8) Prediction of Categorical Data using Decision Tree Algorithm through WEKA using any datasets. a) Tree b) Preprocess c) Logistic**



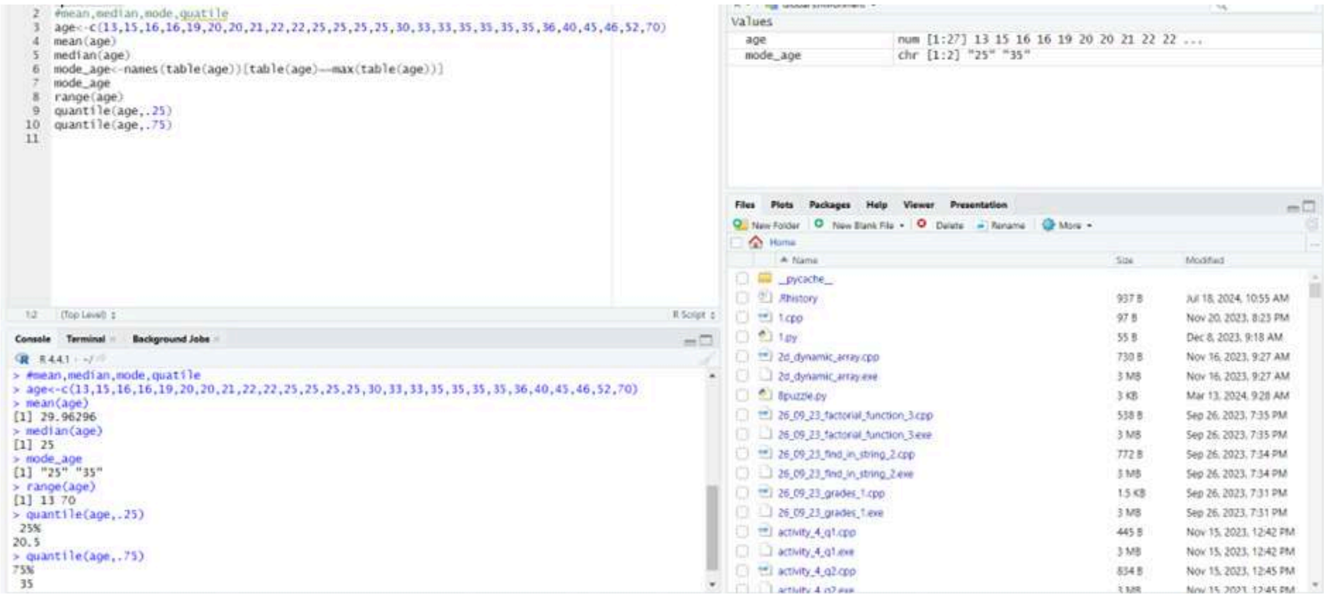


Weka Explorer
Preprocess Classify Cluster Associate Select attributes Visualize
Open file... Open URL... Open DB... Generate... Undo Edit... Save...
Filter
Choose None
Apply Stop
Current relation
Relation: pima_diabetes
Instances: 768
Attributes: 9
Sum of weights: 768
Selected attribute
Name: preg
Missing: 0 (0%)
Distinct: 17
Type: Numeric
Unique: 2 (0%)
Statistic Value
Minimum 0
Maximum 17
Mean 3.845
StdDev 3.37
Attributes
All None Invert Pattern
No. Name
1 preg
2 plas
3 pres
4 skin
5 insu
6 mass
7 pedi
8 age
9 class
Class: class (Nom)
Visualize All
Remove
Status
OK
Log

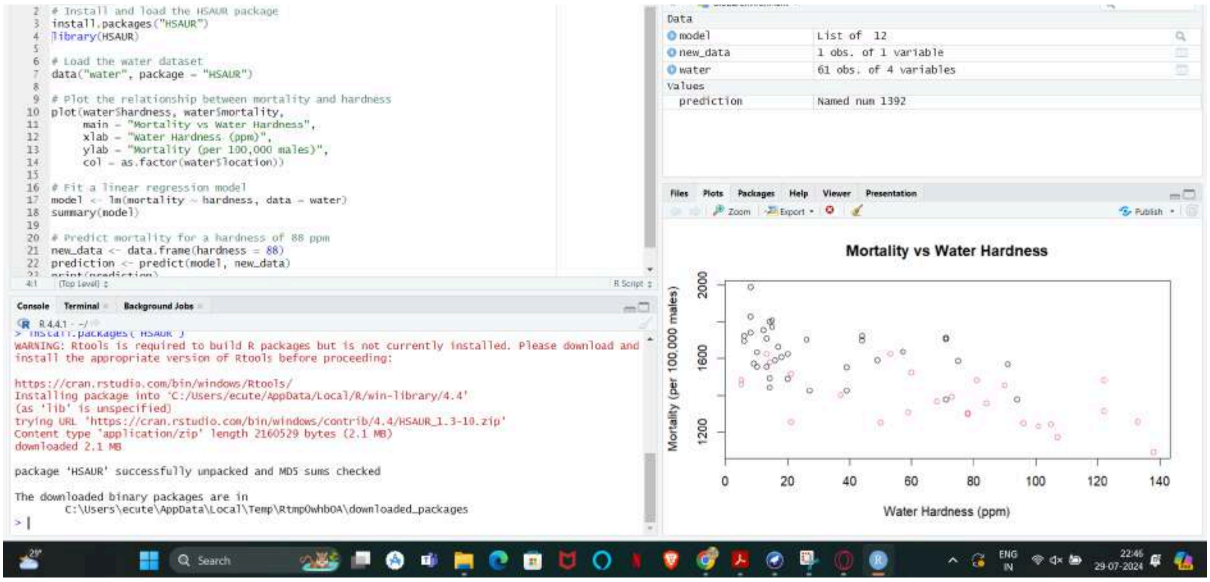



Weka Explorer
Preprocess Classify Cluster Associate Select attributes Visualize
Classifier
Choose Logistic -R 1.0E-8 -M -1 -num-decimal-places 4
Test options
Use training set
Supplied test set
Cross-validation Folds 10
Percentage split % 66
More options...
(Nom) class
Start Stop
Result list (right-click for options)
11:34:03 - trees.DecisionStump
11:37:47 - functions.Logistic
Classifier output
plas 0.9654
pres 1.0134
skin 0.9994
insu 1.0012
mass 0.9142
pedi 0.3866
age 0.9852
Time taken to build model: 0.03 seconds
=== Stratified cross-validation ===
=== Summary ===
Correctly Classified Instances 593 77.2135 %
Incorrectly Classified Instances 175 22.7865 %
Kappa statistic 0.4734
Mean absolute error 0.3094
Root mean squared error 0.3954
Relative absolute error 68.0818 %
Root relative squared error 82.9651 %
Total Number of Instances 768
=== Detailed Accuracy By Class ===
TP Rate FP Rate Precision Recall F-Measure MCC ROC Area PRC Area Class
0.880 0.429 0.793 0.880 0.834 0.480 0.832 0.892 tested_negative
0.571 0.120 0.718 0.571 0.636 0.480 0.832 0.715 tested_positive
Weighted Avg. 0.772 0.321 0.767 0.772 0.765 0.480 0.832 0.831
=== Confusion Matrix ===
a b <-- classified as
440 60 | a = tested_negative
115 153 | b = tested_positive
Status
OK
Log

**9) Suppose that the data for analysis includes the attribute age. The age values for the data tuples are (in increasing order) 13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70.**
**Can you find (roughly) the first quartile (Q1) and the third quartile (Q3) of the data?**



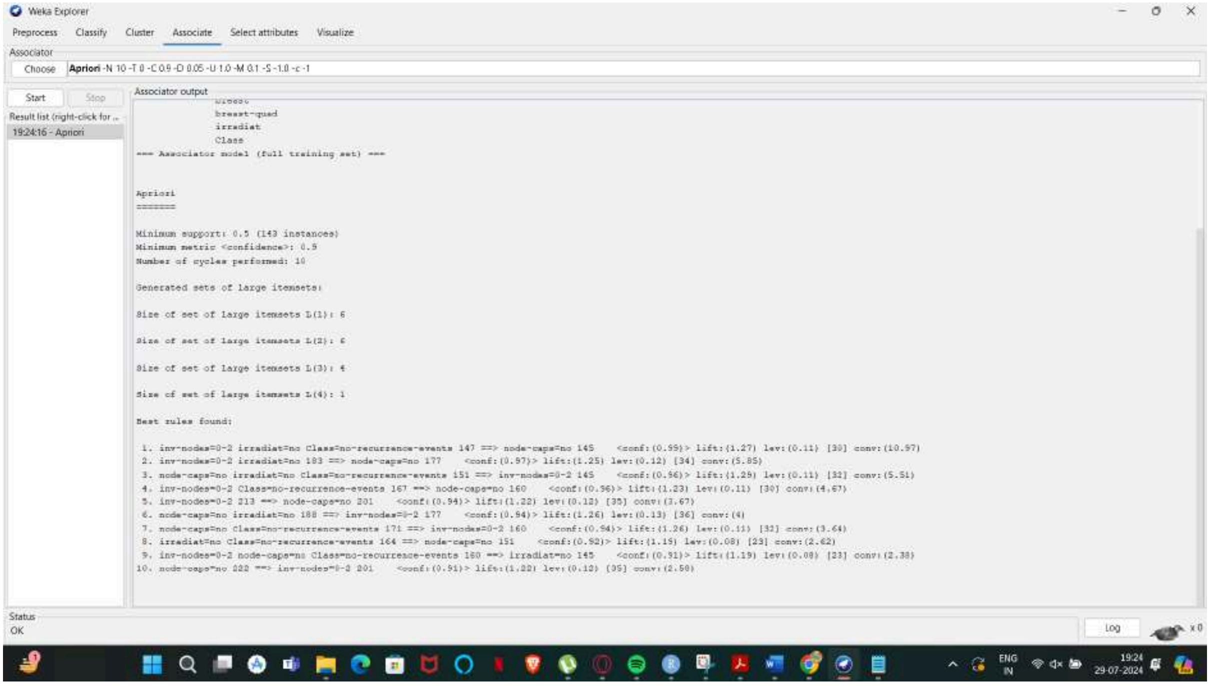
**10) Download the Dataset "water" From R dataset Link.Find out whether there is a linear relation between attributes"mortality" and"hardness" by plot function.Fit the Data into the Linear Regression model. Predict the mortality for the hardness=88.**

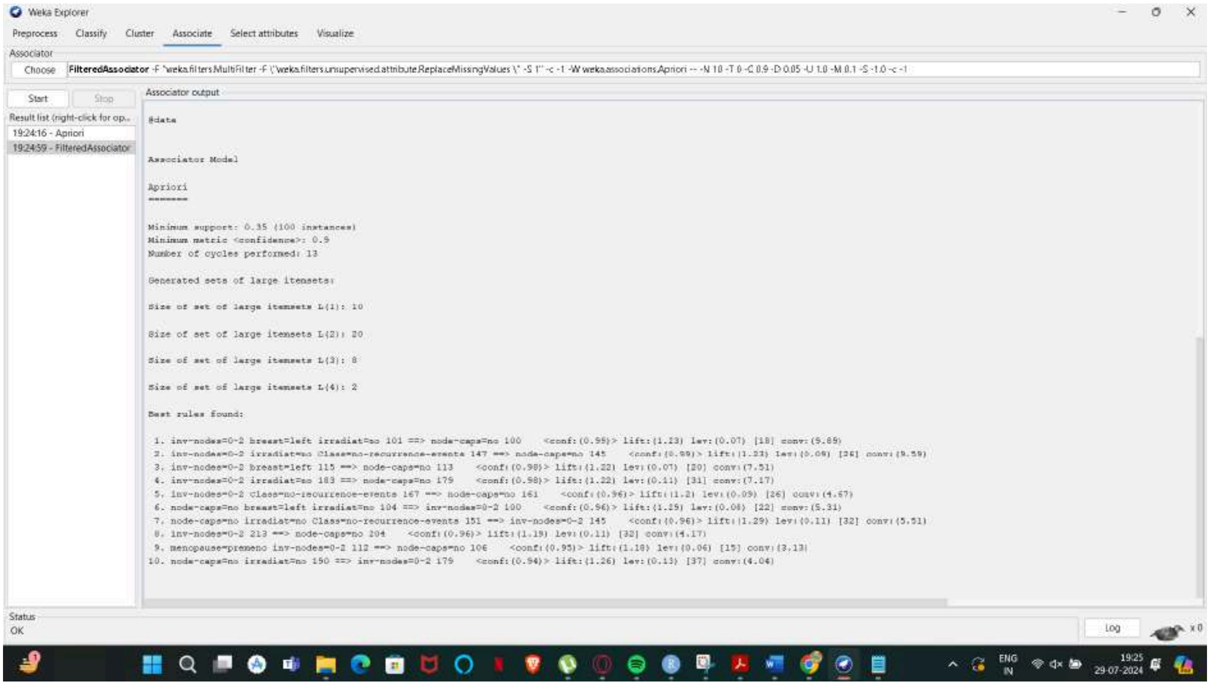



**11) Create the dataset using ARFF file format:**

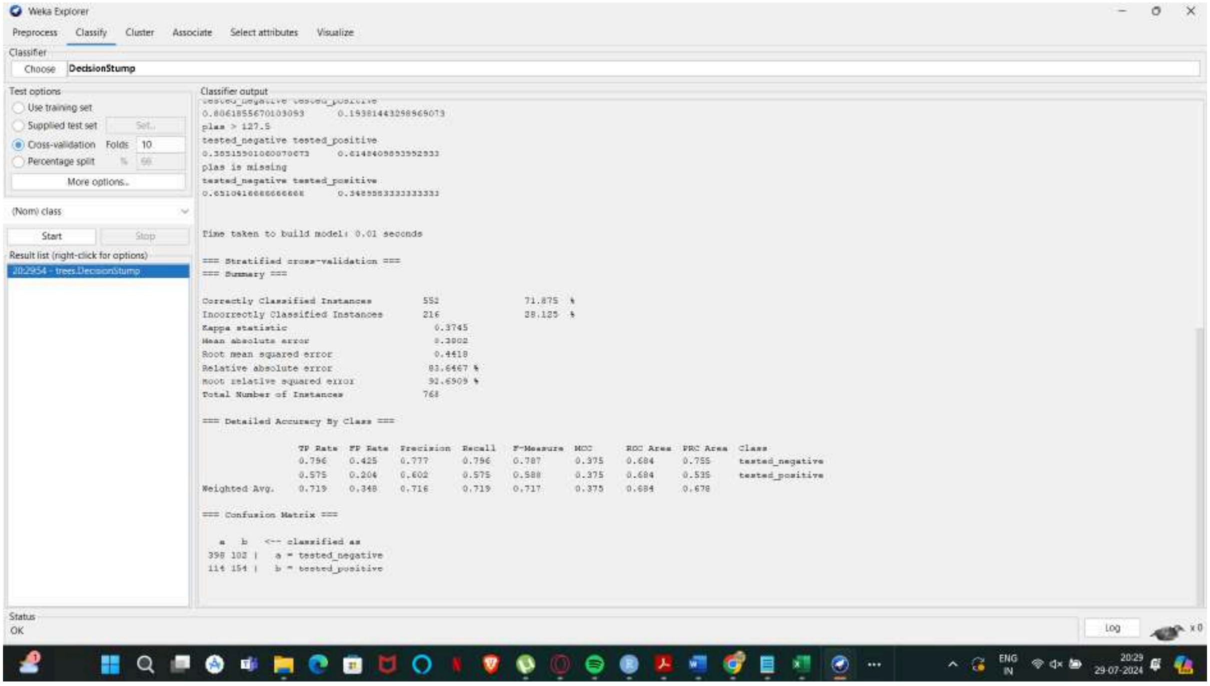
**a.Find the frequent itemsets and generate association rules on this. Assume that minimum support threshold (s = 33.33%) and minimum confident threshold (c = 60%).**

**b.List the various rule generated by apriori and FP tree algorthim ,mention wheather accepted or rejcted.**

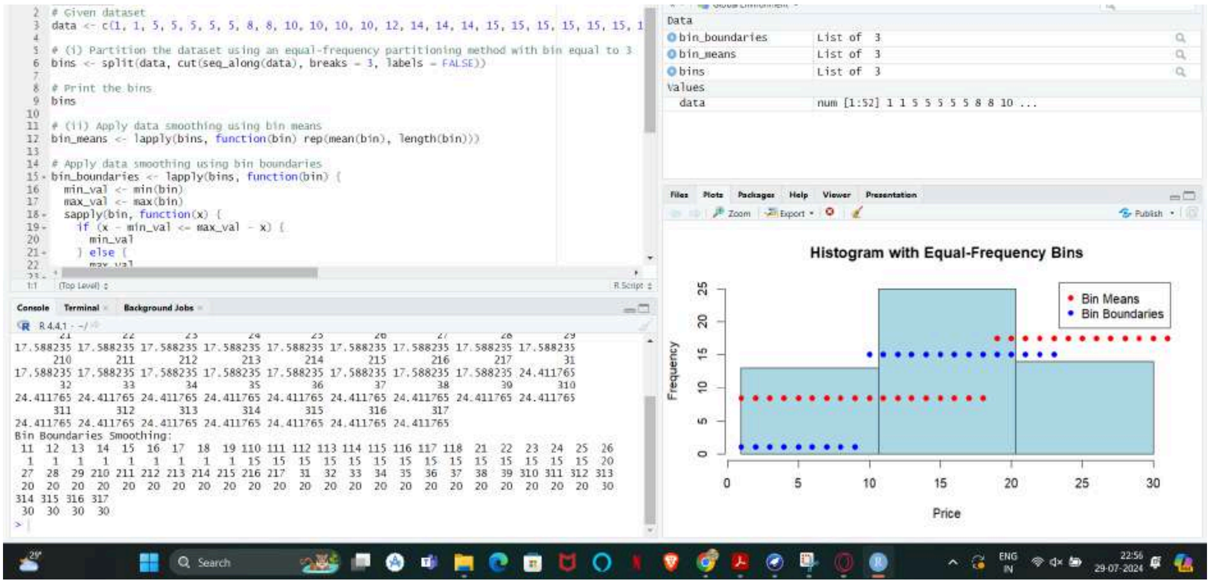


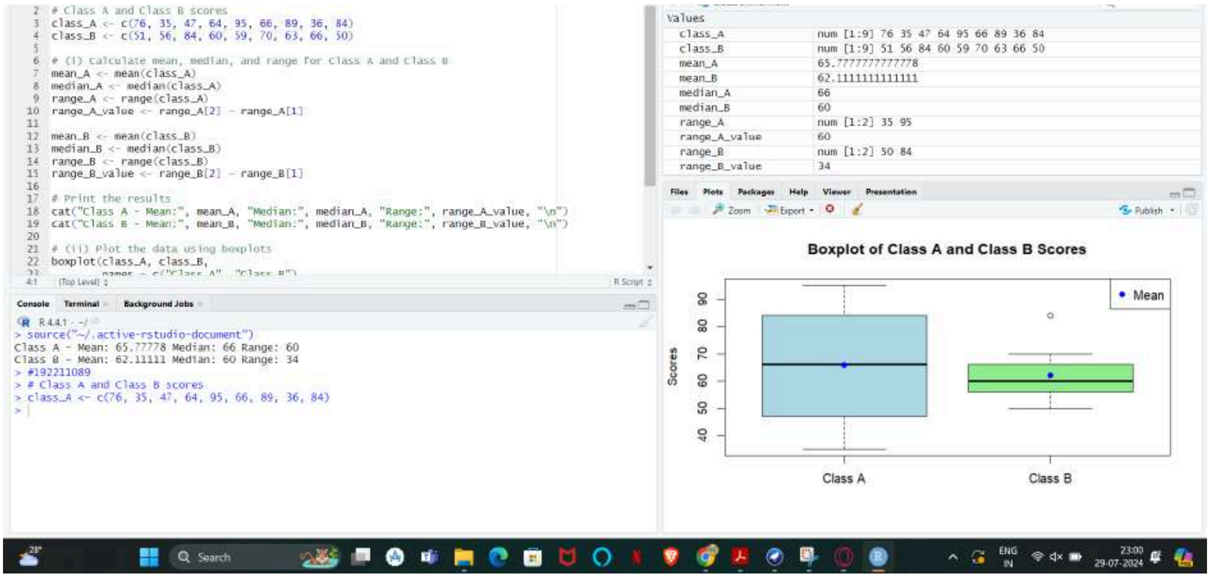
**12) Prediction of Categorical Data using Rule base classification and decision tree classification through WEKA using any datasets. Compare the accuracy using two algorithm and plot the graph**



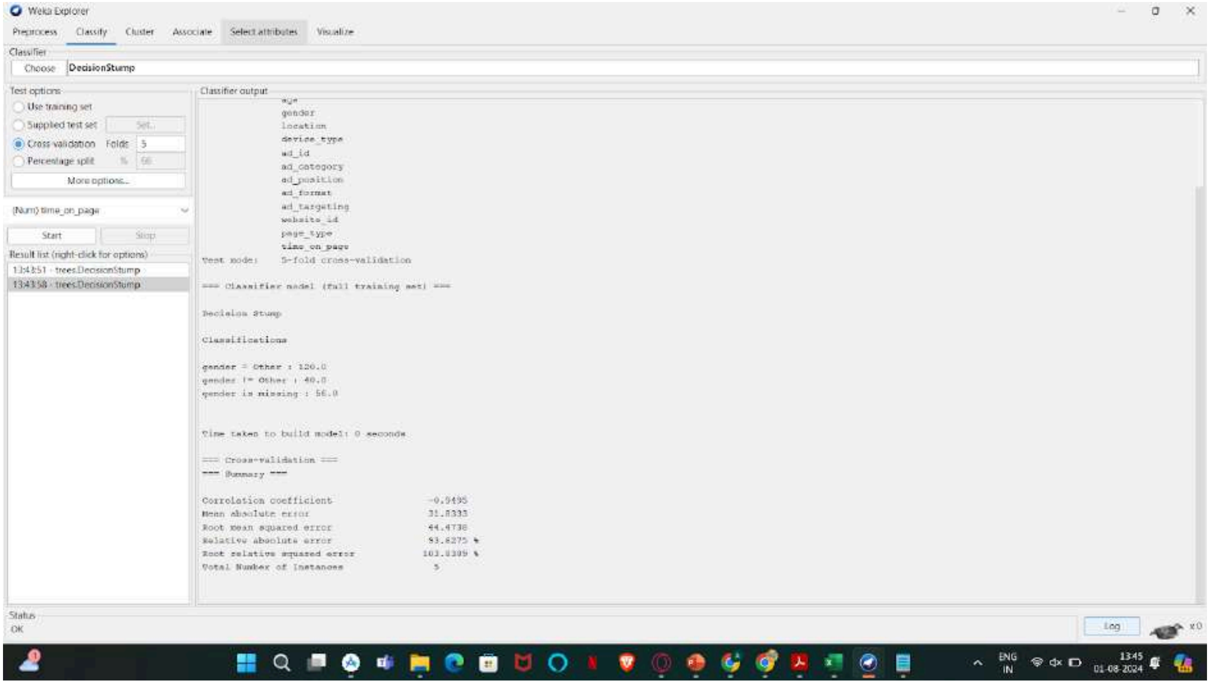

13) **Imagine that you have selected data from the All Electronics data warehouse for analysis. The data set will be huge! The following data are a list of All Electronics prices for commonly sold items (rounded to the nearest dollar). The numbers have been sorted: 1, 1, 5, 5, 5, 5, 5, 8, 8, 10, 10, 10, 10, 12, 14, 14, 14, 15, 15, 15, 15, 15, 15, 18, 18, 18, 18, 18, 18, 18, 18, 20, 20, 20, 20, 20, 20, 20, 21, 21, 21, 21, 25, 25, 25, 25, 25, 28, 28, 30, 30, 30. (i) Partition the dataset using an equal-frequency partitioning method with bin equal to 3 (ii) apply data smoothing using bin means and bin boundary. (iii) Plot Histogram for the above frequency division**



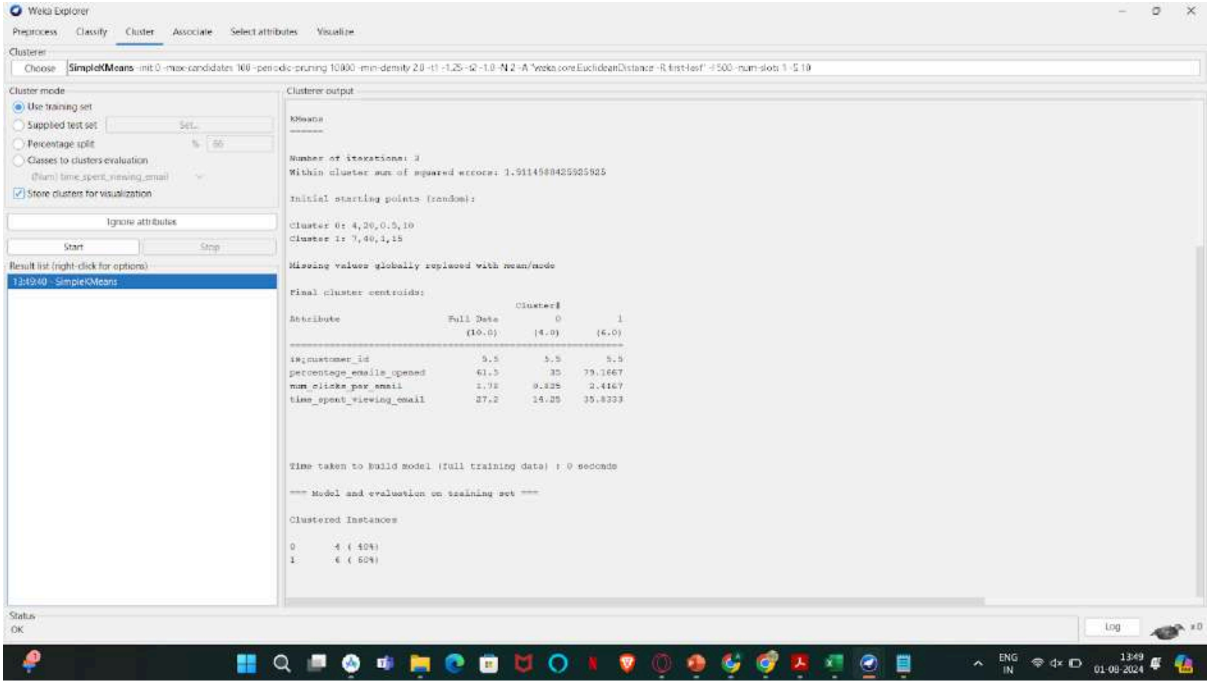

14) **Two Maths teachers are comparing how their Year 9 classes performed in the end of year exams. Their results are as follows: Class A: 76, 35, 47, 64, 95, 66, 89, 36, 8476,35,47,64,95,66,89,36,84 Class B: 51, 56, 84, 60, 59, 70, 63, 66, 5051,56,84,60,59,70,63,66,50 (i) Find which class had scored higher mean, median and range. (ii) Plot above in boxplot and give the inferences**



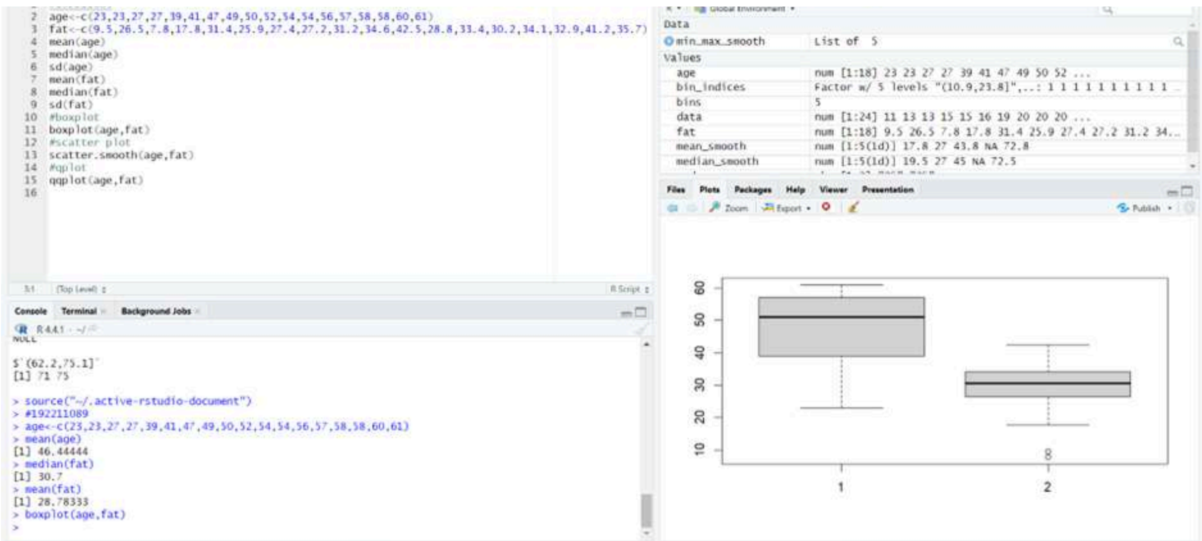

15) **Consider a Binary classification model that can be used to predict whether one or more ads on the website will be clicked or not. The models are used to optimize the ad inventory on websites by selecting which ads will have a better chance of being clicked.**



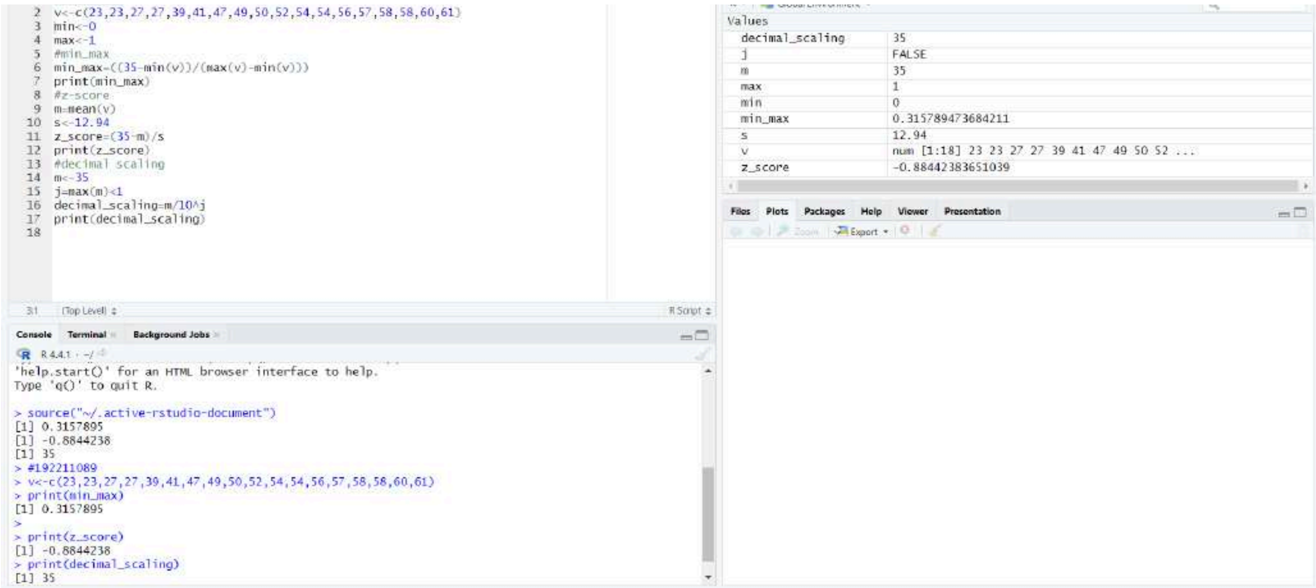
16) **Consider that Many businesses use cluster analysis to identify consumers who are similar to each other so they can tailor their emails sent to consumers in such a way that maximizes their revenue. Consider a business may collect the following information about consumers: Percentage of emails opened Number of clicks per email Time spent viewing email Using these metrics, a business can perform various cluster analyses to identify consumers who use email in similar ways and tailor the types of emails and frequency of emails they send to different clusters of customers. Compare the performance of the applied clustering algorithm.**



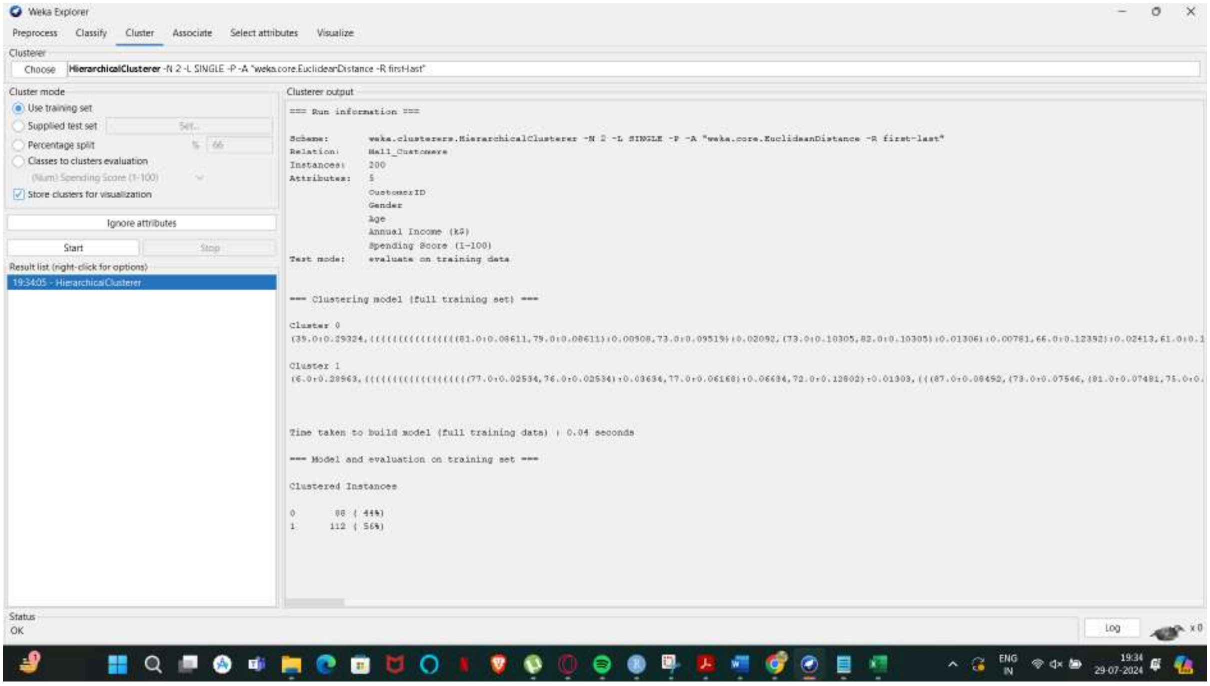

**17) Suppose that a hospital tested the age and body fat data for 18 randomly selected adults with the following results:**

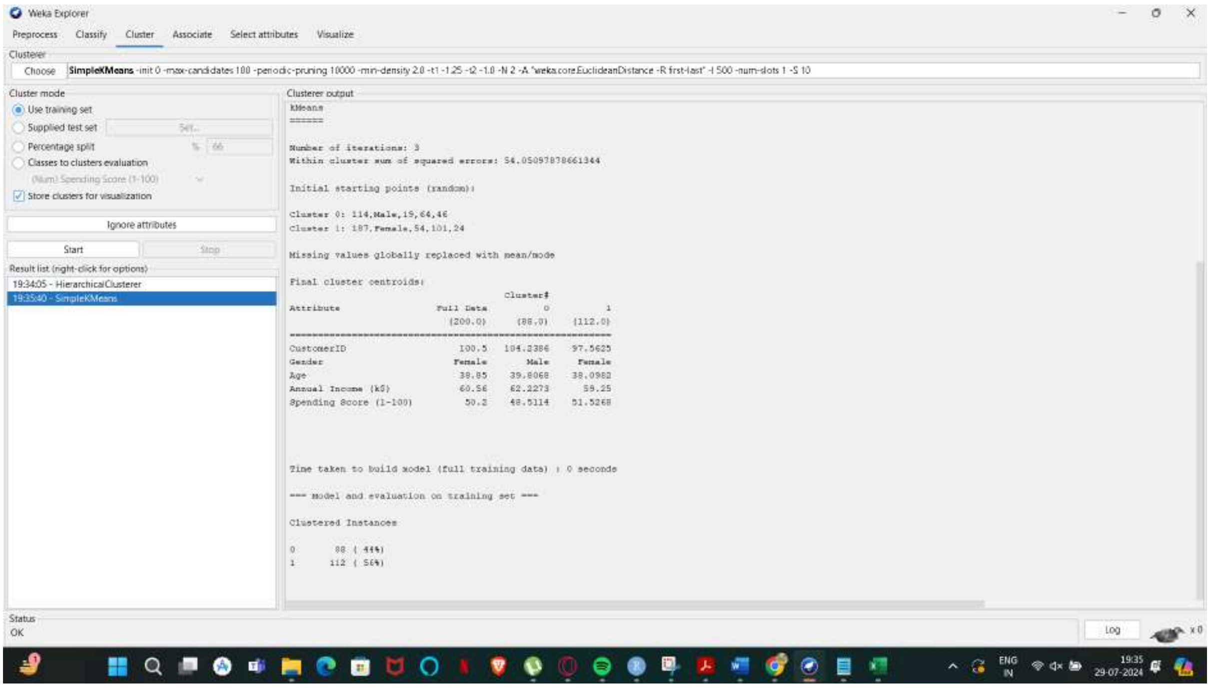



18) Suppose that a hospital tested the age and body fat data for 18 randomly selected adults with the following results:




**19) Consider that you are owning a supermarket mall and through membership cards, you have some basic data about your customers like Customer ID, age, gender, annual income and spending score. For the above scenario, the Problem Statement was You want to understand the customers who can easily converge [Target Customers] so that the data can be given to the marketing team and plan the strategy accordingly. For the above scenario prepare a dataset and perform Clustering Analysis to segment the customers in the Mall. There are clearly Five segments of Customers based on their Annual Income and Spending Score namely Usual Customers, Priority Customers, Senior Citizen Target Customers, and Young Target Customers.Sample data**

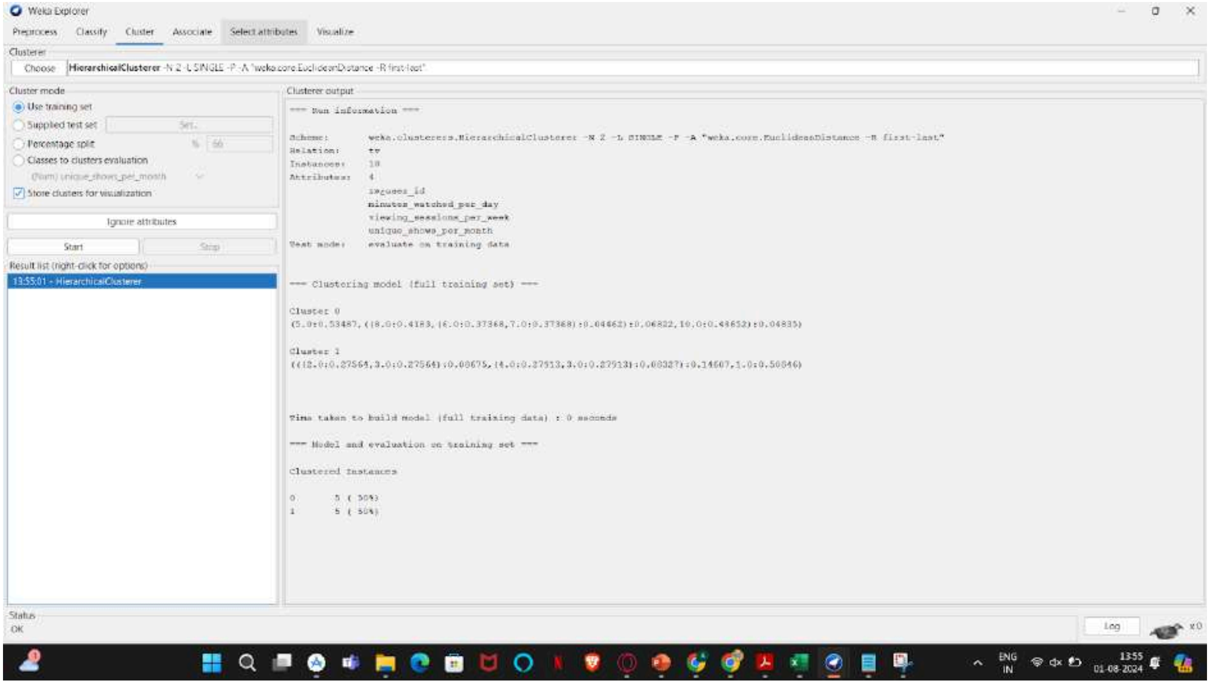



**20) Streaming services often use clustering analysis to identify viewers who have similar behavior. The streaming service may collect the following data about individuals: Minutes watched per day**

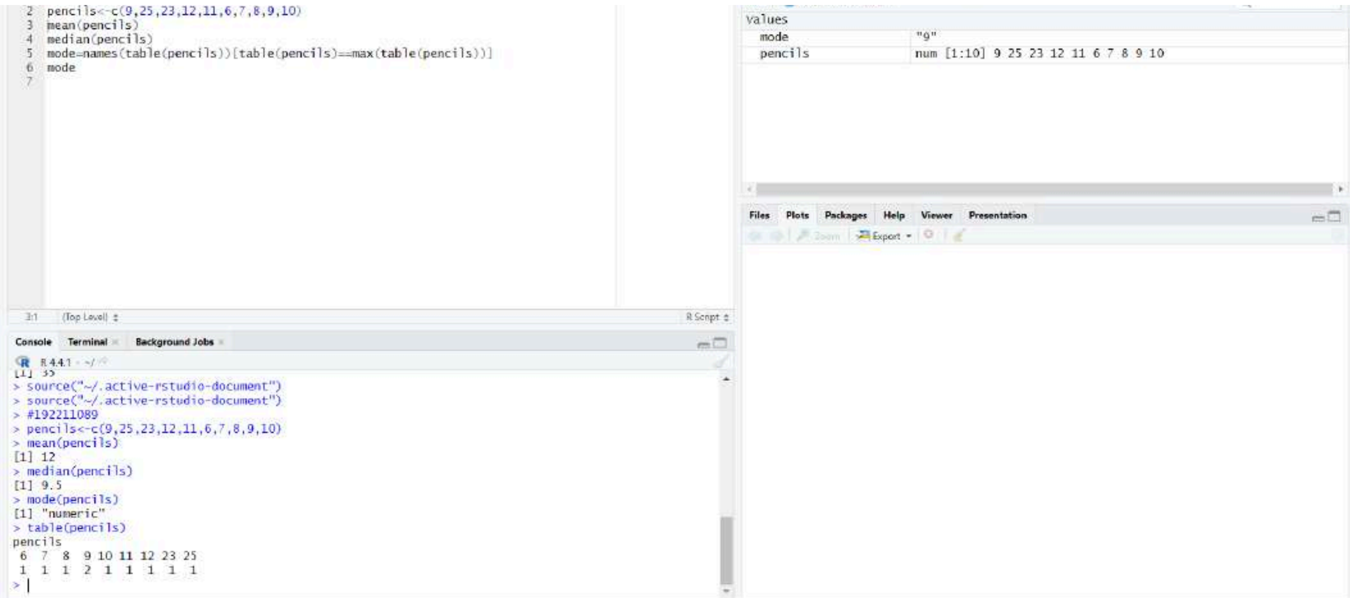
**Total viewing sessions per week Number of unique shows viewed per month Using these metrics, a streaming service can perform cluster analysis to identify high-usage and**

**low-usage users so that they can know whom they should spend most of their advertising dollars on. Apply the Hierarchical clustering algorithm and EM clustering algorithm to identify and compare the performance of the clustering technique.**

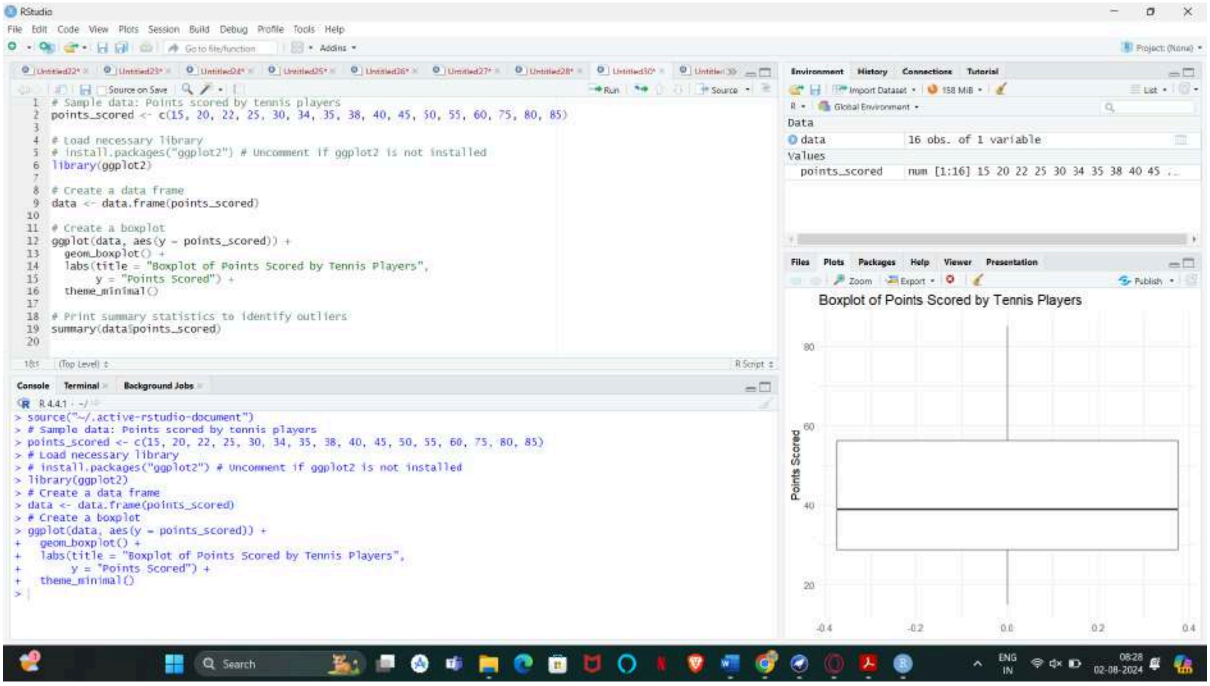


**21)The following values are the number of pencils available in the different boxes. Create a**

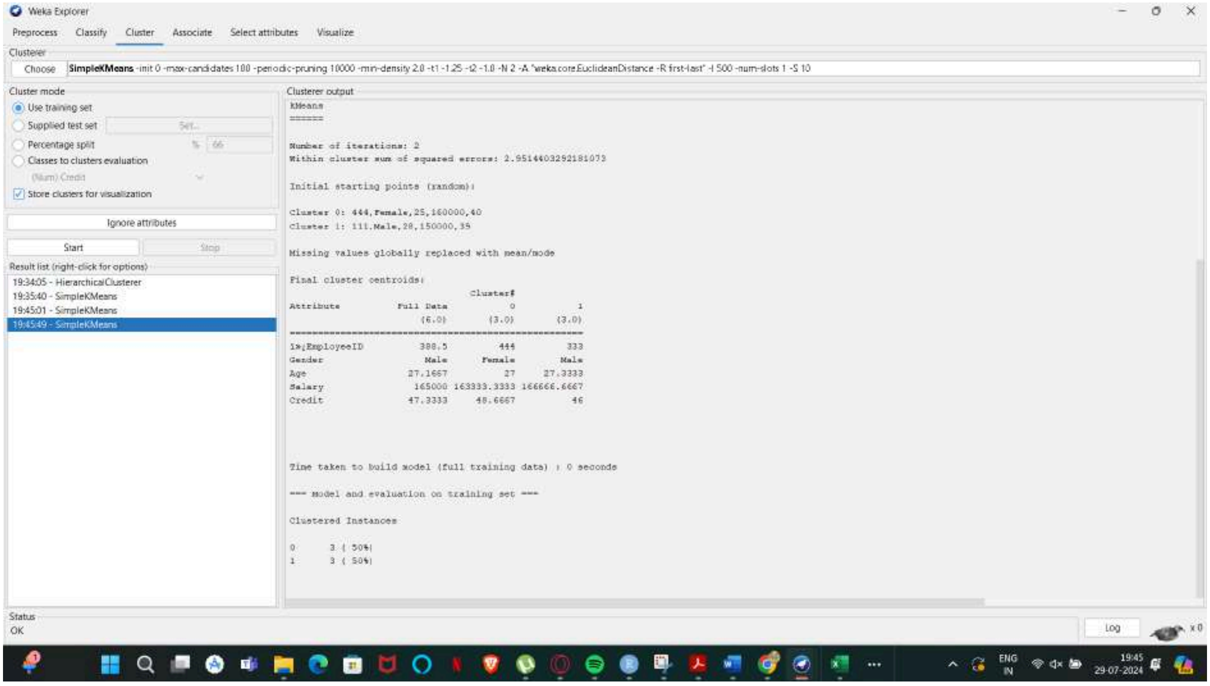
**vector and find out the mean, median and mode values of set of pencils in the given data.**



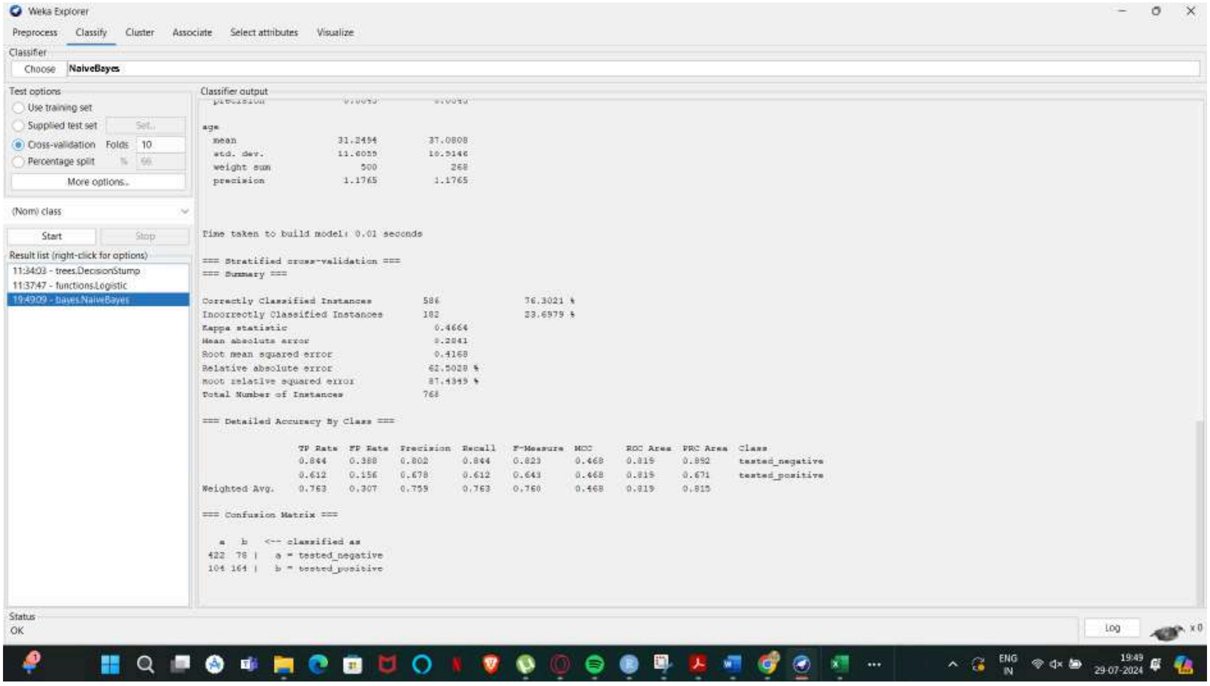
**22) Assume the Tennis coach wants to determine if any of his team players are scoring outliers. To visualize the distribution of points scored by his players, then how can he decide to develop the box plot? Give suitable example using Boxplot visualization technique.**

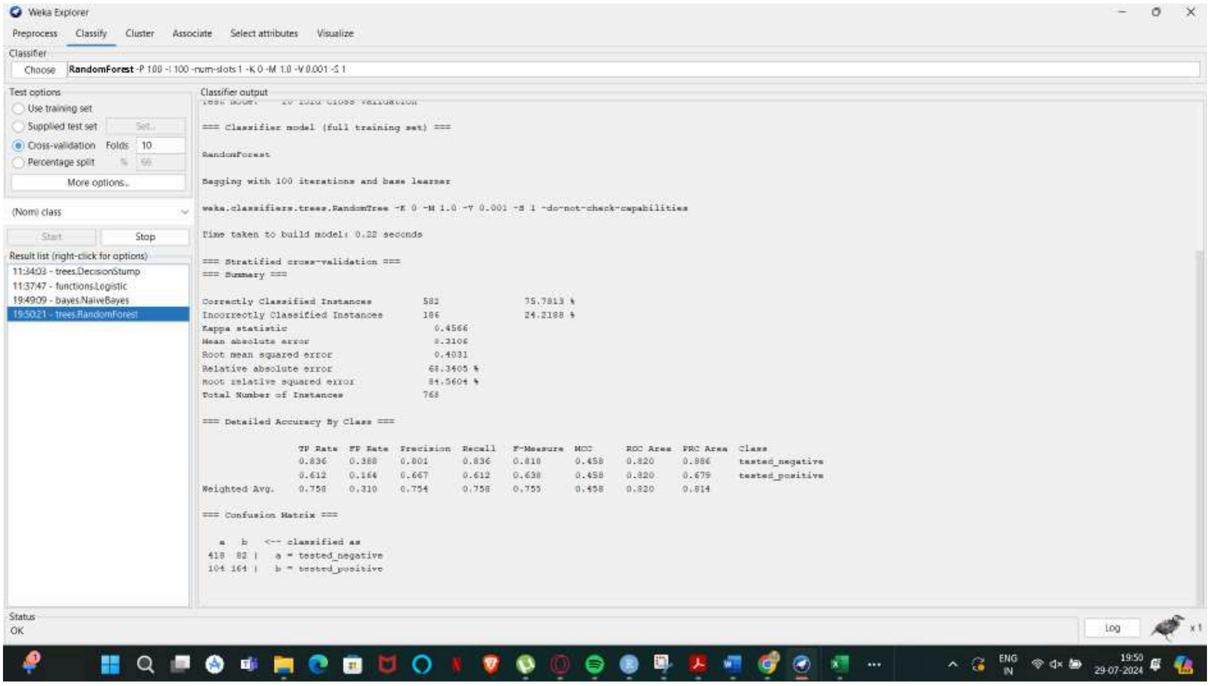



**23) Create the following dataset using CSV file format. To perform cluster analysis using K- Means in WEKA. To change the cluster size and plot the graph and illustrate the visualization of cluster.**

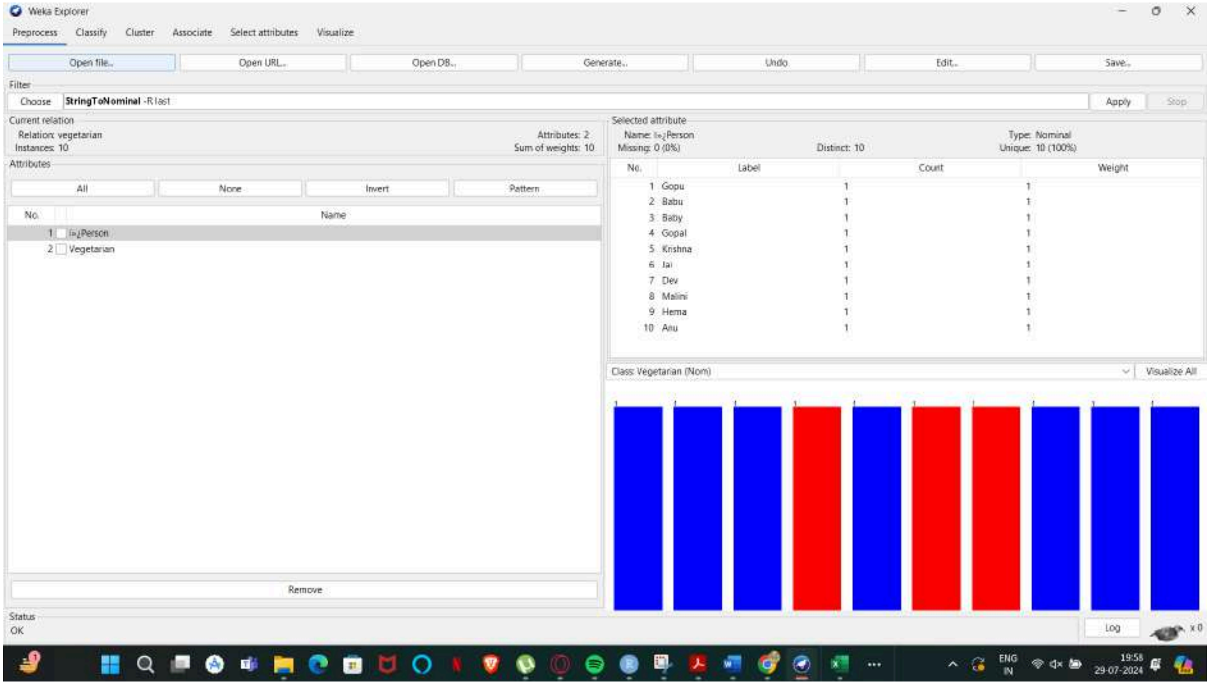


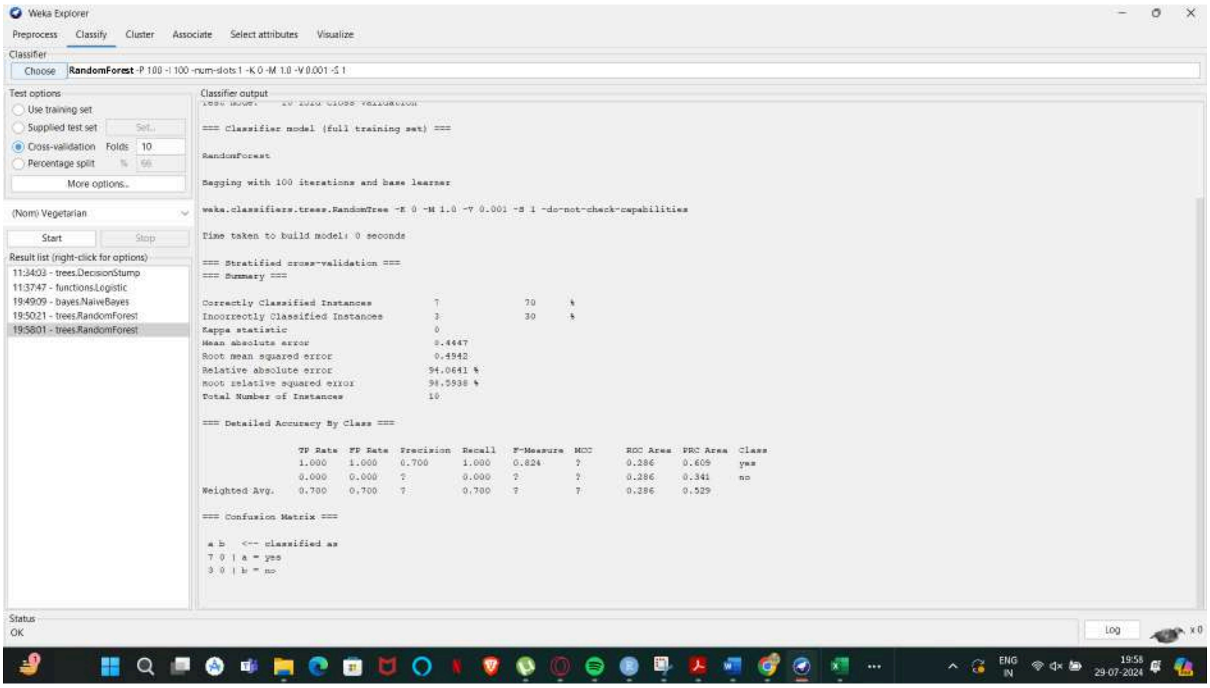
**24) Prediction of categorical data using Naïve Bayes classification through WEKA using any datasets. Compare the Naïve Bayes algorithm with SVM using the summary of results given by the classifiers and plot the graph.**



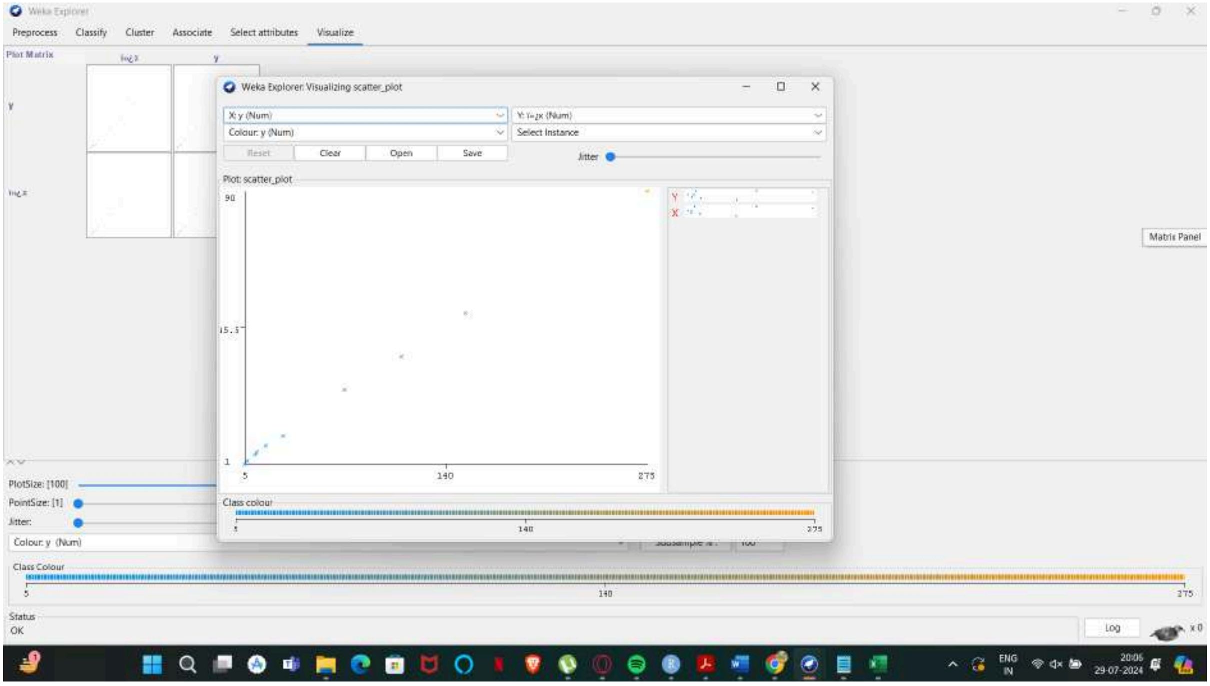




**25) The following list of persons with vegetarian or not details given in the table. How will you find out how many of them are vegetarian and how many of them are non-vegetarian? Which type of the person total count is greater value?**

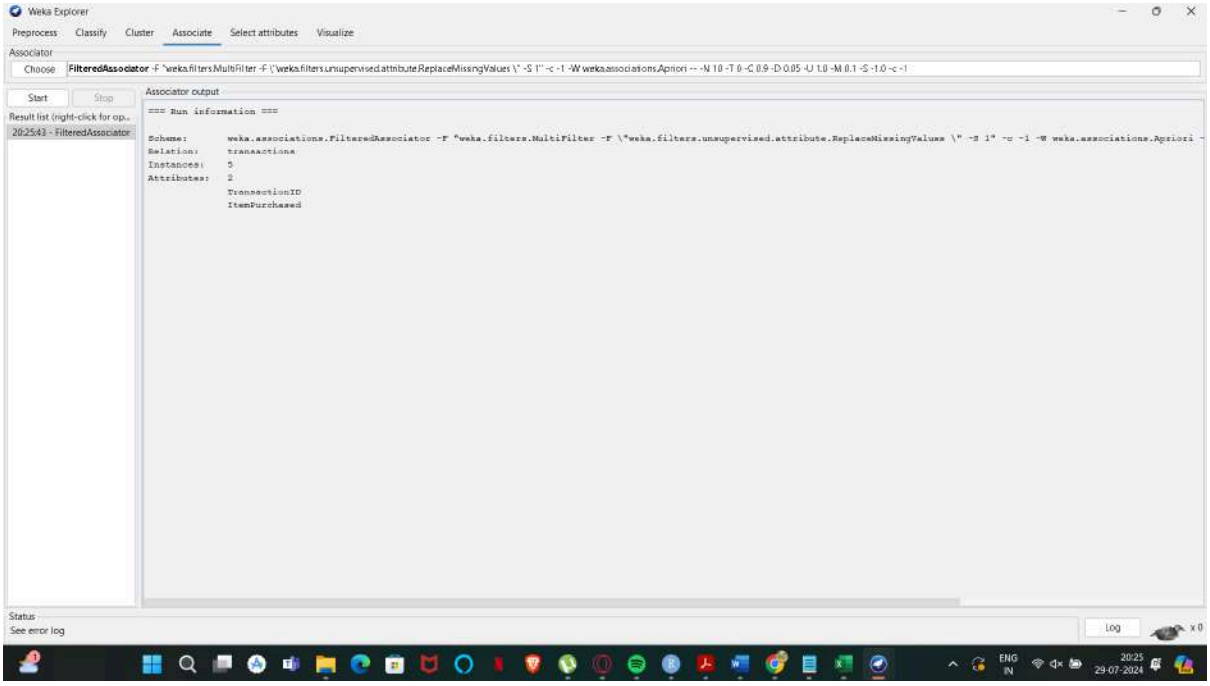




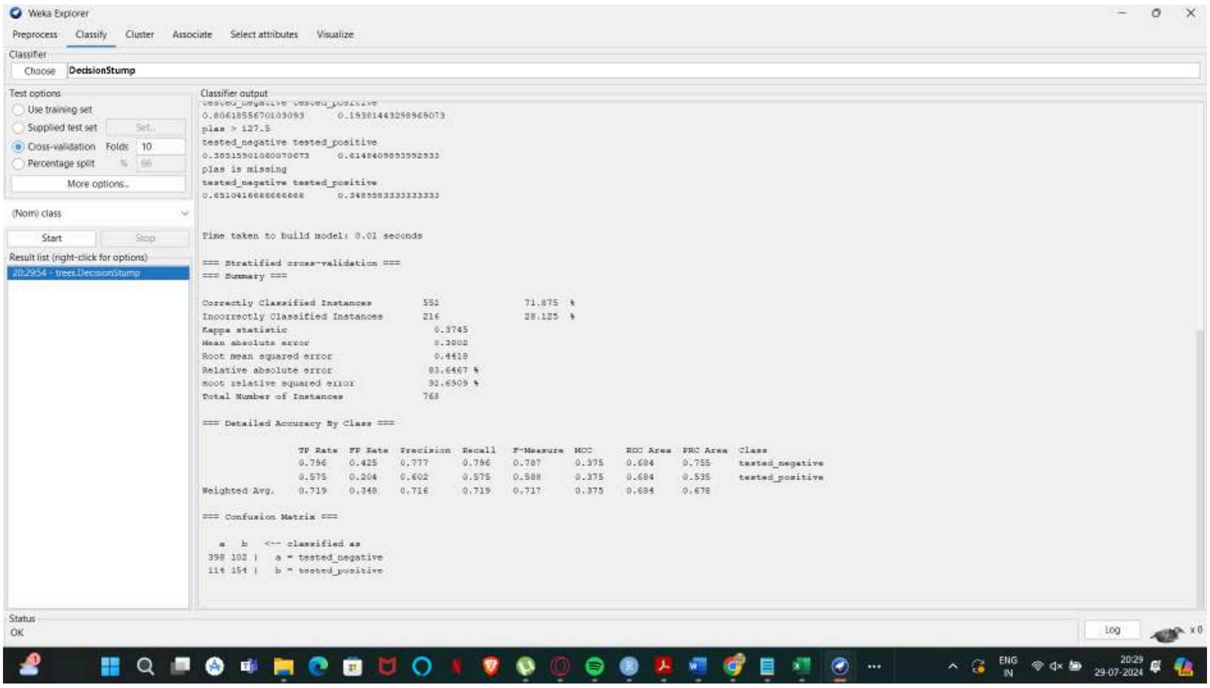
**26) The following table would be plotted as (x,y) points, with the first column being the x values as number of mobile phones sold and the second column being the y values as money. To use the scatter plot for how many mobile phones sold.**

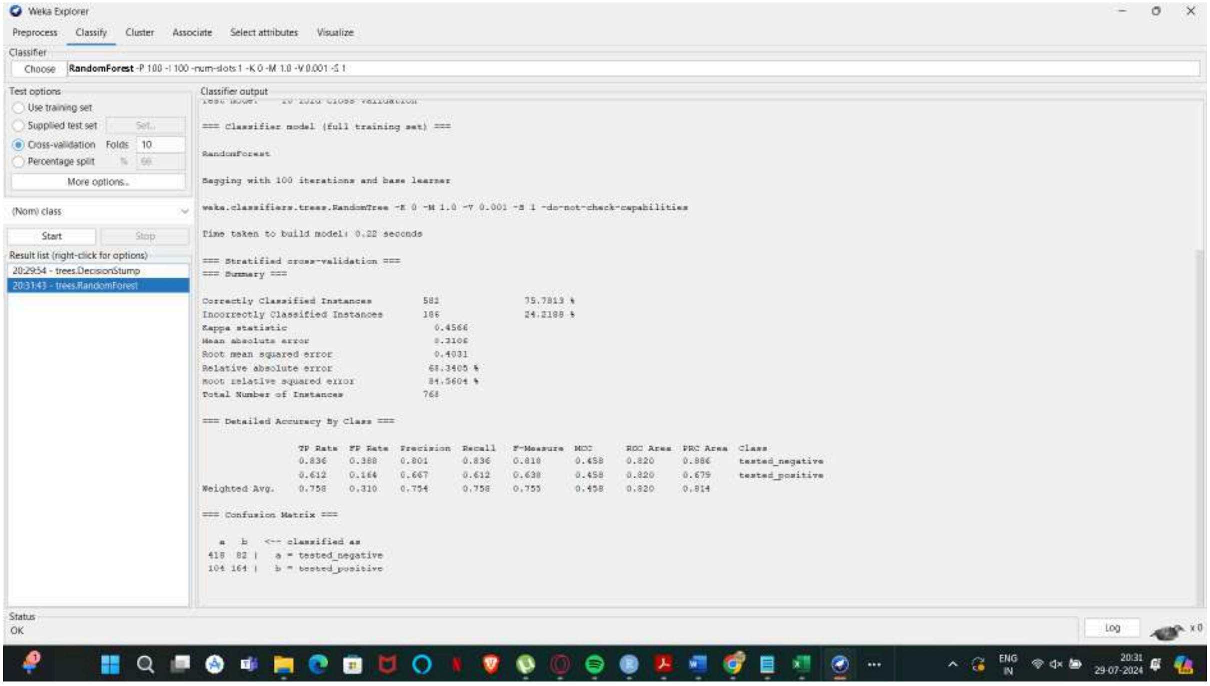


**27) Generate rules using FP growth algorithm using the given dataset which has the following transactions with items purchased: Consider the values as support=50% and confidence=75%.**




**28) Prediction of Diabetes Data using Decision tree classifier in WEKA. Compare it with Support Vector Machine classifier. Show the result accuracy and F1 measure calculation .Plot the graph and explain the summary of results.**



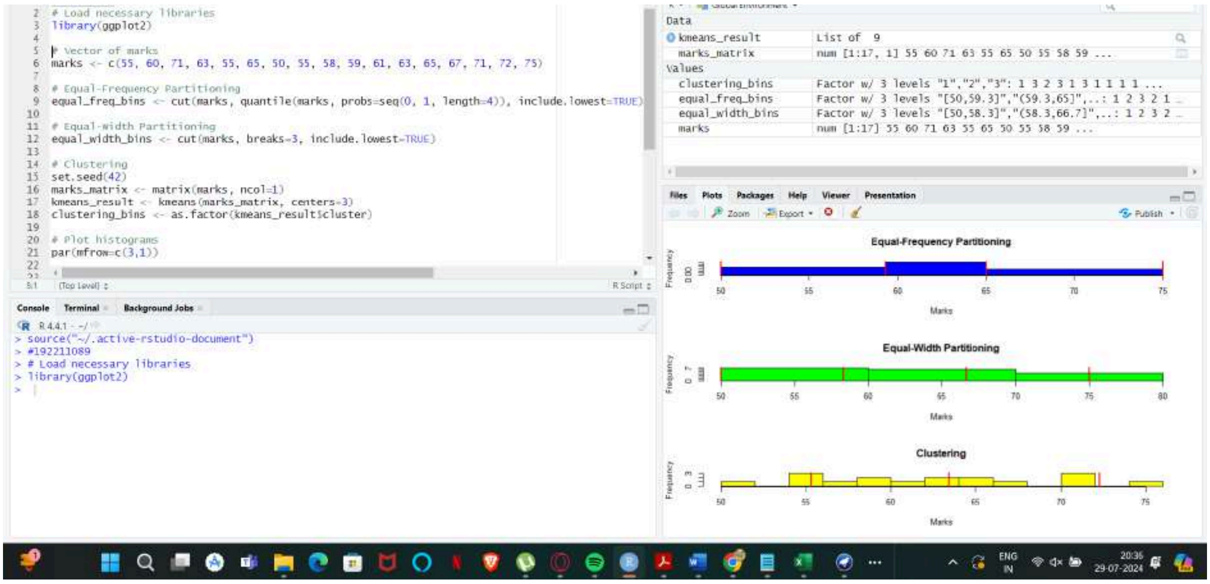



**29) Implement of the R script using marks scored by a student in his model exam has been sorted as follows: 55, 60, 71, 63, 55, 65, 50, 55,58,59,61,63,65,67,71,72,75. Partition them into three bins by each of the following methods. Plot the data points using histogram.**

**(a) equal-frequency (equi-depth) partitioning**

**(b) equal-width partitioning**

**(c) clustering**

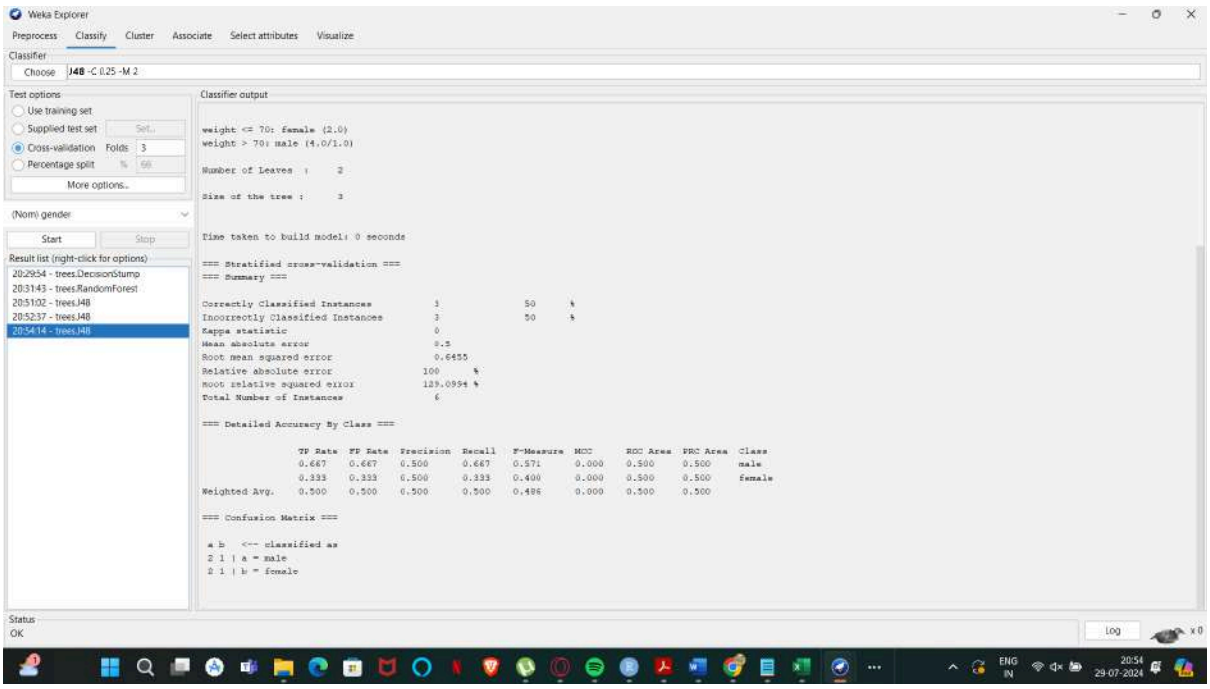



30) **Consider this Decision tree :**

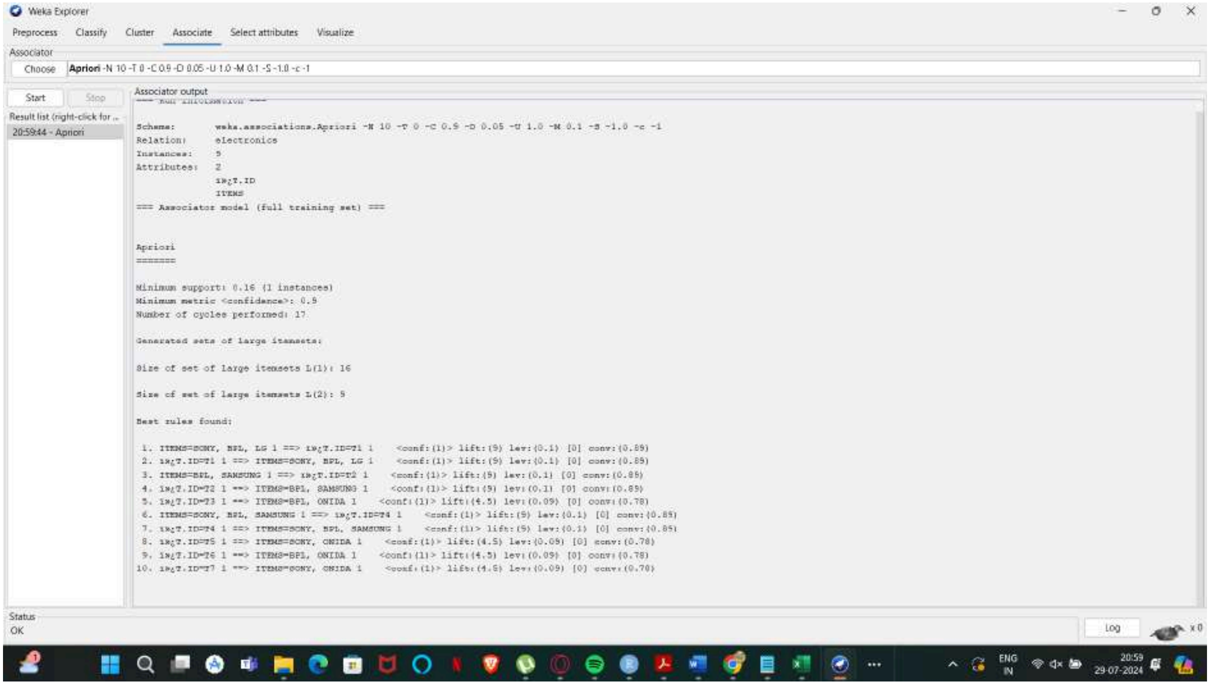
**a)create the data set for the below tree using ARFF format and calculate accuracy and decision for the same**

**b) Using this decision tree generate the rules based on rule based induction.**

**c) Compare both the algorithms and plot the confusion matrix.**



**31) Create an ARFF file for the table below and implement for the Apriori Algorithm and FP growth algorithm and compare the rules generated by both the algorithms. Identify the unique rules generated by the above algorithms.**



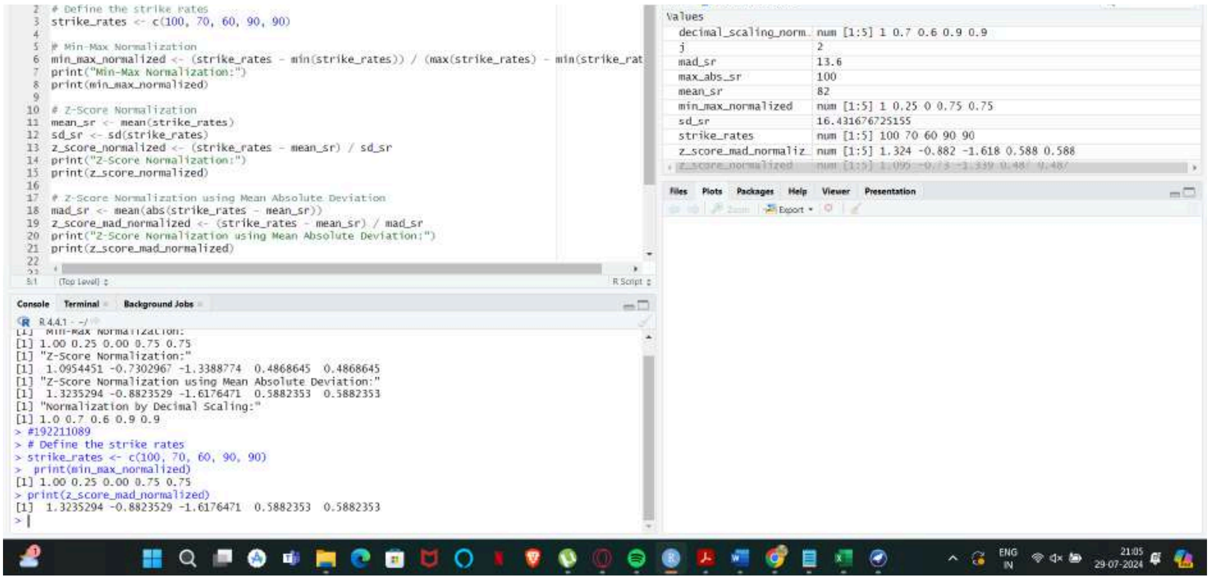

**32) The given are the strike-rates scored by a batsman in season 1 in different tournaments. 100, 70, 60, 90, 90**

**(a) min-max normalization by setting min = 0 and max = 1**

**(b) z-score normalization**

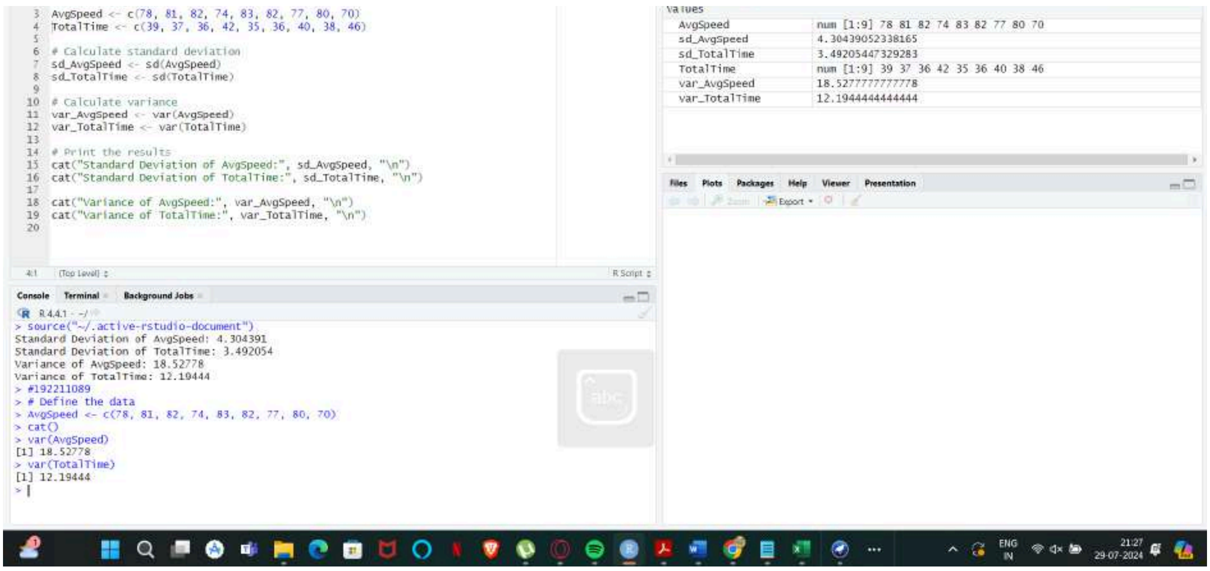
**(c) z-score normalization using the mean absolute deviation instead of standard deviation**

**(d) normalization by decimal scaling**

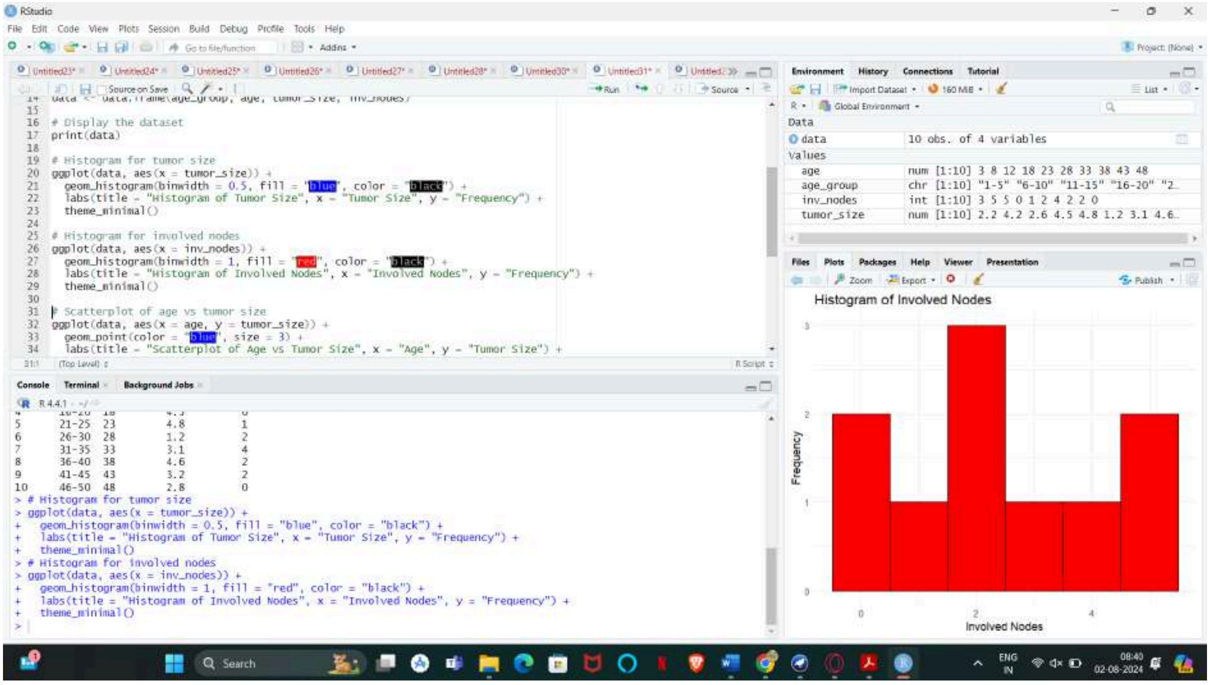



**33) Suppose some car is tested for the AvgSpeed and TotalTime data for 9 randomly selected car with the following result**

**a) Calculate the standard deviation of AvgSpeed and TotalTime.**
**b) Calculate the Variance of AvgSpeed and TotalTime for the above dataset.**



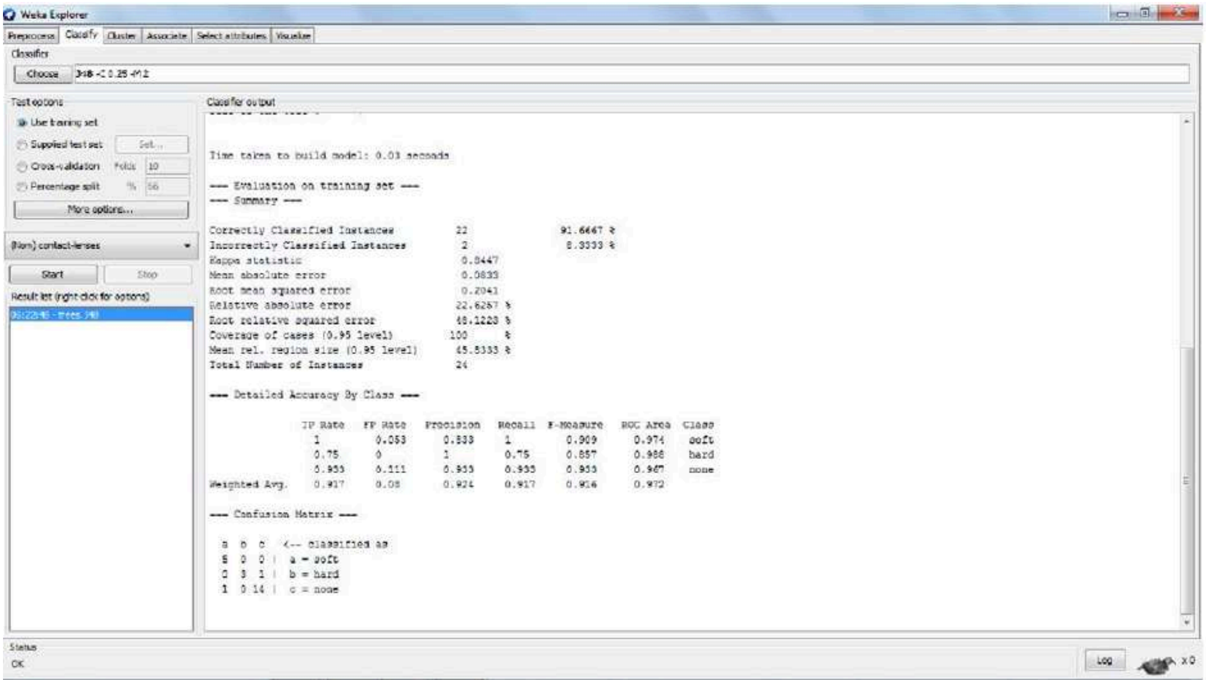

**34) Consider a person want to take a censes / plot for the breast-cancer affected people through the years.Create a own dataset with this parameters age, tumorsize,inv-nodes**
**[example between age 1-5 = no.of.count, 6-10=no.of.count,etc]**
**Draw the Histogram, scatterplot,boxplot.**



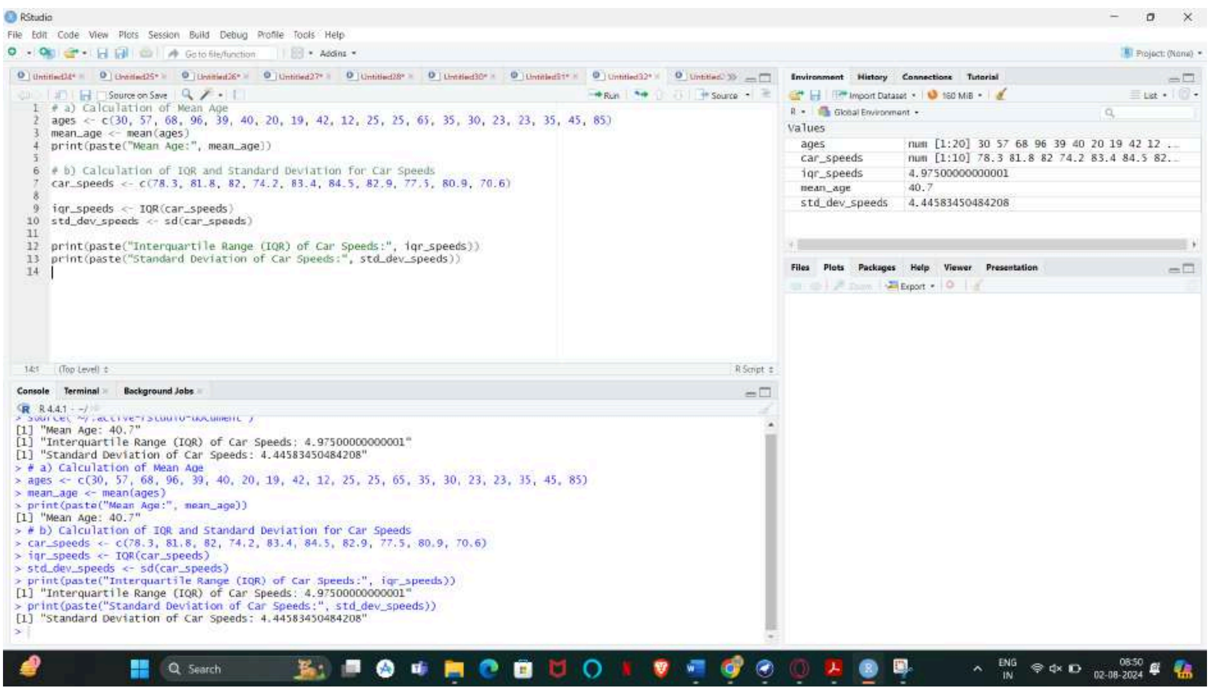

**35)Create the Confusion matrix using this scenario:**
**A shepherd boy gets bored tending the town's flock. To have some fun, he cries out, "Wolf!" even though no wolf is in sight. The villagers run to protect the flock, but then get really mad when they realize the boy was playing a joke on them.One night, the shepherd boy sees a real wolf approaching the flock and calls out, "Wolf!" The villagers refuse to be fooled again and stay in their houses. The hungry wolf turns the flock into lamb chops. The town goes hungry. Panic ensues.**

| | |
|---|---|
| **True Positive (TP):**<br>• Reality: A wolf threatened.<br>• Shepherd said: "Wolf."<br>• Outcome: Shepherd is a hero. | **False Positive (FP):**<br>• Reality: No wolf threatened.<br>• Shepherd said: "Wolf."<br>• Outcome: Villagers are angry at shepherd for waking them up. |
| **False Negative (FN):**<br>• Reality: A wolf threatened.<br>• Shepherd said: "No wolf."<br>• Outcome: The wolf ate all the sheep. | **True Negative (TN):**<br>• Reality: No wolf threatened.<br>• Shepherd said: "No wolf."<br>• Outcome: Everyone is fine. |

36) **Create the ARFF data set for the below mentioned dataset perform the bayes theorem in addition to that compare the same with decision tree. identify the effiecient classifier with accuracy with F1 Score.**




37) **a) Suppose that the "Diabetes data set " data for analysis includes the attribute age. The age values for the data are (in increasing order) 30, 57, 68, 96, 39, 40, 20, 19, 42, 12, 25, 25, 65, 35, 30, 23, 23, 35, 45, 85. What is the mean?**
**b) Suppose that the speed car is mentioned in different driving style.**

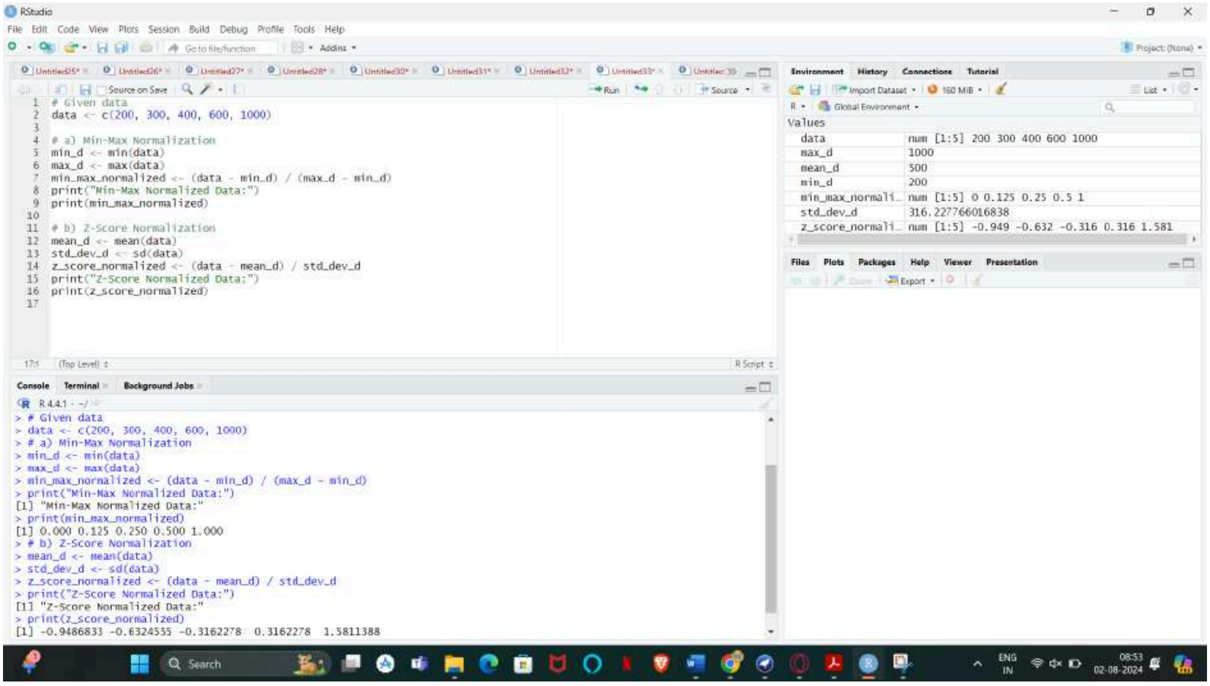


**38) a)Let us consider one example to make the calculation method clear. Assume that the minimum and maximum values for the feature F are $50,000 and $100,000 correspondingly. It needs to range F from 0 to 1. In accordance with min-max normalization, v = $80,**

**b) Use the two methods below to normalize the following group of data: 200, 300, 400, 600, 1000**

**(a) min-max normalization by setting min = 0 and max = 1**

**(b) z-score normalization**




**39)Consider this table**

**TID items bought**

**T100 {M, O, N, K, E, Y}**

**T200 {D, O, N, K, E, Y }**

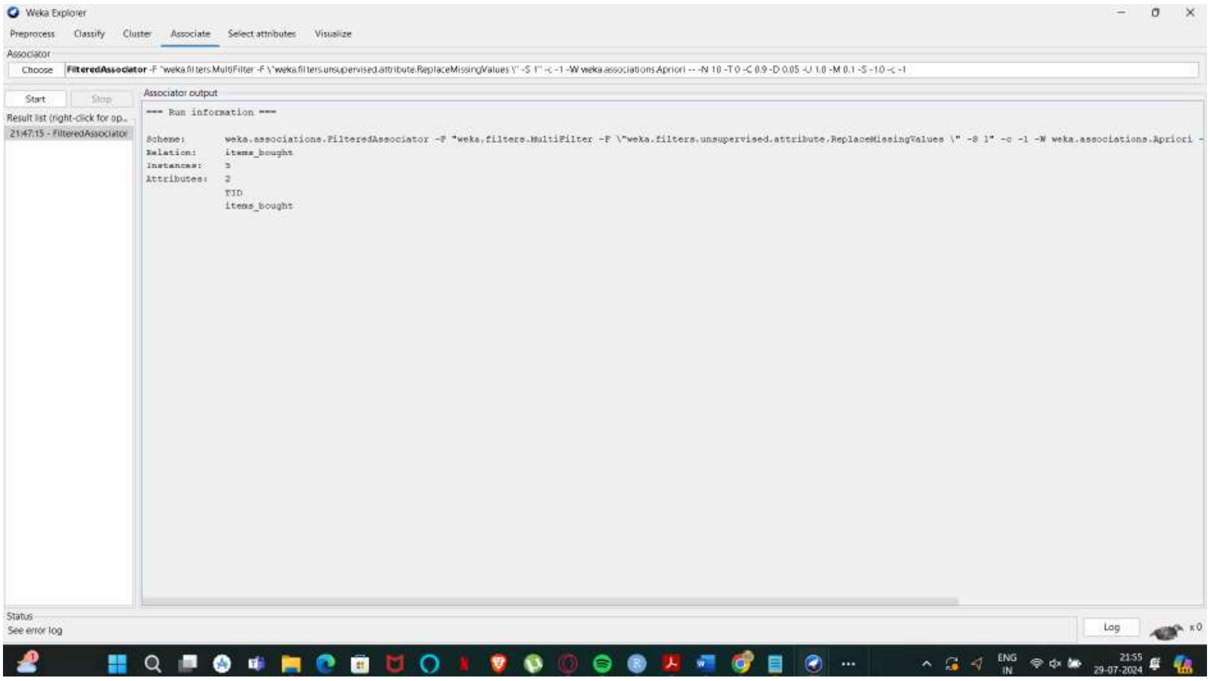
**T300 {M, A, K, E}**

**T400 {M, U, C, K, Y}**

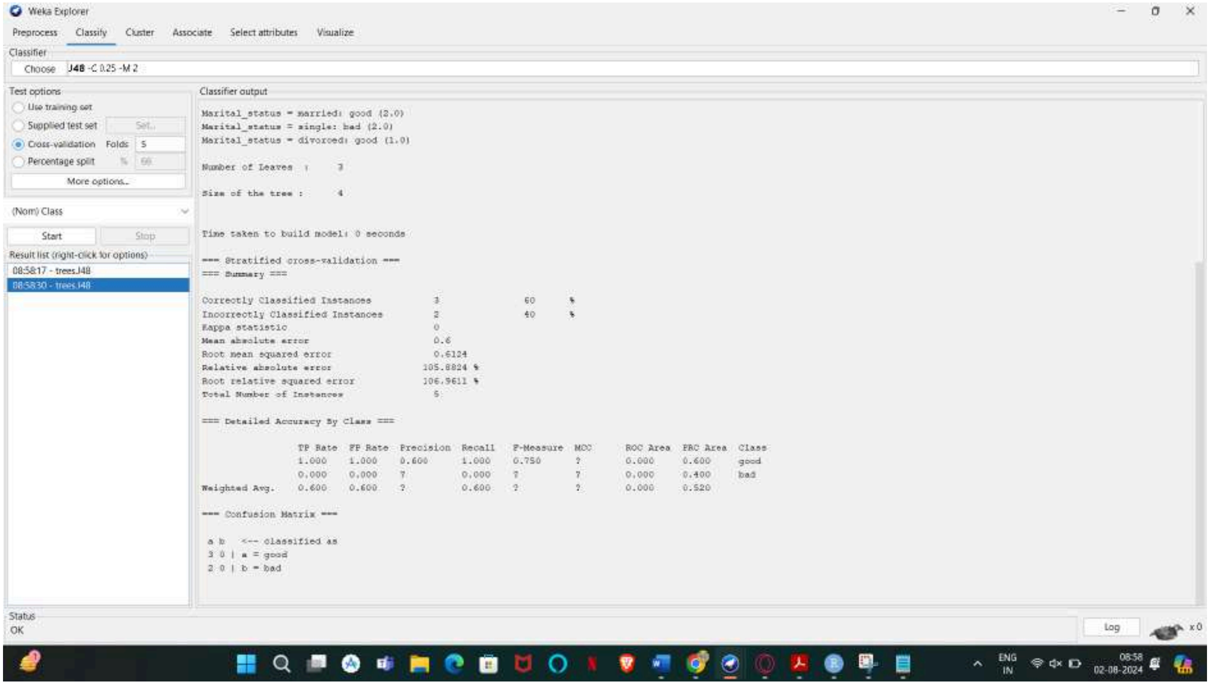
**T500 {C, O, O, K, I ,E}**

**(a) Find all frequent item set using Apriori and FP-growth, respectively. Compare the efficiency of the two mining processes.**

**(b) List all of the strong association rules (with support s and confidence c) matching the following metarule, where X is a variable representing customers, and itemi denotes variables representing items (e.g., "A", "B", etc.):**
**$\forall x \in$ transaction, buys(X, item1) $\wedge$ buys(X, item2) $\Rightarrow$ buys(X, item3)**




**40) Suppose we want to classify potential bank customers as good creditors or bad creditors for loan applications. We have a training dataset describing past customers using the following attributes: Marital status {married, single, divorced}, Gender {male, female}, Age {[18..30[, [30..50[, [50..65[, [65+]}, Income {[10K..25K[, [25K..50K[, [50K..65K[, [65K..100K[, [100K+]}. Using Weka tool solve this problem.**


Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
J48 -C 0.25 -M 2
Test options
Use training set
Supplied test set
Set...
Cross-validation
Folds
5
Percentage split
%
66
More options...
(Nom) Class
Start
Stop
Result list (right-click for options)
08:58:17 - trees.J48
08:58:30 - trees.J48
Classifier output
Marital_status = married: good (2.0)
Marital_status = single: bad (2.0)
Marital_status = divorced: good (1.0)
Number of Leaves : 3
Size of the tree : 4
Time taken to build model: 0 seconds
=== Stratified cross-validation ===
=== Summary ===
Correctly Classified Instances 3 60 %
Incorrectly Classified Instances 2 40 %
Kappa statistic 0
Mean absolute error 0.6
Root mean squared error 0.6124
Relative absolute error 105.8824 %
Root relative squared error 106.9611 %
Total Number of Instances 5
=== Detailed Accuracy By Class ===
TP Rate FP Rate Precision Recall F-Measure MCC ROC Area PRC Area Class
1.000 1.000 0.600 1.000 0.750 ? 0.000 0.600 good
0.000 0.000 ? 0.000 ? ? 0.000 0.400 bad
Weighted Avg. 0.600 0.600 ? 0.600 ? ? 0.000 0.520
=== Confusion Matrix ===
a b <-- classified as
3 0 | a = good
2 0 | b = bad
Status
OK
Log
x 0
ENG
IN
08:58
02-08-2024