PREDICTING HEART DISEASE FOR PEOPLE CONSUMING ALCOHOL USING DATAMINIG TECHNIQUES

PROJECT REPORT

Submitted in fulfilment for the J Component of ITA5007
Data Mining and Business Intelligence

CAL COURSE
In
M.C.A
By

K.BHARATH (18MCA0043)
J.SUNIL KUMAR (18MCA0065)
R.MAYUR (18MCA0019)

Under the guidance of
Prof. SathiyaMoorthy E
SITE

School of Information Technology and Engineering

Winter Semester 2018-19

TABLE OF CONTENTS

Review Report

1. Title:-

PREDICTING HEART DISEASE FOR PEOPLE CONSUMING ALCOHOL USING DATAMINIG TECHNIQUES

2. Abstract:-

- No of instances are “541 instances”.
- This dataset has totally a collection of 12 attributes & a class label
- There are two class labels Alcholic and Non Alcholic.
- Several constraints were considered for the selection of these instances
    - from a larger database.
- For example, tuples having empty cells are deleted.

3. Objective:-

Main objective is to provide some solution for the real-world problems using classification & clustering algorithms like naïve-bayes & k-means algorithms.

4. Data Set Description:-

- No of columns : 12
- No of rows:541
- No. of Attributes: 09
- “blood pressure”numeric
    - ii.“CTC kg” numeric
    - iii.”Cholestrol level” numeric
    - iv.”Behavoiur” numeric
    - v.”slope” numeric
    - vi.”vessels” numeric
    - vii.”thal” numeric

4.1 Input Value:

Description about each attribute:

i.”blood pressure” numeric:

Max Value:169
Min Value:108

ii.”CTC”numeric:

Max Value:564
Min Value:162

iii.”heart rate “numeric:

Max Value: 200
Min Value: 100

iv.”cholestrol”numeric:

Max Value: 564
Min Value: 102

v.”old peak” numeric:

MaxValue: 2.8
Min Value: 0

vi.“Slope”numeric:

Asymmetric coefficient of the kernel

Min Value: 0
Max Value: 2

4.2 Target Value:

Class Label:

{Canadian,Kama,Rosa}

Alcohol and non-alcohol are the types of behavoiur

Number of instances :

Alcholic(230) and Non-Alcoholic(230)

5. Literature survey:

Daniel Lowd,Pedro Domingos,Naive Bayes Models for Probability Estimation[1],they proposed that Naive Bayes are seldom used for general probabilistic learning and inference (i.e., for estimating and computing arbitrary joint, conditional and marginal distributions .But, for a wide range of benchmark datasets, naive Bayes models learned using EM have accuracy and learning time comparable to Bayesian networks with context-specific independence.

Jiangtao Ren , Sau Dan Lee , Xianlu Chen , Ben Kao , Reynold Cheng and David Cheung,Naive Bayes Classification of Uncertain Data[2],they proposed the key solution is to extend the class conditional probability estimation in the Bayes model to handle pdf's. Extensive experiments on UCI datasets show that the accuracy of naive Bayes model can be improved by taking into account the uncertainty information.

Ashok AravindanS. M. Anzar,Compression-Based Averaging of Selective Naive Bayes Classifier[3],they explained The limits of Bayesian model averaging in the case of the naive Bayes assumption and introduce a new weighting scheme based on the ability of the models to conditionally compress the class labels. The weighting scheme on the models reduces to a weighting scheme on the variables, and finally results in a naive Bayes classifier with "soft variable selection".

Andrew McCallum,Kamal Nigam,A Comparison of Event Models for Naive Bayes Text Classification[4],It aims to clarify the confusion by describing the differences and details of these multi-variate Bernoulli model and multinomial model, and by empirically comparing their corpora.They found that the multi-variate Bernoulli performs well with small vocabulary sizes, but that the multinomial performs usually performs even better at larger vocabulary sizes providing on average a 27% reduction in error over the multi-variate Bernoulli model at any vocabulary size.

Mahesh Kini M, Saroja Devi H, Prashant G Desai, Niranjan Chiplunkar,Text Mining Approach to Classify Technical Research Documents using Naïve Bayes[5],they proposed implementations of Naïve Bayesian (NB) approach for the automatic classification of Documents restricted to Technical

Research documents based on their text contents and its results analysis. We also discuss a comparative analysis of Weighted Bayesian classifier approach with the Naive Bayes classifier.

Maria-Florina Balcan,Yingyu Liang,Pramod Gupta,Robust Hierarchical Clustering[6],they analyzed a new robust algorithm for bottom-up agglomerative clustering. Algorithm can be used to cluster accurately in cases where the data satisfies a number of natural properties and where the traditional agglomerative algorithm fails.

K.Sasirekha, P.Baby,Agglomerative Hierarchical Clustering Algorithm- A Review[7],They proposed that Agglomerative algorithm is a "bottom up" approach: each observation starts in its own cluster, and pairs of clusters are merged as one moves up the hierarchy.Divisive: This is a "top down" approach: all observations start in one cluster, and splits are performed recursively as one moves down the hierarchy.

Ying Zhao and George Karypis,Evaluation of Hierarchical Clustering Algorithms for Document Datasets[8],They evaluate different hierarchical clustering algorithms and toward this goal compare various partitional and agglomerative approaches.

K.Ranjini ,N.Rajalingam ,Performance Analysis of Hierarchical Clustering Algorithm[9],theyproposed theimplementation of agglomerative and divisive clustering algorithms applied on various types of data. Visual programming is used for implementation and running time of the algorithms using different linkagesto different types of data are taken for analysis.

Akshay Krishnamurthy,Sivaraman Balakrishnan,Min Xu,Aarti Singh,Efficient Active Algorithms for Hierarchical Clustering[10],they proposed a general framework for active hierarchical clustering that repeatedly runs an off-the-shelf clustering algorithm on small subsets of the data and comes with guarantees on performance, measurement complexity and runtime complexity.

6. A table with dataset description:-

Data Set Characteristics:	Multivariate	Number of Instances:	210	Area:	Life
Attribute Characteristics:	Real	Number of Attributes:	7	Date Donated	2012-09-29
Associated Tasks:	Classification, Clustering	Missing Values?	N/A	Number of Web Hits:	146049

7. Sample database of dataset:-

age	gender	chest_pai	blood_pre	cholestora	blood_sug	ecg	heart_rate	exercise	oldpeak	slope	vessels	thal	alcholic
70	1	4	130	322	0	2	109	0	2.4	2	3	3	alcholic
67	0	3	115	564	0	2	160	0	1.6	2	0	7	no alcholic
57	1	2	124	261	0	0	141	0	0.3	1	0	7	alcholic
64	1	4	128	263	0	0	105	1	0.2	2	1	7	no alcholic
74	0	2	120	269	0	2	121	1	0.2	1	1	3	alcholic
65	1	4	120	177	0	0	140	0	0.4	1	0	7	alcholic
56	1	3	130	256	1	2	142	1	0.6	2	1	6	no alcholic
59	1	4	110	239	0	2	142	1	1.2	2	1	7	alcholic
60	1	4	140	293	0	2	170	0	1.2	2	2	7	no alcholic
63	0	4	150	407	0	2	154	0	4	2	3	7	alcholic
59	1	4	135	234	0	0	161	0	0.5	2	0	7	alcholic
53	1	4	142	226	0	2	111	1	0	1	0	7	no alcholic
44	1	3	140	235	0	2	180	0	0	1	0	3	no alcholic
61	1	1	134	234	0	0	145	0	2.6	2	2	3	alcholic
57	0	4	128	303	0	2	159	0	0	1	1	3	no alcholic
71	0	4	112	149	0	0	125	0	1.6	2	0	3	no alcholic
46	1	4	140	311	0	0	120	1	1.8	2	2	7	alcholic
53	1	4	140	203	1	2	155	1	3.1	3	0	7	alcholic
64	1	1	110	211	0	2	144	1	1.8	2	0	3	no alcholic
40	1	1	140	199	0	0	178	1	1.4	1	0	7	alcholic
67	1	4	120	229	0	2	129	1	2.6	2	2	7	no alcholic
48	1	2	130	245	0	2	180	0	0.2	2	0	3	no alcholic

8. <u>Testing & Training:</u>

Instances information :

Total Number of Instances: 541

- 324(60% of 541 Instances) & 216 (40% of 541 instances)

8.1 Training Data set:

324 Training Instances (60% of 200 Instances) are present in Training Data set.

8.2 Test Data set:

216 Test Instances (40% of 200 instances) are present in Test Data set.

9. Innovation:

Naive Bayes algorithm here proposed can be used for Real time Prediction,Multi class Prediction,Text classification/ Spam Filtering/ Sentiment Analysis,Text classification/ Spam Filtering/ Sentiment Analysis.

10. Tool used for execution of algorithms:

R:The R project for statistical computing

11. Algorithms Description & Methodology:

Naive Bayes Algorithm:

What is Naïve Bayes?

- Statistical method for classification.
- Supervised learning method.
- Assumes an underlying probabilistic model, the Bayes theorem.
- Can solve problems involving both categorical and continuous valued attributes.
- Named after Thomas bayes, who proposed the bayes theorem.

It uses Bayesian Theorem

$P(H|X) = p(X|H)\ P(H) / p(X)$

Hierarchial Clustering:

Hierarchical clustering algorithm is of two types:

i)Agglomerative Hierarchical clustering algorithm or AGNES (agglomerative nesting) and

ii)Divisive Hierarchical clustering algorithm or DIANA (divisive analysis).

Both this algorithm are exactly reverse of each other.

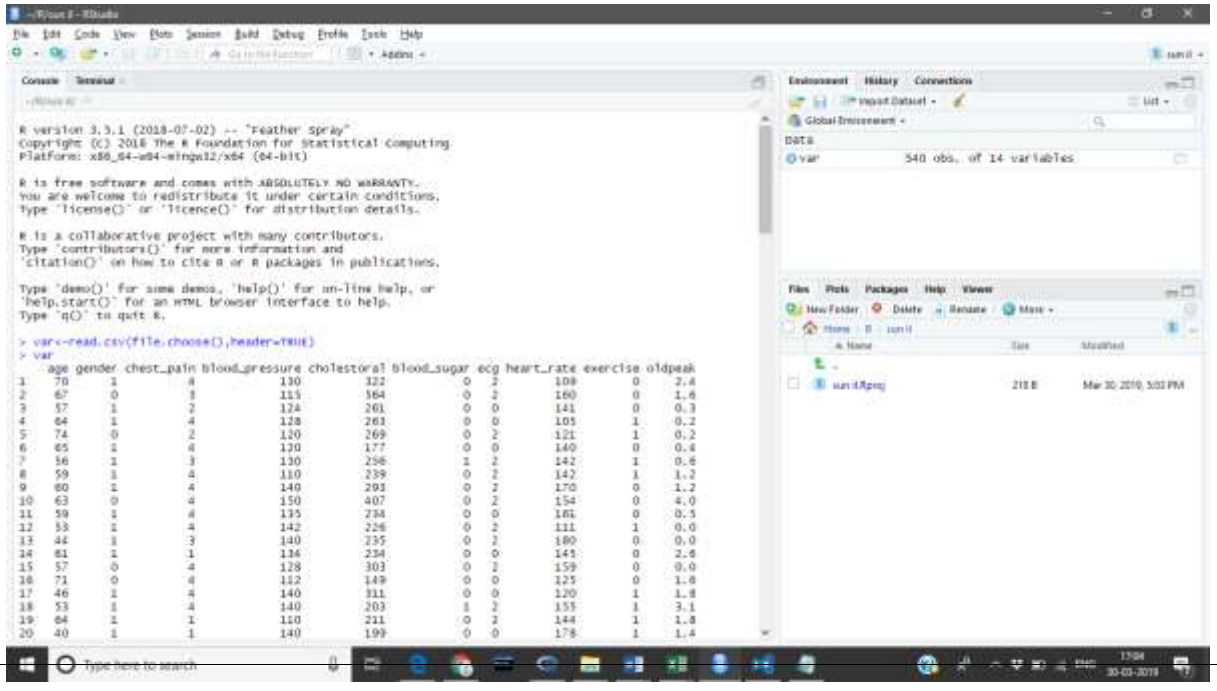
Agglomerative Hierarchical clustering -This algorithm works by grouping the data one by one on the basis of the nearest distance measure of all the pair-wise distance between the data point. Again distance between the data point is recalculated.

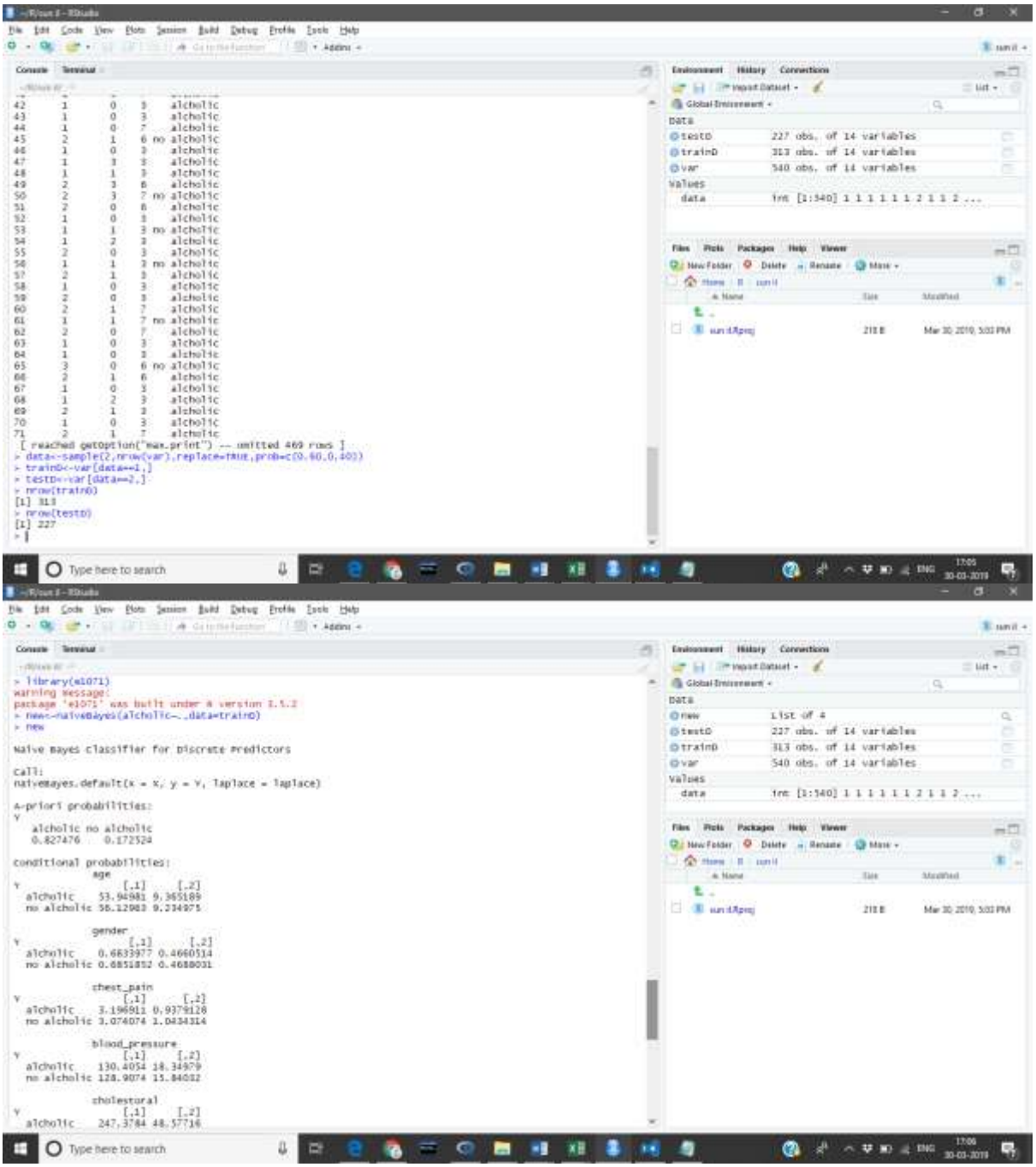
12. Code – Results & discussion:

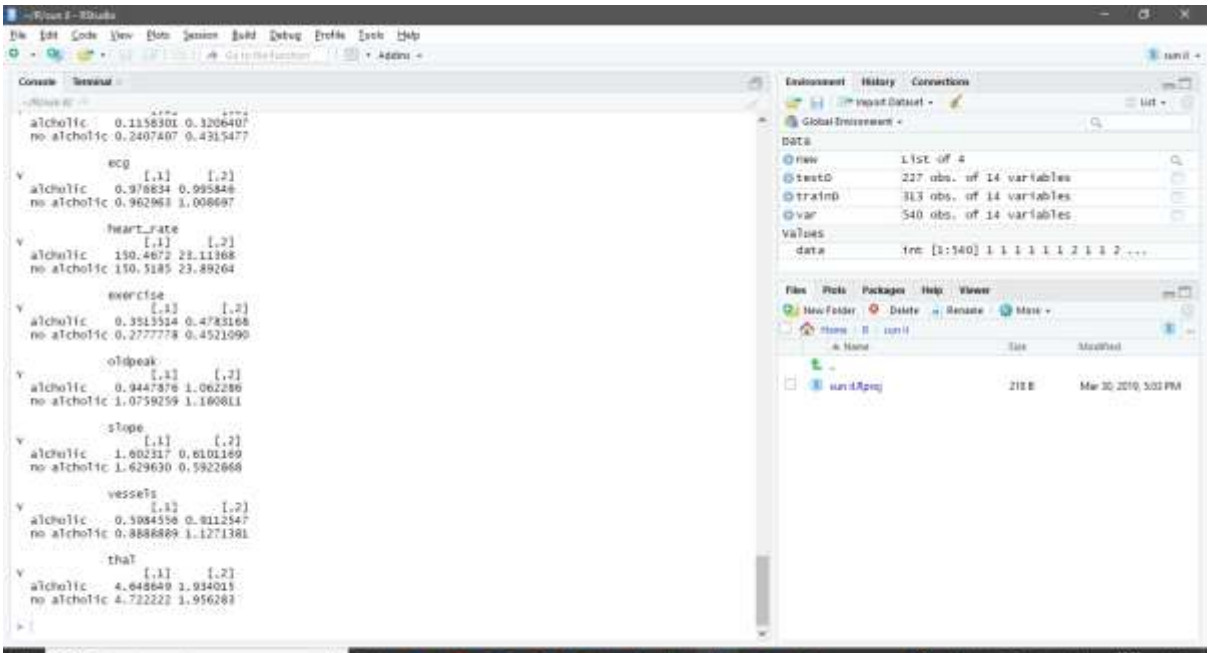
Naive Bayes Algorithm for analysis of geometrical properties of kernels belonging to seeds.

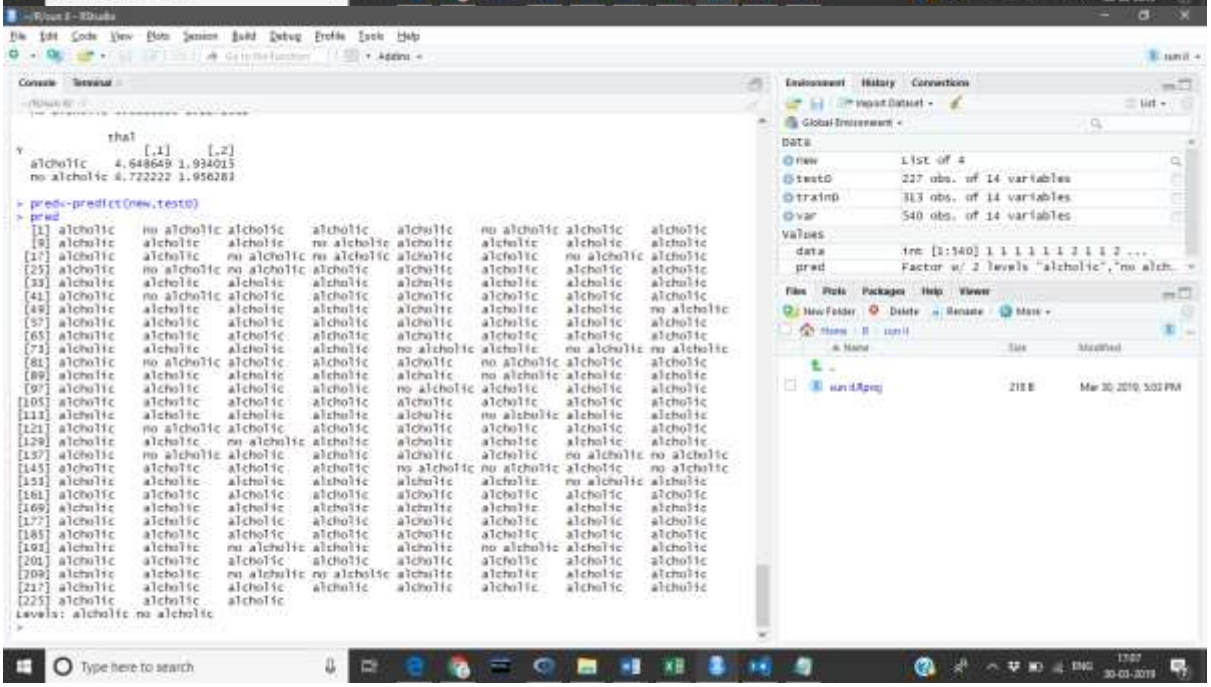
```
var<-read.csv(file.choose(),header=TRUE)
var
data<-sample(2,nrow(var),replace=TRUE,prob=c(0.60,0.40))
trainD<-var[data==1,]
testD<-var[data==2,]
nrow(trainD)
nrow(testD)
library(e1071)
new<-naiveBayes(alcholic~.,data=trainD)
new
pred<-predict(new,testD)
pred
library(rminer)
mmetric(testD$alcholic,pred,c("ACC","PRECISION","TPR","F1"))
library(rpart)
mod=rpart(alcholic~.,data=trainD)
pred=predict(mod,type="class")
table(pred)
table(pred,trainD$alcholic)
clusters1<-hclust(dist(var[,3:4]))
plot(clusters1)
clustcut1<-cutree(clusters1,3)
clustcut1
table(clustcut1,var$alcholic)
library(ggplot2)
ggplot(var,aes(blood_pressure,heart_rate,color=var$alcholic))+geom_point()
savehistory("~/R/sun il/esd.Rhistory")
```

SCREENSHOTS:




42 1 0 3 alcholic
43 1 0 3 alcholic
44 1 0 7 alcholic
45 2 1 6 no alcholic
46 1 0 3 alcholic
47 1 1 3 alcholic
48 1 1 3 alcholic
49 2 3 6 alcholic
50 2 3 7 no alcholic
51 2 0 6 alcholic
52 1 0 3 alcholic
53 1 1 3 no alcholic
54 1 2 3 alcholic
55 2 0 3 alcholic
56 1 1 3 no alcholic
57 2 1 3 alcholic
58 1 0 3 alcholic
59 2 0 3 alcholic
60 2 1 7 alcholic
61 1 1 7 no alcholic
62 2 0 7 alcholic
63 1 0 3 alcholic
64 1 0 3 alcholic
65 3 0 6 no alcholic
66 2 1 6 alcholic
67 1 0 3 alcholic
68 1 2 3 alcholic
69 2 1 3 alcholic
70 1 0 3 alcholic
71 2 1 7 alcholic
[reached getOption("max.print") -- omitted 469 rows]
> data=sample(2,nrow(var),replace=TRUE,prob=c(0.60,0.40))
> trainD<-var[data==1,]
> testD<-var[data==2,]
> nrow(trainD)
[1] 313
> nrow(testD)
[1] 227
Type here to search
Data
testD 227 obs. of 14 variables
trainD 313 obs. of 14 variables
var 540 obs. of 14 variables
Values
data int [1:540] 1 1 1 1 1 1 2 1 1 2 ...
> library(e1071)
Warning message:
package 'e1071' was built under R version 3.5.2
> new<-naiveBayes(alcholic~.,data=trainD)
> new
Naive Bayes Classifier for Discrete Predictors
Call:
naiveBayes.default(x = X, y = Y, laplace = laplace)
A-priori probabilities:
Y
alcholic no alcholic
0.827476 0.172524
Conditional probabilities:
age
Y [,1] [,2]
alcholic 53.94981 9.365189
no alcholic 56.12903 9.234975
gender
Y [,1] [,2]
alcholic 0.6833977 0.4660514
no alcholic 0.6851852 0.4688031
chest_pain
Y [,1] [,2]
alcholic 3.196911 0.9379128
no alcholic 3.074074 1.0434314
blood_pressure
Y [,1] [,2]
alcholic 130.4054 18.34879
no alcholic 128.9074 15.84032
cholestoral
Y [,1] [,2]
alcholic 247.3784 48.57716
new List of 4
testD 227 obs. of 14 variables
trainD 313 obs. of 14 variables
var 540 obs. of 14 variables
Values
data int [1:540] 1 1 1 1 1 1 2 1 1 2 ...

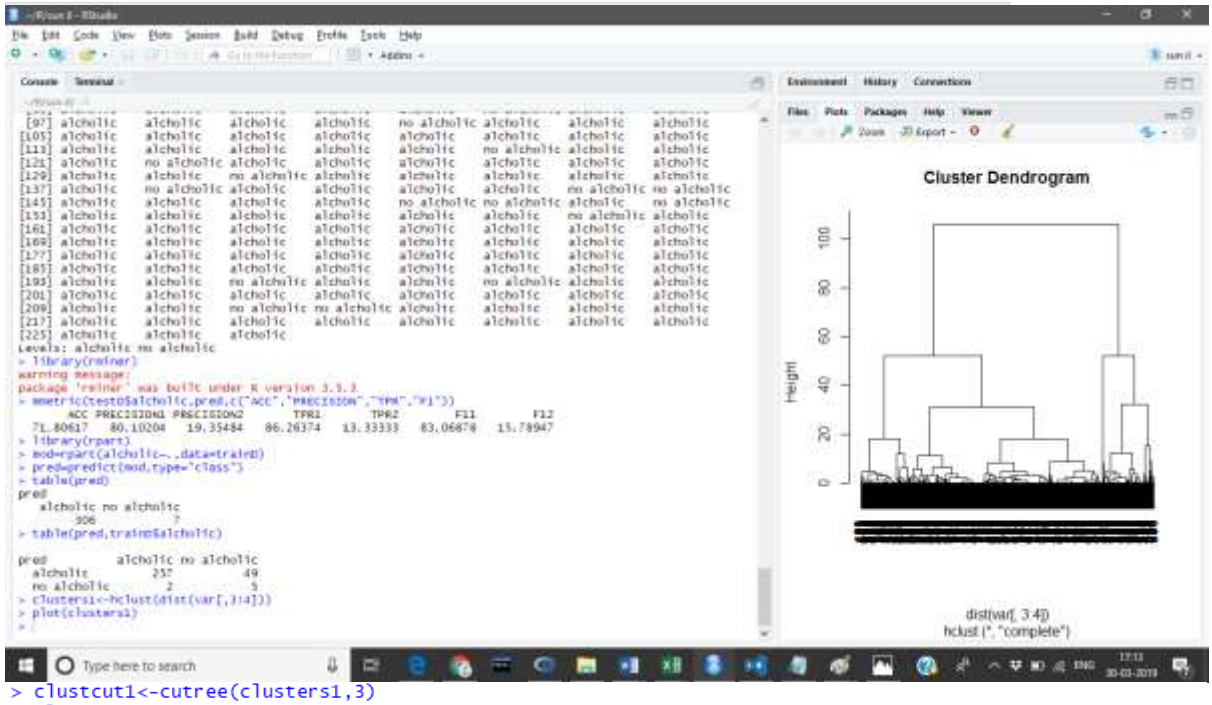




```
> library(rminer)
Warning message:
package 'rminer' was built under R version 3.5.3
> mmetric(testD$alcholic,pred,c("ACC","PRECISION","TPR","F1"))
       ACC PRECISION1 PRECISION2        TPR1        TPR2         F11         F12
  71.80617    80.10204    19.35484    86.26374    13.33333    83.06878    15.78947
>
```

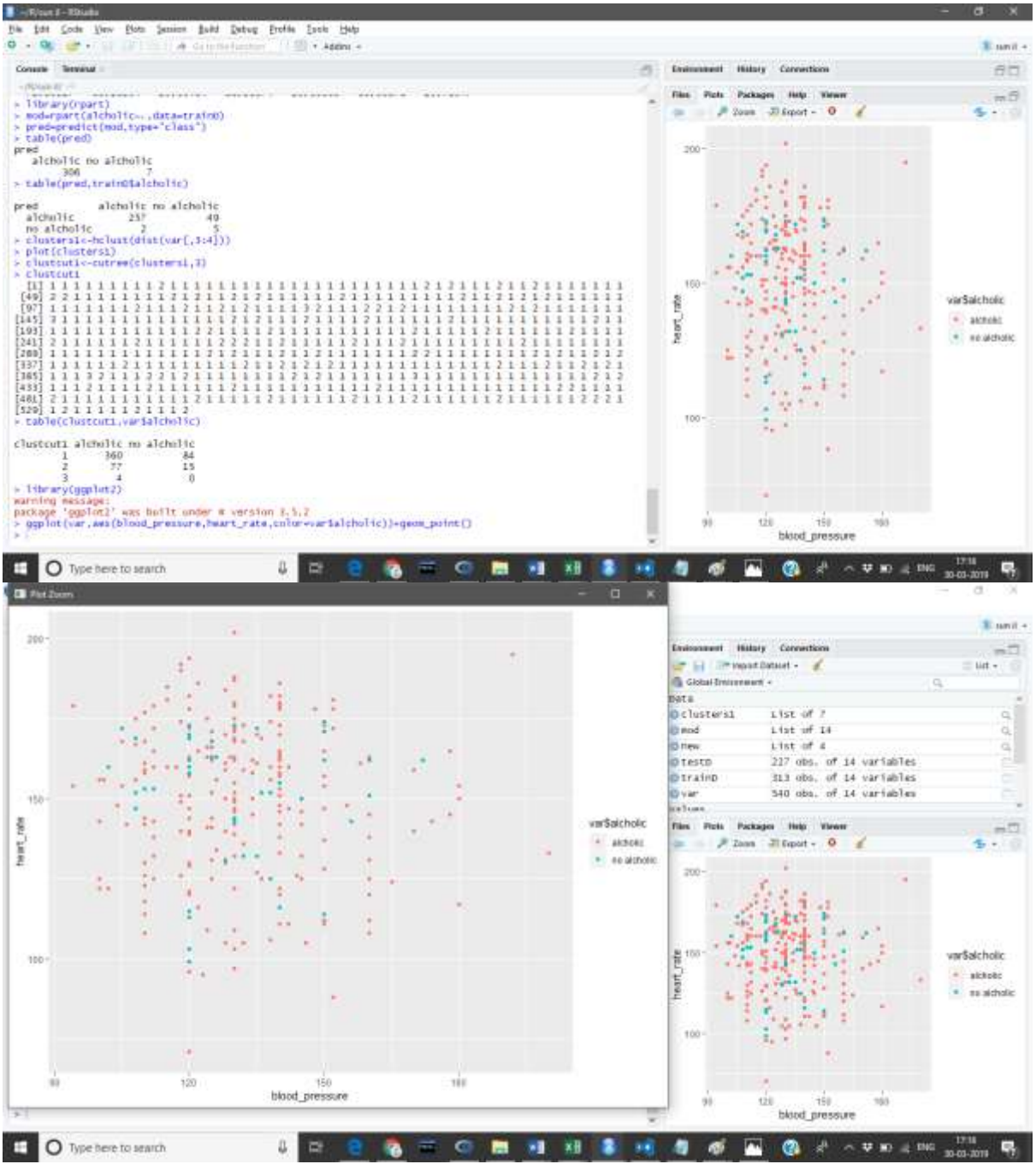
```
> library(rpart)
> mod=rpart(alcholic~.,data=trainD)
> pred=predict(mod,type="class")
> table(pred)
pred
   alcholic no alcholic
        306           7
>
```

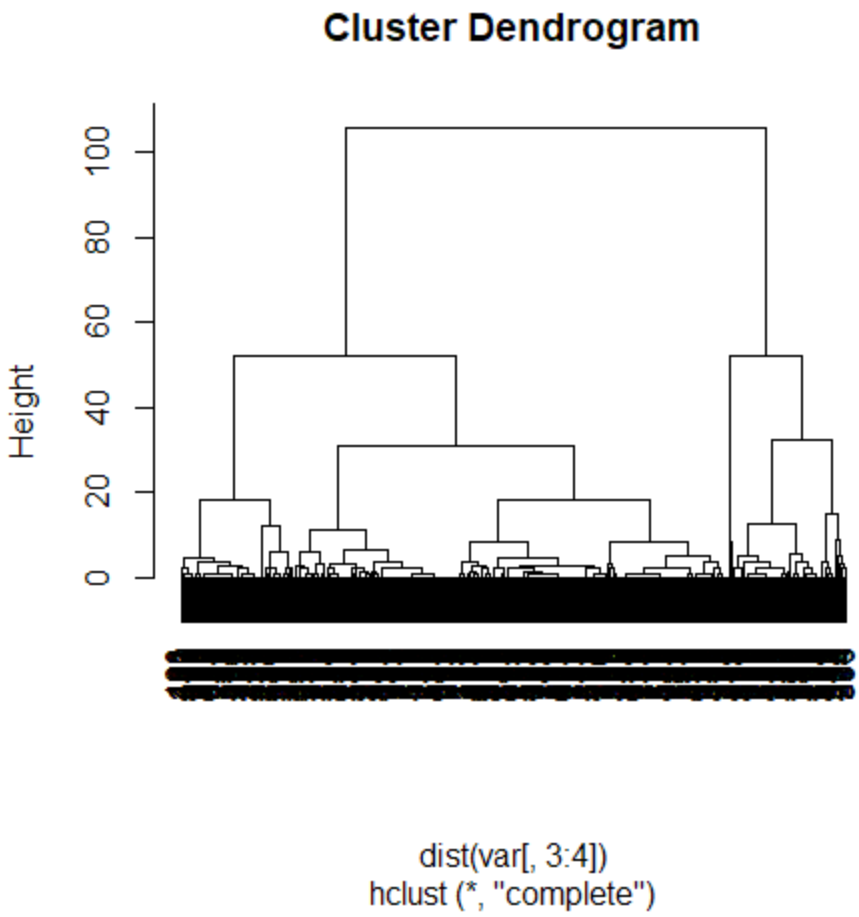
```
> table(pred,trainD$alcholic)

pred          alcholic no alcholic
  alcholic         257           49
  no alcholic        2            5
> 
```




```
> clustcut1<-cutree(clusters1,3)
> clustcut1
  [1] 1 1 1 1 1 1 1 1 1 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 1 2 1 1 1 2 1 1 2 1 1 1 1 1 1 1
 [49] 2 2 1 1 1 1 1 1 1 1 2 1 2 1 1 2 1 2 1 1 1 1 1 1 2 1 1 1 1 1 1 1 1 2 1 1 2 1 2 1 2 1 1 1 1 1
 [97] 1 1 1 1 1 1 1 1 2 1 1 1 2 1 1 2 1 2 1 1 1 1 3 2 1 1 1 2 2 1 2 1 1 1 1 1 1 1 1 2 1 2 1 1 1 1 1 1
[145] 3 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 1 2 1 1 1 2 1 1 1 1 2 1 1 1 1 1 1 2 1 1 1 1 1 1 1 1 1 1 2 1 1
[193] 1 1 1 1 1 1 1 1 1 1 1 1 1 2 2 1 1 1 1 2 1 1 1 1 1 1 1 1 1 1 2 1 1 1 1 1 2 1 1 1 1 1 1 2 1 1 1 1
[241] 2 1 1 1 1 1 1 1 2 1 1 1 1 1 1 2 2 2 1 1 2 1 1 1 1 1 2 1 1 1 1 2 1 1 1 1 1 1 1 1 1 2 1 1 1 1 1 1
[289] 1 1 1 1 1 1 1 1 1 1 1 1 1 2 1 2 1 1 1 2 1 1 2 1 1 1 1 1 1 1 2 2 1 1 1 1 1 1 1 1 2 1 2 1 1 2 1 2
[337] 1 1 1 1 1 1 1 2 1 1 1 1 1 1 1 1 1 2 1 1 2 1 2 1 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 1 1 1 2 1 1 2 1 2 1
[385] 1 1 1 3 2 1 1 1 2 2 1 2 1 1 1 1 1 1 1 1 1 2 1 2 1 1 1 1 1 1 1 3 1 1 1 1 1 1 1 1 1 1 1 1 2 1 2
[433] 1 1 1 2 1 1 1 1 2 1 1 1 1 1 1 2 1 1 1 1 1 1 1 1 1 1 1 2 1 1 1 1 1 1 1 1 1 1 1 1 2 2 1 1 1 1
[481] 2 1 1 1 1 1 1 1 1 1 1 1 2 1 1 1 1 1 1 1 2 1 1 1 1 2 1 1 1 1 1 1 2 1 1 1 1 1 1 2 2 2 1
[529] 1 2 1 1 1 1 1 1 2 1 1 1 2
> 
```

```
> table(clustcut1,var$alcholic)

clustcut1 alcholic no alcholic
        1       360           84
        2        77           15
        3         4            0
> 
```


> library(rpart)
> mod=rpart(alcholic~.,data=train0)
> pred=predict(mod,type="class")
> table(pred)
pred
alcholic no alcholic
306 7
> table(pred,train0$alcholic)
pred alcholic no alcholic
alcholic 257 49
no alcholic 2 5
> clusters1<-hclust(dist(var[,3:4]))
> plot(clusters1)
> clustcut1<-cutree(clusters1,3)
> clustcut1
> table(clustcut1,var$alcholic)
clustcut1 alcholic no alcholic
1 360 84
2 77 15
3 4 0
> library(ggplot2)
warning message:
package 'ggplot2' was built under R version 3.5.2
> ggplot(var,aes(blood_pressure,heart_rate,color=var$alcholic))+geom_point()
heart_rate
blood_pressure
var$alcholic
alcholic
no alcholic
Type here to search
Plot Zoom
Environment
History
Connections
Import Dataset
Global Environment
Data
clusters1 List of 7
mod List of 14
new List of 4
testo 227 obs. of 14 variables
train0 313 obs. of 14 variables
var 540 obs. of 14 variables
Values
Files
Plots
Packages
Help
Viewer
Zoom
Export


Cluster Dendrogram
Height
0
20
40
60
80
100
dist(var[, 3:4])
hclust (*, "complete")

13. Conclusion:

Here, in our project we proposed two effective algorithms, with their implementation namely Hierarchial clustering algorithm and Naïve Bayes classification algorithm. For Naïve Bayes algorithm we got better accuracy of 78.04% for our Geometrical properties of seeds dataset. Even the clustering algorithm, which we have implemented can easily segment into number of clusters need by plotting the centroids of each cluster and also visualizes the distance values of the data.

14. References:

[1.]Daniel Lowd,Pedro Domingos,Naive Bayes Models for Probability Estimation

[2.]Jiangtao Ren , Sau Dan Lee , Xianlu Chen , Ben Kao , Reynold Cheng and David Cheung,Naive Bayes Classification of Uncertain Data

[3.]Ashok AravindanS. M. Anzar,Compression-Based Averaging of Selective Naive Bayes Classifier

[4.]Andrew McCallum,Kamal Nigam,A Comparison of Event Models for Naive Bayes Text Classification

[5.]Mahesh Kini M, Saroja Devi H, Prashant G Desai, Niranjan Chiplunkar,Text Mining Approach to Classify Technical Research Documents using Naïve Bayes

[6.]Maria-Florina Balcan,Yingyu Liang,Pramod Gupta,Robust Hierarchical Clustering

[7.]K.Sasirekha, P.Baby,Agglomerative Hierarchical Clustering Algorithm- A Review

[8.]Ying Zhao and George Karypis,Evaluation of Hierarchical Clustering Algorithms for Document Datasets