Data Analysis

Mehedi Hassan

06/09/2020

# Import Data

# import Data  
Response\_data <- read\_excel("finalll.xlsx")

# Descriptive analysis

str(Response\_data)

Classes ‘tbl\_df’, ‘tbl’ and 'data.frame': 539 obs. of 22 variables:

$ Age : Factor w/ 17 levels "17","18","19",..: 8 8 7 6 5 5 7 8 10 9 ...

$ Gender : chr "male" "male" "male" "Female" ...

$ Agree\_Society\_Rule : chr "yes" "no" "yes" "yes" ...

$ Think\_About\_You : chr "extrovert" "extrovert" "extrovert" "extrovert" ...

$ Have\_Smart\_Phone : chr "yes" "yes" "yes" "yes" ...

$ Like\_Spend\_Time\_With\_friend : chr "yes" "yes" "yes" "yes" ...

$ Happy\_Current\_position : chr "No" "yes" "yes" "yes" ...

$ Tensed\_About\_Carrer : chr "yes" "yes" "no" "yes" ...

$ Discuss\_problem\_anyone : chr "yes" "yes" "yes" "no" ...

$ Feel\_Comfortable\_being\_alone : chr "no" "yes" "no" "no" ...

$ Try\_Committed\_Suicide : chr "no" "no" "no" "no" ...

$ suicide\_is\_the\_solution\_all\_problems : chr "no" "no" "no" "no" ...

$ happy\_with\_family\_members : chr "yes" "yes" "yes" "yes" ...

$ consider\_yourself\_family\_burden : chr "no" "yes" "yes" "yes" ...

$ Where\_feel\_more\_comfortable : chr "family" "family" "friends" "family" ...

$ Family\_Members : chr "3" "4" "4" "3" ...

$ participated\_extra\_curricular\_activities: chr "no" "yes" "yes" "no" ...

$ difficulties\_in\_your\_educational\_system : chr "yes" "yes" "yes" "yes" ...

$ Sleep\_in\_a\_day : chr "8 hours" "less than 6 hours" "6 hours" "8 hours" ...

$ SSC\_Result : num 4.94 4.2 5 4.56 5 4.33 5 5 5 4.5 ...

$ HSC\_Result : num 4 4.5 4.25 4.25 4.5 3.5 4.6 4.2 3.9 4.1 ...

$ University\_CGPA : num 3.5 3.73 3.06 2.96 3.21 2.96 2.9 3.66 3.63 3.57 ...

# Summary of the data

summary(Response\_data)

Age Gender Agree\_Society\_Rule Think\_About\_You Have\_Smart\_Phone Like\_Spend\_Time\_With\_friend

21 : 88 Female:180 no :183 extrovert:308 no : 37 no : 58

22 : 81 male :359 yes:356 introvert:231 yes:502 yes:481

24 : 80

23 : 77

25 : 63

20 : 49

(Other):101

Happy\_Current\_position Tensed\_About\_Carrer Discuss\_problem\_anyone Feel\_Comfortable\_being\_alone

No :183 no :155 no :210 no :209

yes:356 yes:384 yes:329 yes:330

Try\_Committed\_Suicide suicide\_is\_the\_solution\_all\_problems happy\_with\_family\_members

no :459 no :472 no : 70

yes: 80 yes: 67 yes:469

consider\_yourself\_family\_burden Where\_feel\_more\_comfortable Family\_Members

no :330 family :336 4 :198

yes:209 friends:160 5 :115

others : 43 3 :111

6 : 61

7 : 20

8 : 10

(Other): 24

participated\_extra\_curricular\_activities difficulties\_in\_your\_educational\_system Sleep\_in\_a\_day

no :152 no :175 6 hours :102

yes:387 yes:364 7 hours :150

8 hours :140

above 8 hours : 73

less than 6 hours: 74

SSC\_Result HSC\_Result University\_CGPA Family\_Member

5 :251 5 :116 3 : 69 4 :198

4.5 : 40 4 : 67 3.5 : 39 5 :115

4 : 27 4.5 : 46 2.5 : 17 3 :111

4.88 : 23 3.5 : 21 3.1 : 14 6 : 61

4.44 : 15 4.33 : 14 3.25 : 12 7 : 20

4.56 : 13 4.75 : 14 4 : 12 8 : 10

(Other):170 (Other):261 (Other):376 (Other): 24

# Featured Attributes are selected using x-square and P-value for categorical/independent variables

Some important functions are below: 1. subset() : After satisfying a particular condition, the subset functions returns vector, data frame etc. 2. mean() : For using the trimmed-mean. 3. chisq.test() : For the chi-Square test it has been used. 4. Apriori(): It provides the rule of association mining. 5. plot(): A generic object to plot the objects of R. 6. inspect(): It represents all the summary of option, plot and statistics.

##Attach scientific notation in printing with this code  
options(scipen = 0) # if remove than use scipen=999  
  
new\_data<-Response\_data  
X.Squared<-c(1:ncol(new\_data))  
P\_value<-c(1:ncol(new\_data))  
df<-c(1:ncol(new\_data))  
for (i in 1:ncol(new\_data)){  
 a<-chisq.test(table(new\_data[i]))  
 X.Squared[i]<-a$statistic  
 P\_value[i]<-a$p.value  
 df[i]<-a$parameter  
}  
features\_data<-data.frame(X=rep("outcomes",ncol(new\_data)),Y=names(new\_data),X.Squared,df,P\_value)

X Y X.Squared df P\_value

1 outcomes Age 570.60482 16 3.892189e-111

2 outcomes Gender 59.44527 1 1.257430e-14

3 outcomes Agree\_Society\_Rule 55.52690 1 9.218873e-14

4 outcomes Think\_About\_You 11.00000 1 9.111189e-04

5 outcomes Have\_Smart\_Phone 401.15955 1 3.079747e-89

6 outcomes Like\_Spend\_Time\_With\_friend 331.96475 1 3.588113e-74

7 outcomes Happy\_Current\_position 55.52690 1 9.218873e-14

8 outcomes Tensed\_About\_Carrer 97.29314 1 5.978724e-23

9 outcomes Discuss\_problem\_anyone 26.27273 1 2.964455e-07

10 outcomes Feel\_Comfortable\_being\_alone 27.16327 1 1.869778e-07

11 outcomes Try\_Committed\_Suicide 266.49536 1 6.587916e-60

12 outcomes suicide\_is\_the\_solution\_all\_problems 304.31354 1 3.784578e-68

13 outcomes happy\_with\_family\_members 295.36364 1 3.372133e-66

14 outcomes consider\_yourself\_family\_burden 27.16327 1 1.869778e-07

15 outcomes Where\_feel\_more\_comfortable 242.14100 2 2.628751e-53

16 outcomes Family\_Members 1894.44527 18 0.000000e+00

17 outcomes participated\_extra\_curricular\_activities 102.45826 1 4.405538e-24

18 outcomes difficulties\_in\_your\_educational\_system 66.27273 1 3.926541e-16

19 outcomes Sleep\_in\_a\_day 48.28200 4 8.242874e-10

20 outcomes SSC\_Result 8180.35065 69 0.000000e+00

21 outcomes HSC\_Result 3021.06494 84 0.000000e+00

22 outcomes University\_CGPA 1611.21521 133 4.916456e-252

23 outcomes Family\_Member 1894.44527 18 0.000000e+00

## Apriori Algorithm

As the Apriori algorithm has been applied for association rule mining, Where, Support = 45%, Conﬁdence = 95%.

## Apriori Algorithm

As the Apriori algorithm has been applied for association rule mining, Where, Support = 45%, Conﬁdence = 95%.

library(arules)

## Loading required package: Matrix

##   
## Attaching package: 'arules'

## The following objects are masked from 'package:base':  
##   
## abbreviate, write

library(arulesViz)

## Loading required package: grid

## Registered S3 method overwritten by 'seriation':  
## method from   
## reorder.hclust gclus

# Finding association rules  
Rules <- apriori(Response\_data)

Apriori

Parameter specification:

confidence minval smax arem aval originalSupport maxtime support minlen maxlen target ext

0.8 0.1 1 none FALSE TRUE 5 0.1 1 10 rules FALSE

Algorithmic control:

filter tree heap memopt load sort verbose

0.1 TRUE TRUE FALSE TRUE 2 TRUE

Absolute minimum support count: 53

set item appearances ...[0 item(s)] done [0.00s].

set transactions ...[382 item(s), 539 transaction(s)] done [0.00s].

sorting and recoding items ... [53 item(s)] done [0.00s].

creating transaction tree ... done [0.00s].

checking subsets of size 1 2 3 4 5 6 7 8 9 10 done [0.18s].

writing ... [262897 rule(s)] done [0.12s].

creating S4 object ... done [0.15s].

> summary(Rules)

set of 262897 rules

rule length distribution (lhs + rhs):sizes

1 2 3 4 5 6 7 8 9 10

5 234 2788 14127 38514 64039 69715 48593 20310 4572

Min. 1st Qu. Median Mean 3rd Qu. Max.

1.000 6.000 7.000 6.647 8.000 10.000

summary of quality measures:

support confidence lift count

Min. :0.1002 Min. :0.8000 Min. :0.8834 Min. : 54.00

1st Qu.:0.1095 1st Qu.:0.9091 1st Qu.:1.0587 1st Qu.: 59.00

Median :0.1243 Median :0.9545 Median :1.0985 Median : 67.00

Mean :0.1422 Mean :0.9378 Mean :1.1999 Mean : 76.64

3rd Qu.:0.1540 3rd Qu.:0.9823 3rd Qu.:1.1376 3rd Qu.: 83.00

Max. :0.9314 Max. :1.0000 Max. :8.8361 Max. :502.00

mining info:

data ntransactions support confidence

Response\_data 539 0.1 0.8

Rule<-apriori(Response\_data,parameter = list(support=0.45,confidence=0.95))

Apriori

Parameter specification:

confidence minval smax arem aval originalSupport maxtime support minlen maxlen target ext

0.95 0.1 1 none FALSE TRUE 5 0.45 1 10 rules FALSE

Algorithmic control:

filter tree heap memopt load sort verbose

0.1 TRUE TRUE FALSE TRUE 2 TRUE

inspect(head(sort(Rule,by="lift"),8))

**From a huge numbers of rules I inspected 8 major rules.**

**SL No> LHS> RHS> Support> Confidence >Lift>Count**

[1] {Have\_Smart\_Phone=yes,

Like\_Spend\_Time\_With\_friend=yes,

Where\_feel\_more\_comfortable=family} => {happy\_with\_family\_members=yes} 0.5491651 0.9801325 1.126421 296

[2] {Have\_Smart\_Phone=yes,

Like\_Spend\_Time\_With\_friend=yes,

suicide\_is\_the\_solution\_all\_problems=no,

Where\_feel\_more\_comfortable=family} => {happy\_with\_family\_members=yes} 0.5194805 0.9790210 1.125144 280

[3] {Have\_Smart\_Phone=yes,

Like\_Spend\_Time\_With\_friend=yes,

Try\_Committed\_Suicide=no,

Where\_feel\_more\_comfortable=family} => {happy\_with\_family\_members=yes} 0.5083488 0.9785714 1.124627 274

[4] {Have\_Smart\_Phone=yes,

Like\_Spend\_Time\_With\_friend=yes,

Try\_Committed\_Suicide=no,

suicide\_is\_the\_solution\_all\_problems=no,

Where\_feel\_more\_comfortable=family} => {happy\_with\_family\_members=yes} 0.4897959 0.9777778 1.123715 264

[5] {Like\_Spend\_Time\_With\_friend=yes,

Where\_feel\_more\_comfortable=family} => {happy\_with\_family\_members=yes} 0.5677180 0.9776358 1.123552 306

[6] {Have\_Smart\_Phone=yes,

Try\_Committed\_Suicide=no,

Where\_feel\_more\_comfortable=family} => {happy\_with\_family\_members=yes} 0.5324675 0.9761905 1.121891 287

[7] {Like\_Spend\_Time\_With\_friend=yes,

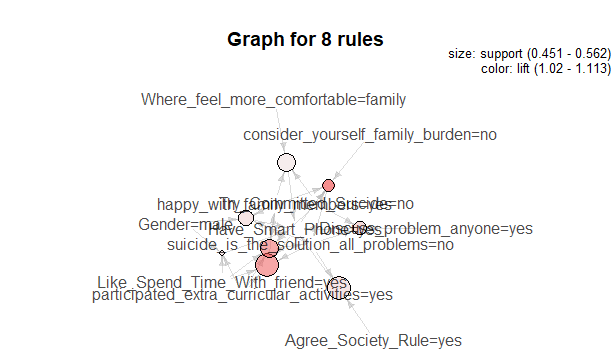
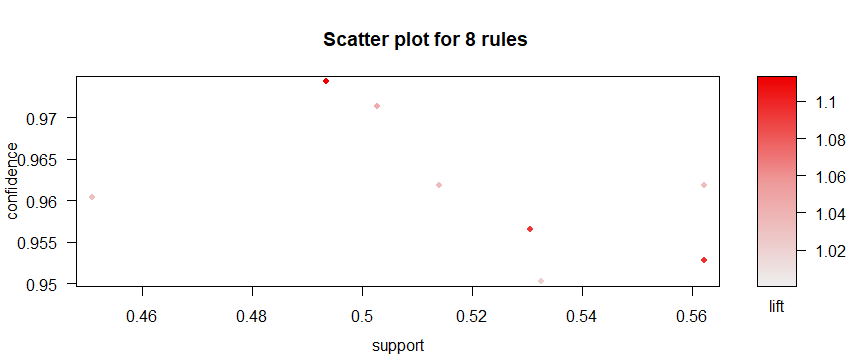
suicide\_is\_the\_solution\_all\_problems=no,

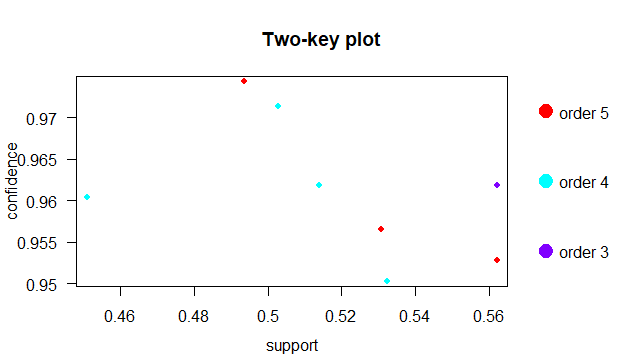
Where\_feel\_more\_comfortable=family} => {happy\_with\_family\_members=yes} 0.5306122 0.9761092 1.121797 286

[8] {Like\_Spend\_Time\_With\_friend=yes,

Try\_Committed\_Suicide=no,

Where\_feel\_more\_comfortable=family} => {happy\_with\_family\_members=yes} 0.5250464 0.9758621 1.121513 283



****

**Have a good thesis Paper!**

**Thank You**

**Mehedi Hassan**

**CSE, North Western University**

**CEO, The Virtual BD**

**+8801902624501**