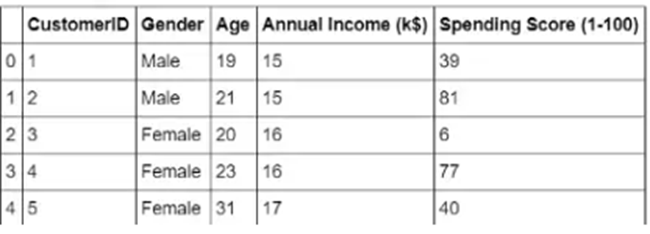
Day 4

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



**Input:**

@relation dataset

@attribute customerid{1,2,3,4,5}

@attribute gender{male,female}

@attribute age{19,21,20,23,31}

@attribute income{15,16,17}

@attribute score{usual,priority,senior,young}

@data

1 male 19 15 usual

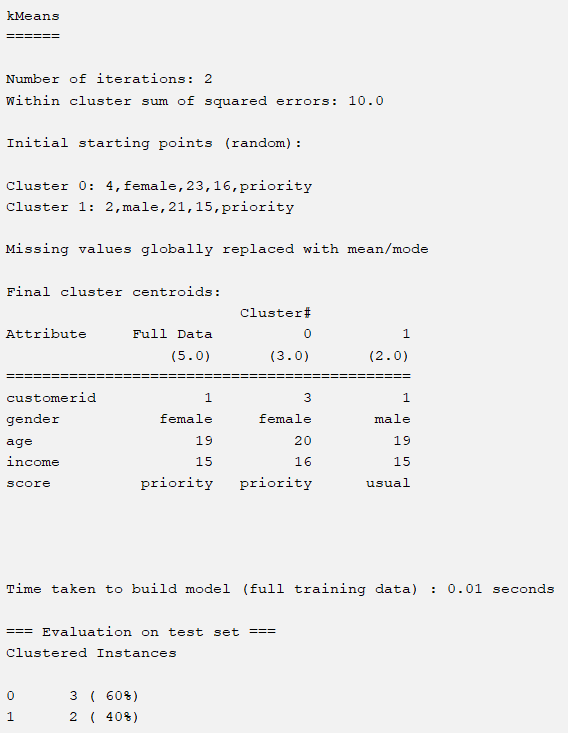
2 male 21 15 priority

3 female 20 16 young

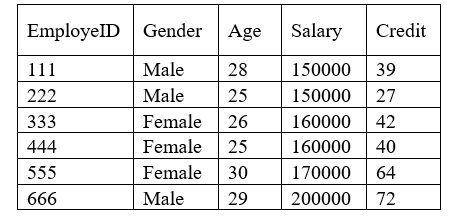
4 female 23 16 priority

5 female 31 17 senior

**Output:**

****

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 the cluster.



**Input:**

@relation dataset

@attribute id{1,2,3,4,5,6}

@attribute gender{male,female}

@attribute age{28,25,26,30,29}

@attribute salary{15,16,17,20}

@attribute credit{39,27,42,40,64,72}

@data

1 male 28 15 39

2 male 25 15 27

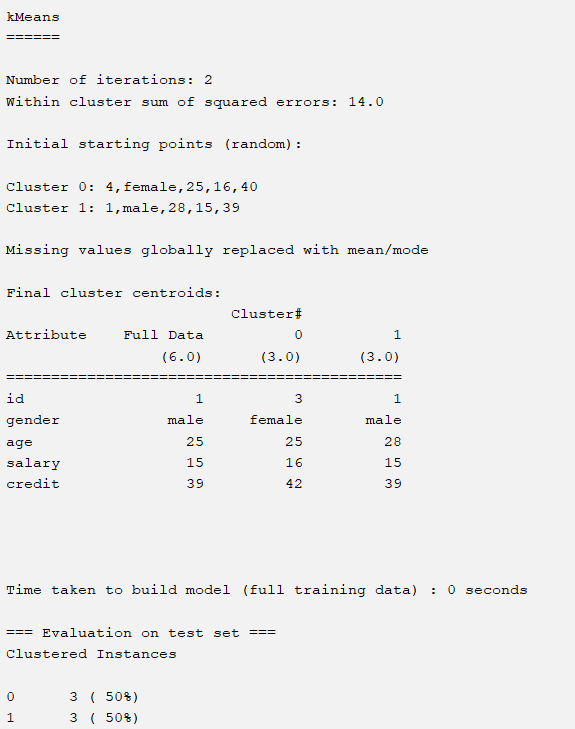
3 female 26 16 42

4 female 25 16 40

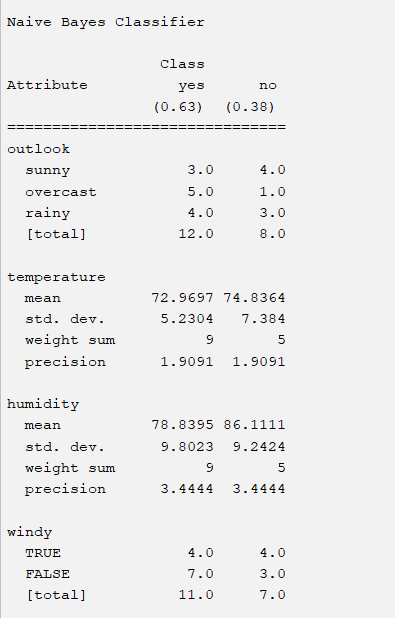
5 female 30 17 64

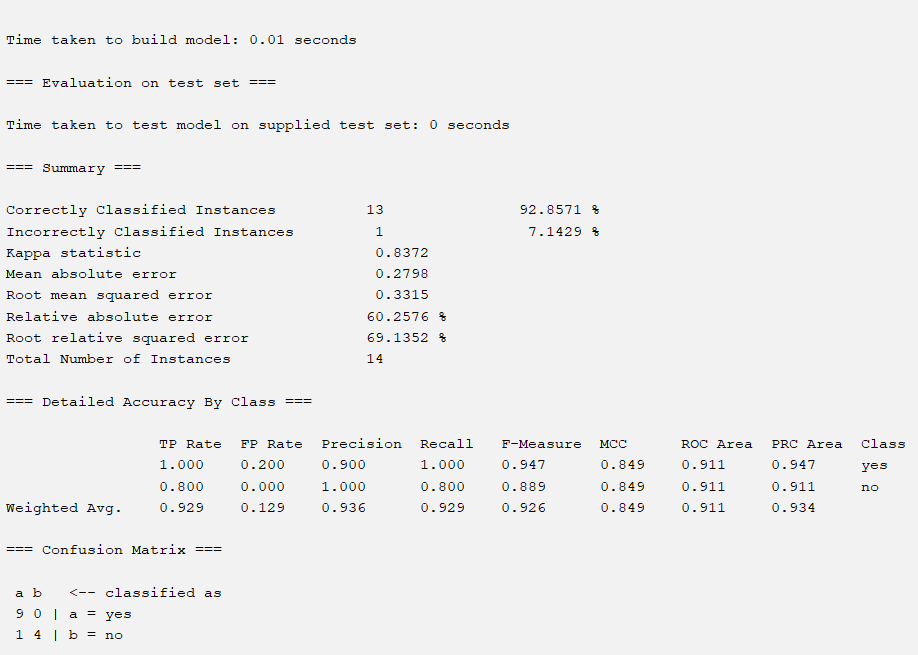
6 male 29 20 72

**Output:**

****

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.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?

**Input:**

@relation dataset

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

@attribute vegeterian{yes,no}

@data

gopu yes

babu yes

baby yes

gopal no

krishna yes

jai no

dev no

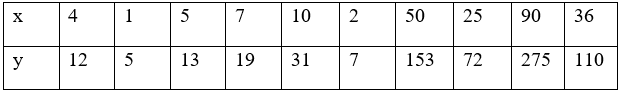
malini yes

hema yes

anu yes

**Output:**

 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.



**Input:**

@relation dataset

@attribute x{4,1,5,7,10,2,50,25,90,36}

@attribute y{12,5,13,19,31,7,153,72,275,110}

@data

4,12

1,5

5,13

7,19

10,31

2,7

50,153

25,72

90,275

36,110

R programming:

x\_values <- c(4, 1, 5, 7, 10, 2, 50, 25, 90, 36)

y\_values <- c(12, 5, 13, 19, 31, 7, 153, 72, 275, 110)plot(x\_values, y\_values, main="Scatter Plot of Mobile Phones Sold",

xlab="Number of Mobile Phones Sold", ylab="Money",

pch=16, col="blue")

grid()

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%.



**Input:**

@relation dataset

@attribute transid{1,2,3,4,5}

@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

1,true,true,true,true,false,false

2,true,true,false,true,false,false

3,true,false,false,false,true,true

4,true,false,false,true,true,false

5,false,true,false,true,true,false

**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.

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

**Input:**

marks <- c(55, 60, 71, 63, 55, 65, 50, 55, 58, 59, 61, 63, 65,

bins\_a <- cut(marks, breaks = 3, labels = c("Low", "Medium", "High"))

bins\_b <- cut(marks, breaks = seq(min(marks), max(marks), length.out = 4), labels = c("Low", "Medium", "High"))

k <- 3

clusters <- kmeans(matrix(marks), centers = k)

bins\_c <- cut(clusters$centers[clusters$cluster], breaks = 3, labels = c("Low", "Medium", "High"))

par(mfrow = c(1, 3))

hist(marks, main = "Equal-frequency (equi-depth) partitioning", col = "skyblue", breaks = 3)

hist(marks, main = "Equal-width partitioning", col = "lightgreen", breaks = seq(min(marks), max(marks), length.out = 4))

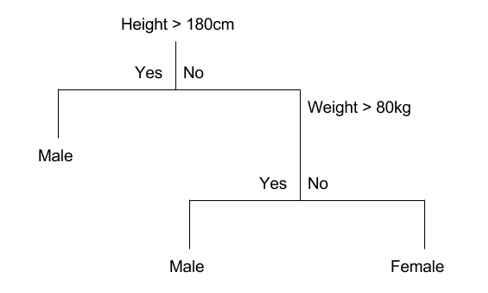
hist(marks, main = "Clustering", col = "lightpink", breaks = 3)

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%

****

**Input:**

@relation dataset

@attribute id{1,2,3,4,5,6,7,8,9}

@attribute sony{true,false}

@attribute bpl{true,false}

@attribute lg{true,false}

@attribute samsung{true,false}

@attribute onida{true,false}

@data

1,true,true,true,false,false

2,false,true,false,true,false

3,false,true,false,false,true

4,true,true,false,true,false

5,true,false,false,false,true

6,false,true,false,false,true

7,true,false,false,false,true

8,true,true,true,false,true

9,true,true,false,false,true

**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

**Input:**

strike\_rates <- c(100, 70, 60, 90, 90)

min\_max\_normalization <- function(x) {

(x - min(x)) / (max(x) - min(x))

}

normalized\_min\_max <- min\_max\_normalization(strike\_rates)

z\_score\_normalization <- function(x) {

(x - mean(x)) / sd(x)

}

normalized\_z\_score <- z\_score\_normalization(strike\_rates)

mad\_normalization <- function(x) {

(x - mean(x)) / mad(x)

}

normalized\_mad <- mad\_normalization(strike\_rates)

decimal\_scaling\_normalization <- function(x) {

x / 10^(ceiling(log10(max(x))))

}

normalized\_decimal\_scaling <- decimal\_scaling\_normalization(strike\_rates)

cat("Original Data:", strike\_rates, "\n\n")

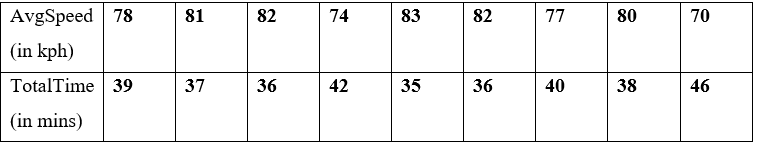
cat("(a) Min-Max Normalization:", normalized\_min\_max, "\n")

cat("(b) Z-Score Normalization:", normalized\_z\_score, "\n")

cat("(c) Z-Score Normalization (MAD):", normalized\_mad, "\n")

cat("(d) Normalization by Decimal Scaling:", normalized\_decimal\_scaling, "\n")

 12.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.

**Input:**

avg\_speed <- c(78, 81, 82, 74, 83, 82, 77, 80, 70)

total\_time <- c(39, 37, 36, 42, 35, 36, 40, 38, 46)

sd\_avg\_speed <- sd(avg\_speed)

sd\_total\_time <- sd(total\_time)

var\_avg\_speed <- var(avg\_speed)

var\_total\_time <- var(total\_time)

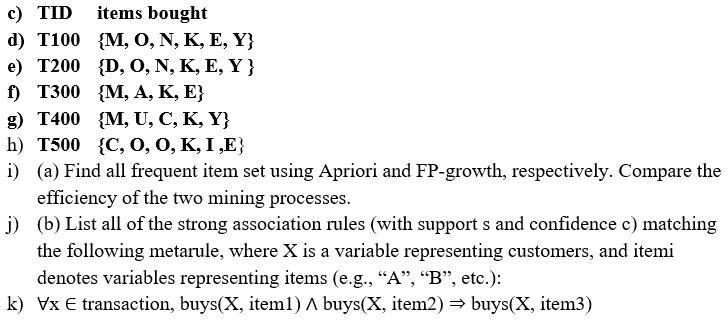
cat("Standard Deviation of AvgSpeed:", sd\_avg\_speed, "\n")

cat("Standard Deviation of TotalTime:", sd\_total\_time, "\n\n")

cat("Variance of AvgSpeed:", var\_avg\_speed, "\n")

cat("Variance of TotalTime:", var\_total\_time, "\n")

13. Consider the table



Input:

@relation dataset

@attribute M{true,false}

@attribute O{true,false}

@attribute N{true,false}

@attribute K{true,false}

@attribute E{true,false}

@attribute Y{true,false}

@attribute D{true,false}

@attribute A{true,false}

@attribute U{true,false}

@attribute C{true,false}

@attribute I{true,false}

@data

true true true true true true false false false false false false

false true true true true true true false false false false false

true false false true true false false false true false false false

true false false true false true false false false true true false

false true false true true false false false false false true true