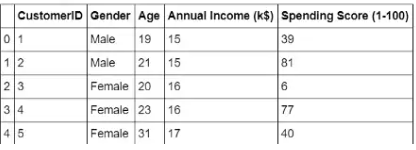
Experiment-1

.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.*Sample data



Program

@relation exp1

@attribute customerid numeric

@attribute gender{male,female}

@attribute age numeric

@attribute annual\_income numeric

@attribute spending\_score numeric

@data

1,male,19,15,39

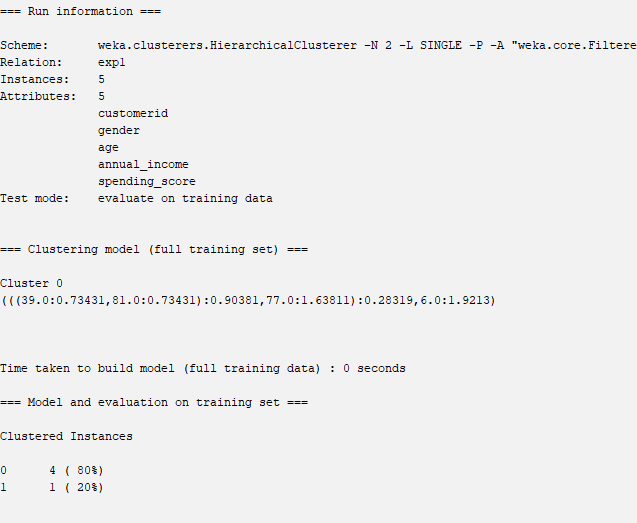
2,male,21,15,81

3,female,20,16,6

4,female,23,16,77

5,female,31,17,40

Output



Experiment-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.

| EmployeID | Gender | Age | Salary | Credit |
| --- | --- | --- | --- | --- |
| 111 | Male | 28 | 150000 | 39 |
| 222 | Male | 25 | 150000 | 27 |
| 333 | Female | 26 | 160000 | 42 |
| 444 | Female | 25 | 160000 | 40 |
| 555 | Female | 30 | 170000 | 64 |
| 666 | Male | 29 | 200000 | 72 |

Program

@relation exp1

@attribute emp\_id numeric

@attribute gender{male,female}

@attribute age numeric

@attribute salary numeric

@attribute credit numeric

@data

111,male,28,1500000,39

222,male,25,1500000,27

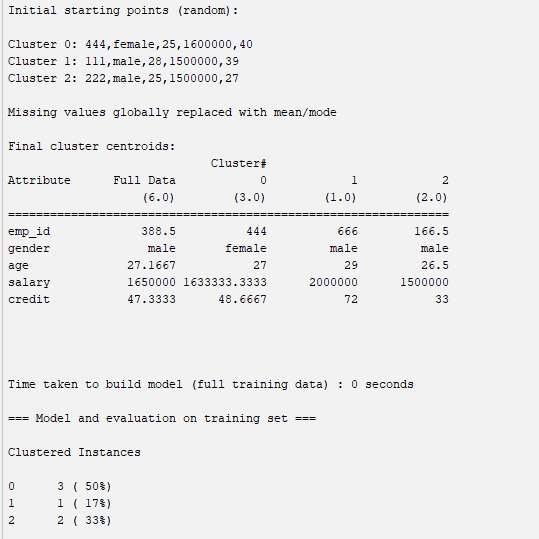
333,female,26,1600000,42

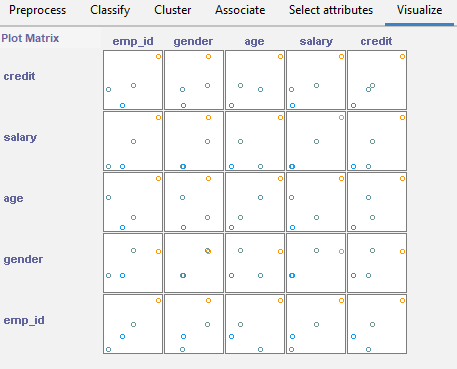
444,female,25,1600000,40

555,female,30,1700000,64

666,male,29,2000000,72

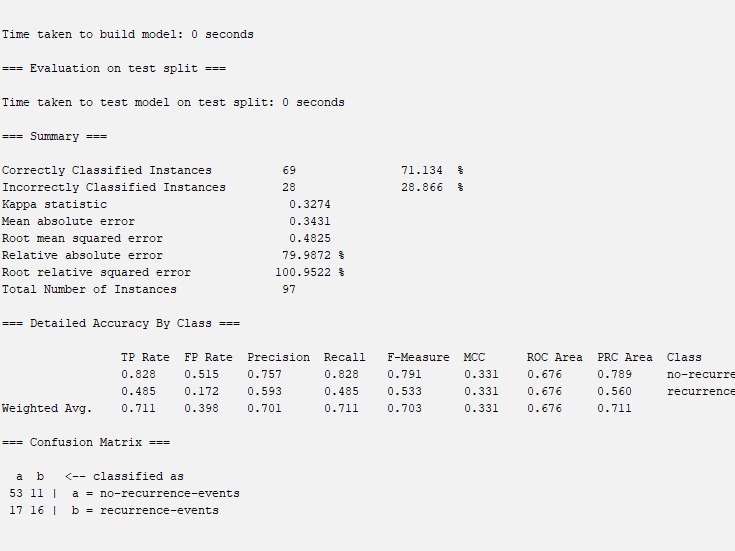
Output





Experiment-3 naivebayes in breast\_cancer

Output



Experiment-4

**T**he 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?

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Person** | Gopu | Babu | Baby | Gopal | Krishna | Jai | Dev | Malini | Hema | Anu |
| **Vegetarian** | yes | yes | yes | no | yes | no | no | yes | yes | yes |

Program

@relation exp1

@attribute person{gopu,babu,baby,gopal,krishna,jai,dev,malini,hema,anu}

@attribute gender{yes,no}

@data

gopu,yes

babu,yes

baby,yes

gopal,no

krishna,yes

jai,no

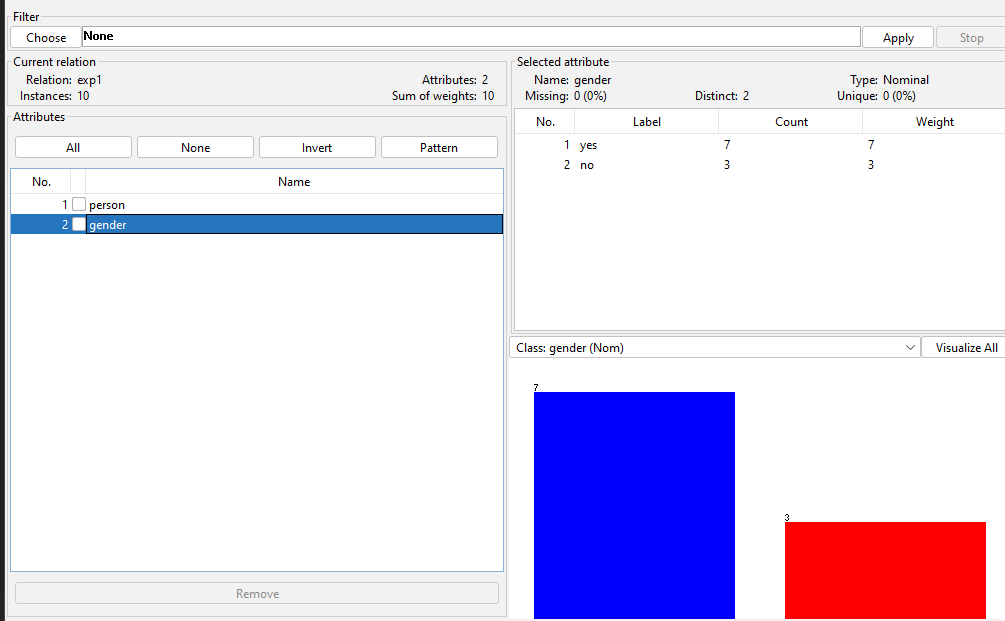
dev,no

malini,yes

hema,yes

anu,yes

output;



Experiment-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%.



Program

@relation exp1

@attribute bread{true,false}

@attribute cheese{true,false}

@attribute egg{true,false}

@attribute juice{true,false}

@attribute milk{true,false}

@attribute yogurt{true,false}

@data

true,true,true,true,false,false

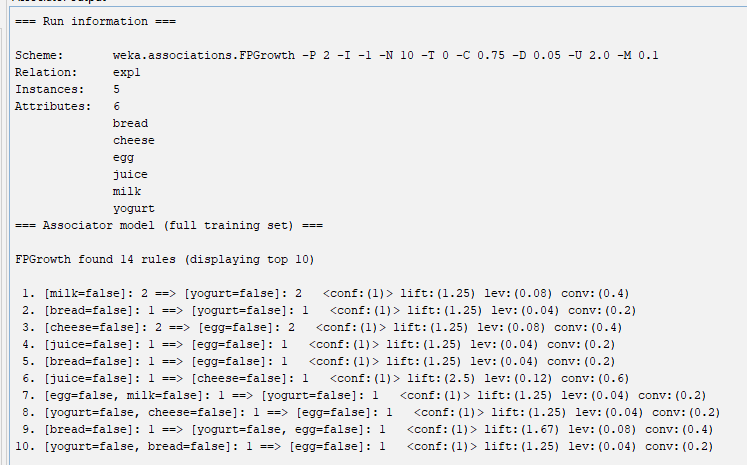
true,true,false,true,false,false

true,false,false,false,true,true

true,false,false,true,true,false

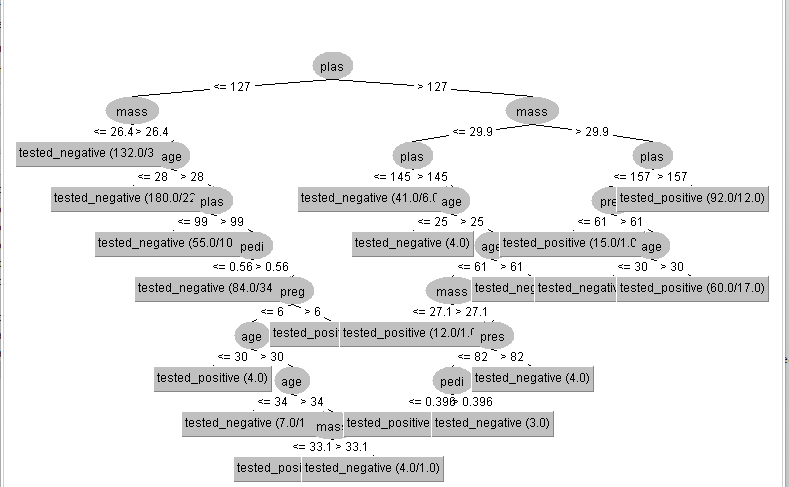
false,true,false,true,true,false

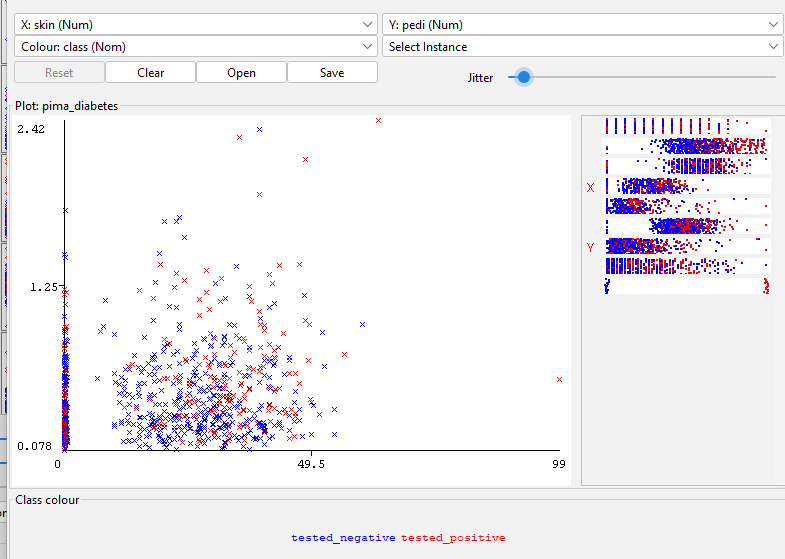
output;



Experiment-7

**.P**rediction 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.





Experiment-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

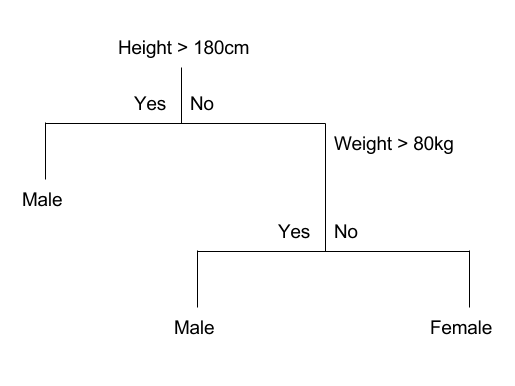
(c) clustering

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.



library(dplyr)

library(tibble)

library(ggplot2)

data <- c(55, 60, 71, 63, 55, 65, 50, 55,58,59,61,63,65,67,71,72,75)

equi\_depth <- quantile(data, probs = c(0, 1/3, 2/3, 1))

equi\_depth\_partitioned <- cut(data, breaks = equi\_depth, labels = c("Low", "Medium", "High"), include.lowest = TRUE)

min\_value <- min(data)

max\_value <- max(data)

width <- (max\_value - min\_value)/3

equal\_width <- seq(min\_value, max\_value, by = width)

equal\_width\_partitioned <- cut(data, breaks = equal\_width, labels = c("Low", "Medium", "High"), include.lowest = TRUE)

kmeans\_model <- kmeans(data, centers = 3)

cluster\_assignments <- as.factor(kmeans\_model$cluster)

levels(cluster\_assignments) <- c("Low", "Medium", "High")

data\_tibble <- tibble(data = data, equi\_depth\_partitioned = equi\_depth\_partitioned, equal\_width\_partitioned = equal\_width\_partitioned, cluster\_assignments = cluster\_assignments)

ggplot(data\_tibble, aes(x = data)) +

geom\_histogram(binwidth = 5) +

facet\_wrap(~equi\_depth\_partitioned, ncol = 1, scales = "free\_x") +

ggtitle("Histogram using Equi-Depth Partitioning")

ggplot(data\_tibble, aes(x = data)) +

geom\_histogram(binwidth = 5) +

facet\_wrap(~equal\_width\_partitioned, ncol = 1, scales = "free\_x") +

ggtitle("Histogram using Equal-Width Partitioning")

ggplot(data\_tibble, aes(x = data)) +

geom\_histogram(binwidth = 5) +

facet\_wrap(~cluster\_assignments, ncol = 1, scales = "free\_x") +

experiment-9

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%

|  |  |
| --- | --- |
| T.ID | ITEMS |
| T1 | SONY, BPL, LG |
| T2 | BPL, SAMSUNG |
| T3 | BPL, ONIDA |
| T4 | SONY, BPL, SAMSUNG |
| T5 | SONY, ONIDA |
| T6 | BPL, ONIDA |
| T7 | SONY, ONIDA |
| T8 | SONY, BPL, ONIDA, LG |
| T9 | SONY, BPL, ONIDA |

program

@relation exp1

@attribute SONY{true,false}

@attribute BPL{true,false}

@attribute LG{true,false}

@attribute SAMSUNG{true,false}

@attribute ONIDA{true,false}

@data

true,true,true,false,false

false,true,false,true,false

false,true,false,false,true

true,true,false,true,false

true,false,false,false,true

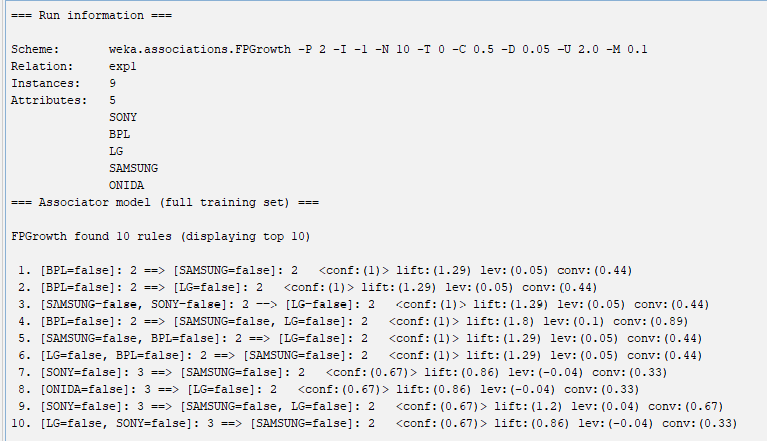
false,true,false,false,true

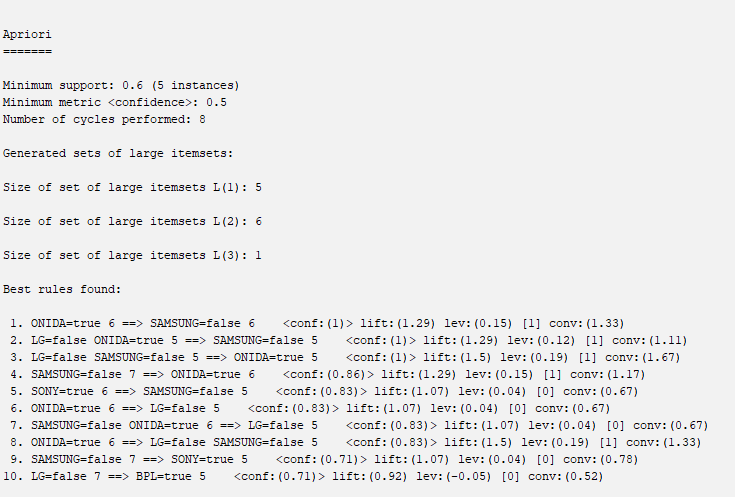
true,false,false,false,true

true,true,true,false,true

true,true,false,false,true

output;





Experiment-10 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** |

1. Calculate the standard deviation of AvgSpeed and TotalTime.
2. Calculate the Variance of AvgSpeed and TotalTime for the above dataset.

speed<-c(78,81,82,74,83,82,77,80,70)

time<-c(39,37,36,42,35,36,40,38,46)

a<-var(speed)

b<-var(time)

ss<-sqrt(var(speed))

st<-sqrt(var(time))

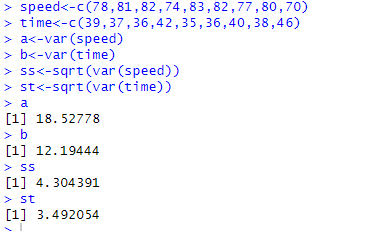
a

b

ss

st

output



Experiment-11

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

1. min-max normalization by setting min = 0 and max = 1
2. z-score normalization
3. z-score normalization using the mean absolute deviation instead of standard deviation
4. normalization by decimal scaling

a<-c(100,70,60,90,90)

min=0

max=1

v<-((a-min(a)/max(a)-min(a))\*max+min)/100

summary(v,method=c("range"))

v

#zscore

w=mean(a)

s=sd(a)

z=(a-w)/s

summary(z)

z

output

