ASSOCIATION RULES ANALYSIS

On Student Alcohol Consumption Dataset

AUTHOR: Vanessa Fotso

JUNE 9, 2020

UMGC – SUMMER 2020

DATA 630 – MACHINE LEARNING

Introduction

The aim of this report is to identify any relationship between students' social factors and their likelihood to consume alcohol. Alcohol consumption in secondary education has been a pertaining problem. Alcohol is in fact the most commonly abused drug among youth in the USA. According to the CDC, individuals aged between 12 and 21years old made approximately 119,000 emergency rooms visits for injuries and conditions related to alcohol in 2013. Disruptive consumption of alcohol among minors may lead to education failure, physical problems, death from alcohol poisoning and unwanted sexual activity. We will be using the association rules mining to analyze the Student Alcohol Consumption dataset. The method uses the Apriori algorithm to identify frequent patterns and find association between attributes in a set. The algorithm generate rules based on statistical metrics such as support and confidence. Support is the probability an antecedent occurs in the dataset divided by the total number of observations. Confidence shows the probability of occurrence of the consequent based on the probability of the antecedents. We hope the Apriori rule method will help us derive the combination of factors that may lead to alcohol consumption among students

Data Description

The Student Alcohol Consumption dataset analyzed using R programming in this report was taken from the archives of the Machine Learning repository of the University of California, Irvine (UCI). The data is a multivariate dataset collected from a survey from high school students with a mix of categorical and numerical variables. we used the str() function to check the data structure and identify the type of variables present. The output in figure 1 reveals that the dataset contains 33 attributes and 395 cases representing students. The variables name as well as the type are also displayed in figure 1. Some variables name include age, gender, family size, education background, daily alcohol consumption, health, absences and grades. A full detail of the variables is listed in the appendix section. We can also note that there is no identification key in the output, so we will not need to drop a column from the data.

```
'data.frame':   395 obs. of  33 variables:
 $ school    : Factor w/ 2 levels "GP","MS": 1 1 1 1 1 1 1 1 1 1 ...
 $ sex       : Factor w/ 2 levels "F","M": 1 1 1 1 1 2 2 1 2 2 ...
 $ age       : int  18 17 15 15 16 16 16 17 15 15 ...
 $ address   : Factor w/ 2 levels "R","U": 2 2 2 2 2 2 2 2 2 2 ...
 $ famsize   : Factor w/ 2 levels "GT3","LE3": 1 1 2 1 1 2 2 1 2 1 ...
 $ Pstatus   : Factor w/ 2 levels "A","T": 1 2 2 2 2 2 2 1 1 2 ...
 $ Medu      : int  4 1 1 4 3 4 2 4 3 3 ...
 $ Fedu      : int  4 1 1 2 3 3 2 4 2 4 ...
 $ Mjob      : Factor w/ 5 levels "at_home","health",..: 1 1 1 2 3 4 3 3 4 3 ...
 $ Fjob      : Factor w/ 5 levels "at_home","health",..: 5 3 3 4 3 3 3 5 3 3 ...
 $ reason    : Factor w/ 4 levels "course","home",..: 1 1 3 2 2 4 2 2 2 2 ...
 $ guardian  : Factor w/ 3 levels "father","mother",..: 2 1 2 2 1 2 2 2 2 2 ...
 $ traveltime: int  2 1 1 1 1 1 1 2 1 1 ...
 $ studytime : int  2 2 2 3 2 2 2 2 2 2 ...
 $ failures  : int  0 0 3 0 0 0 0 0 0 0 ...
 $ schoolsup : Factor w/ 2 levels "no","yes": 2 1 2 1 1 1 1 2 1 1 ...
 $ famsup    : Factor w/ 2 levels "no","yes": 1 2 1 2 2 2 1 2 2 2 ...
 $ paid      : Factor w/ 2 levels "no","yes": 1 1 2 2 2 2 1 1 2 2 ...
 $ activities: Factor w/ 2 levels "no","yes": 1 1 1 2 1 2 1 1 1 2 ...
 $ nursery   : Factor w/ 2 levels "no","yes": 2 1 2 2 2 2 2 2 2 2 ...
 $ higher    : Factor w/ 2 levels "no","yes": 2 2 2 2 2 2 2 2 2 2 ...
 $ internet  : Factor w/ 2 levels "no","yes": 1 2 2 2 1 2 2 1 2 2 ...
 $ romantic  : Factor w/ 2 levels "no","yes": 1 1 1 2 1 1 1 1 1 1 ...
 $ famrel    : int  4 5 4 3 4 5 4 4 4 5 ...
 $ freetime  : int  3 3 3 2 3 4 4 1 2 5 ...
 $ goout     : int  4 3 2 2 2 2 4 4 2 1 ...
 $ Dalc      : int  1 1 2 1 1 1 1 1 1 1 ...
 $ Walc      : int  1 1 3 1 2 2 1 1 1 1 ...
 $ health    : int  3 3 3 5 5 5 3 1 1 5 ...
 $ absences  : int  6 4 10 2 4 10 0 6 0 0 ...
 $ G1        : int  5 5 7 15 6 15 12 6 16 14 ...
 $ G2        : int  6 5 8 14 10 15 12 5 18 15 ...
 $ G3        : int  6 6 10 15 10 15 11 6 19 15 ...
```

Figure 1: Student Alcohol Consumption Dataset Structure

Next, we use the summary() function to understand the data distribution. The output in figure 2 shows six descriptive statistics for numeric variables in the dataset. Those statistics include the minimum value, maximum value, 1st and 3rd quartile, median, and mean. Additionally, we can note that there's no missing values in the dataset; however, given the nature of the analysis being performed, we will need to transform all numerical variables to categorical in order to conduct the Apriori rules method. We can also observe that most numerical variables have small range. The attribute age for example ranges between 15 and 22, and the daily alcohol consumption (dalc) ranges between 1 and . We can also note the large range of the absence variable, having a range between 0 and 75. This variable will be helpful in determining if alcohol use has an impact on absences.

```
> summary(alcohol)
 school   sex          age       address famsize   Pstatus      Medu            Fedu             Mjob
 GP:349   F:208   Min.   :15.0   R: 88   GT3:281   A: 41   Min.   :0.000   Min.   :0.000   at_home : 59
 MS: 46   M:187   1st Qu.:16.0   U:307   LE3:114   T:354   1st Qu.:2.000   1st Qu.:2.000   health  : 34
                  Median :17.0                             Median :3.000   Median :2.000   other   :141
                  Mean   :16.7                             Mean   :2.749   Mean   :2.522   services:103
                  3rd Qu.:18.0                             3rd Qu.:4.000   3rd Qu.:3.000   teacher : 58
                  Max.   :22.0                             Max.   :4.000   Max.   :4.000
       Fjob            reason      guardian     traveltime      studytime        failures      schoolsup
 at_home : 20   course    :145   father: 90   Min.   :1.000   Min.   :1.000   Min.   :0.0000   no :344
 health  : 18   home      :109   mother:273   1st Qu.:1.000   1st Qu.:1.000   1st Qu.:0.0000   yes: 51
 other   :217   other     : 36   other : 32   Median :1.000   Median :2.000   Median :0.0000
 services:111   reputation:105                Mean   :1.448   Mean   :2.035   Mean   :0.3342
 teacher : 29                                 3rd Qu.:2.000   3rd Qu.:2.000   3rd Qu.:0.0000
                                              Max.   :4.000   Max.   :4.000   Max.   :3.0000
 famsup     paid     activities nursery   higher    internet  romantic      famrel          freetime
 no :153   no :214   no :194    no : 81   no : 20   no : 66   no :263   Min.   :1.000   Min.   :1.000
 yes:242   yes:181   yes:201    yes:314   yes:375   yes:329   yes:132   1st Qu.:4.000   1st Qu.:3.000
                                                                        Median :4.000   Median :3.000
                                                                        Mean   :3.944   Mean   :3.235
                                                                        3rd Qu.:5.000   3rd Qu.:4.000
                                                                        Max.   :5.000   Max.   :5.000
     goout            Dalc            Walc           health         absences            G1
 Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   : 0.000   Min.   : 3.00
 1st Qu.:2.000   1st Qu.:1.000   1st Qu.:1.000   1st Qu.:3.000   1st Qu.: 0.000   1st Qu.: 8.00
 Median :3.000   Median :1.000   Median :2.000   Median :4.000   Median : 4.000   Median :11.00
 Mean   :3.109   Mean   :1.481   Mean   :2.291   Mean   :3.554   Mean   : 5.709   Mean   :10.91
 3rd Qu.:4.000   3rd Qu.:2.000   3rd Qu.:3.000   3rd Qu.:5.000   3rd Qu.: 8.000   3rd Qu.:13.00
 Max.   :5.000   Max.   :5.000   Max.   :5.000   Max.   :5.000   Max.   :75.000   Max.   :19.00
       G2              G3
 Min.   : 0.00   Min.   : 0.00
 1st Qu.: 9.00   1st Qu.: 8.00
 Median :11.00   Median :11.00
 Mean   :10.71   Mean   :10.42
 3rd Qu.:13.00   3rd Qu.:14.00
 Max.   :19.00   Max.   :20.00
```

Figure 2: Descriptive Statistics for All Variables.

Since we are interested in determining the effect of alcohol consumption on other variables, it will be interesting to look at the distribution of the daily alcohol distribution (Dalc). We can see from the histogram in figure 3 that the daily alcohol consumption is skewed to the right, with more than 75% of students having a low daily consumption.

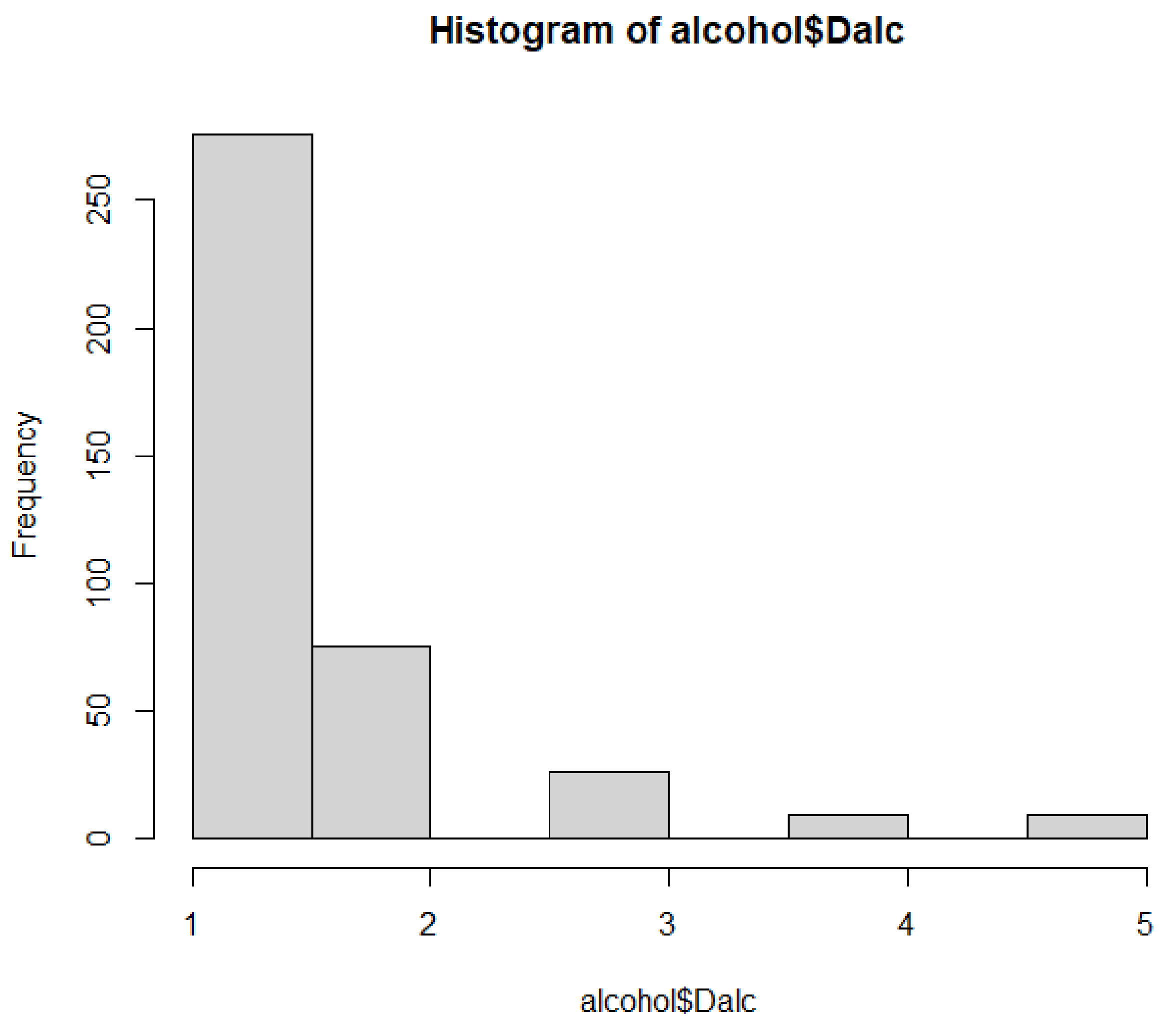

Figure 3: Student Daily Alcohol Consumption

Data Preprocessing

As noted above, in order to use the Apriori Rule method, we will need to convert all the numerical variables to discrete or factor variables. For variables with small range like age and medium use (medu), we will use the factor() function to covert to categorical variables. Because this is a survey dataset, most variables have limited values allowed, favorizing scaled responses (low/high, bad/good, etc.). For variable with relatively large range (G1, G2, & G3), we will use an interval method discretization, and the fixed method discretization for absence variable. The commands for discretization are included in the appendix below. We also added labels to the study time, family relation (famrel), free time, go out, daily alcohol consumption and health variables to facilitate the supervised learning. Figure 4 below display the summary of the all variable in the dataset after conversion of numerical variables to categorical variables.

```
> summary(alcohol)
 school    sex          age        address famsize    Pstatus Medu     Fedu          Mjob            Fjob
 GP:349    F:208    16       :104   R: 88   GT3:281    A: 41   0:  3    0:  2    at_home : 59    at_home : 20
 MS: 46    M:187    17       : 98   U:307   LE3:114    T:354   1: 59    1: 82    health  : 34    health  : 18
                    15       : 82                               2:103    2:115    other   :141    other   :217
                    18       : 82                               3: 99    3:100    services:103    services:111
                    19       : 24                               4:131    4: 96    teacher : 58    teacher : 29
                    20       :  3
                    (Other):  2
        reason        guardian    traveltime        studytime      failures schoolsup famsup      paid     activities
 course     :145    father: 90    1:257      <2hrs        :105    0:312     no :344   no :153   no :214   no :194
 home       :109    mother:273    2:107      2 to 5 hrs :198     1: 50     yes: 51   yes:242   yes:181   yes:201
 other      : 36    other : 32    3: 23      5 to 10 hrs: 65     2: 17
 reputation:105                   4:  8      over 10 hrs: 27     3: 16


 nursery    higher     internet   romantic        famrel          freetime          goout              Dalc
 no : 81    no : 20    no : 66    no :263    very bad :  8    very low : 19    very low : 23    very low :276
 yes:314    yes:375    yes:329    yes:132    bad      : 18    low      : 64    low      :103    low      : 75
                                             fair     : 68    medium   :157    medium   :130    medium   : 26
                                             good     :195    high     :115    high     : 86    high     :  9
                                             very good:106    very high: 40    very high: 53    very high:  9

       Walc            health          absences            G1                  G2                 G3
 very low :151    very bad : 47    [0,5)  :244    [3,6.2)     : 33    [0,3.8)     : 13    [0,4)  : 38
 low      : 85    bad      : 45    [5,10) : 68    [6.2,9.4)  :109    [3.8,7.6)  : 51    [4,8)  : 32
 medium   : 80    fair     : 91    [10,15): 47    [9.4,12.6) :125    [7.6,11.4) :163    [8,12) :163
 high     : 51    good     : 66    [15,75]: 36    [12.6,15.8): 87    [11.4,15.2):135    [12,16):122
 very high: 28    very good:146                   [15.8,19]  : 41    [15.2,19]  : 33    [16,20]: 40
```

Figure 4: Summary of the Discretized Outputs

As we can see in figure 4, all variables are now categorical and we are now ready to run association rules method

Association Rules Analysis

The main goal of the analysis is to generate association rules linked to alcohol consumption. In the previous steps, we labeled the daily consumption of alcohol as very low, low, medium, high and very high. We ran two separate sets of rules one on Dalc: one with low or very low and another one with high or very high, then we compared the results of the two runs.

For te run where Dalc is very low/low, we used a minimum support of 0.2, a minimum confidence level of 0.8, and a minimum length of 2. The result for this run is displayed in figure 5 below.

```
> #Run apriori method rules to get rules for low daily alcohol use
> alclow<-apriori(alcohol, parameter= list(supp=0.2, conf=0.8, minlen=2), appearance=l
ist(rhs=c("Dalc=very low", "Dalc=low"), default="lhs"))
Apriori

Parameter specification:
 confidence minval smax arem  aval originalSupport maxtime support minlen maxlen
        0.8    0.1    1 none FALSE            TRUE       5     0.2      2     10
 target  ext
  rules TRUE

Algorithmic control:
 filter tree heap memopt load sort verbose
    0.1 TRUE TRUE  FALSE TRUE    2    TRUE

Absolute minimum support count: 79

set item appearances ...[2 item(s)] done [0.00s].
set transactions ...[122 item(s), 395 transaction(s)] done [0.06s].
sorting and recoding items ... [67 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 [0.77s].
writing ... [1404 rule(s)] done [0.13s].
creating S4 object  ... done [0.15s].
Warning message:
In apriori(alcohol, parameter = list(supp = 0.2, conf = 0.8, minlen = 2),  :
  Mining stopped (maxlen reached). Only patterns up to a length of 10 returned!
> alclow
set of 1404 rules
> 
```

Figure 5: Apriori Output for Dalc = very low / low On the Left-Hand Side

The above run took 122 items as the method input and returned 1,404 rules.

Next, we ran some commands to sort the rules by lift and eliminate the redundant rules or any subset of a more general rule. The lift parameter help us evaluate the strength of the rule by providing the degree of correlation between the antecedent and the consequent. Figure 6 below display the summary of the rules after the pruning step. The number of rules has decreased from 1,404 to 362 rules, and Figure 7 displays the top 5 strongest rules, arranged by lift value.

```
> summary(rules.pruned)
set of 362 rules

rule length distribution (lhs + rhs):sizes
 2  3  4  5  6  7  8  9 
 2 19 60 92 87 63 33  6 

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  2.000   5.000   6.000   5.641   7.000   9.000 

summary of quality measures:
    support         confidence        coverage           lift      
 Min.   :0.2000   Min.   :0.8000   Min.   :0.2000   Min.   :1.145  
 1st Qu.:0.2101   1st Qu.:0.8054   1st Qu.:0.2557   1st Qu.:1.153  
 Median :0.2278   Median :0.8148   Median :0.2785   Median :1.166  
 Mean   :0.2398   Mean   :0.8249   Mean   :0.2911   Mean   :1.181  
 3rd Qu.:0.2557   3rd Qu.:0.8333   3rd Qu.:0.3089   3rd Qu.:1.193  
 Max.   :0.4228   Max.   :1.0000   Max.   :0.5266   Max.   :1.431  
     count       
 Min.   : 79.00  
 1st Qu.: 83.00  
 Median : 90.00  
 Mean   : 94.72  
 3rd Qu.:101.00  
 Max.   :167.00  

mining info:
    data ntransactions support confidence
 alcohol           395     0.2        0.8
```

Figure 6: Remaining Rules After Pruning Redundant Rules

```
> # preview the top 5 rules by lift
> inspect(head(sort(rules.pruned, by="lift")),n=5)
    lhs                                rhs               support   confidence
[1] {activities=yes,Walc=very low} => {Dalc=very low} 0.2177215 1.0000000
[2] {sex=F,Walc=very low}          => {Dalc=very low} 0.2379747 1.0000000
[3] {Fjob=other,Walc=very low}     => {Dalc=very low} 0.2000000 1.0000000
[4] {romantic=no,Walc=very low}    => {Dalc=very low} 0.2531646 1.0000000
[5] {higher=yes,Walc=very low}     => {Dalc=very low} 0.3645570 1.0000000
[6] {Walc=very low}                => {Dalc=very low} 0.3797468 0.9933775
    coverage  lift     count
[1] 0.2177215 1.431159  86
[2] 0.2379747 1.431159  94
[3] 0.2000000 1.431159  79
[4] 0.2531646 1.431159 100
[5] 0.3645570 1.431159 144
[6] 0.3822785 1.421682 150
>
```

Figure 7: Top 5 Association rules by lift Value

The top five rules suggest that all students (confidence = 1) who have activities and drink less over the weekend are less likely to have high daily alcohol intake.

A set of 362 rules is still quite high. Increasing the minimum support value to 0.35 have reduced the number of rules to 13, thus improving the algorithm efficiency.

We then plot a parallel coordinate of 12 rules to better highlight the rules.

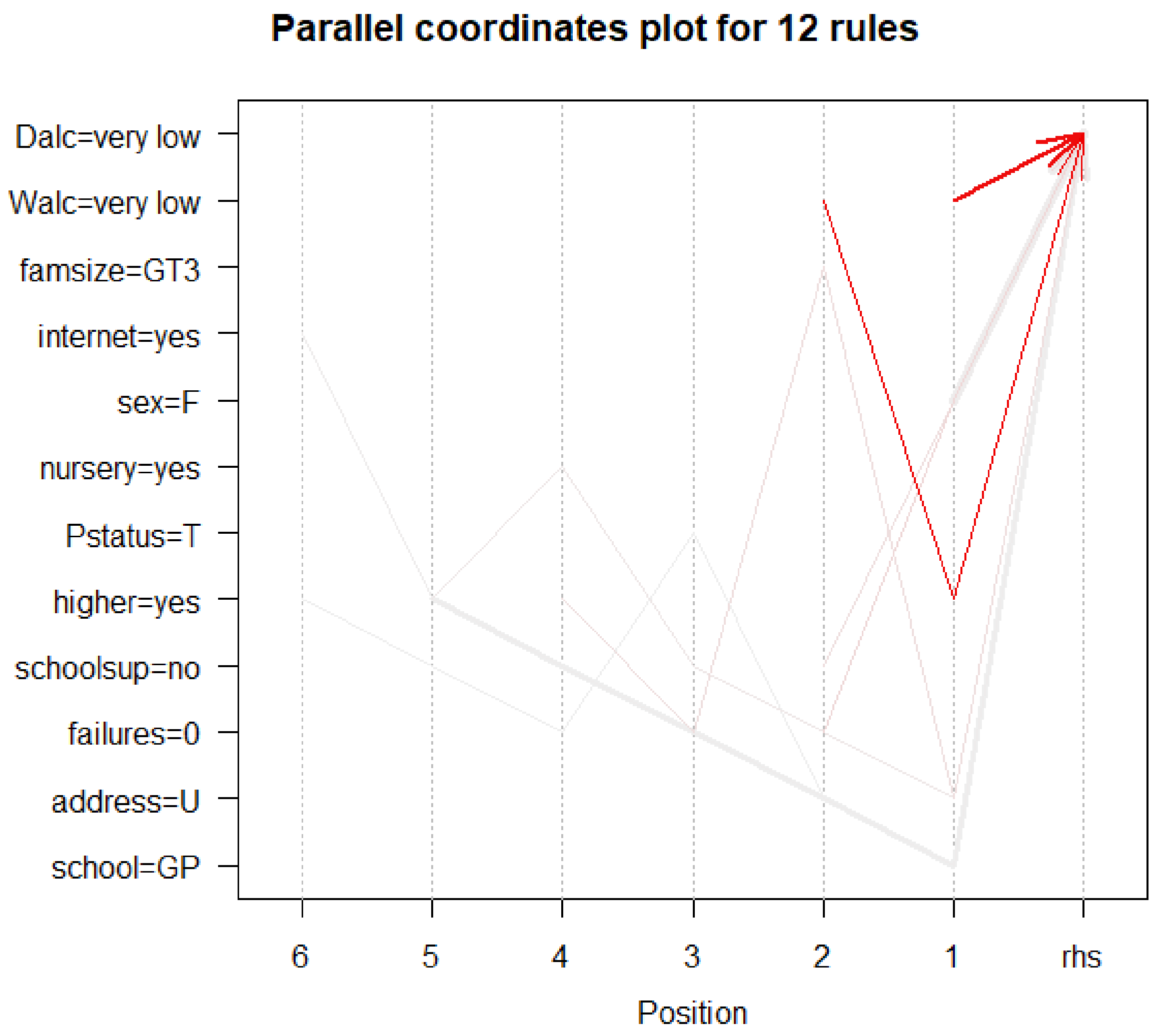

Figure 8: Parallel Coordinates Plot for 12 rules

Figure 8 shows that female students going to GP school who have internet and are not in a relationship are likely to have a very low daily alcohol intake.

Simultaneously, we run the algorithm to generate the rules leading to a medium/high/very high daily alcohol intake, with the same parameter as the first run (a minimum support of 0.2, a minimum confidence level of 0.8, and a minimum length of 2.). However, the algorithm returned zero rules. So, we adjusted the parameters as follow: min-supp = 0.05, confidence = 0.1 and min length = 2, and

the algorithm generated 3 non redundant rules as seen in figure 9 below.

```
> #Run apriori method rules to get rules for high daily alcohol use
> alchigh<-apriori(alcohol, parameter= list(supp=0.05, conf=0.1, minlen=2), appeara
nce=list(rhs=c("Dalc=very high", "Dalc=high","Dalc=medium"), default="lhs"))
Apriori

Parameter specification:
 confidence minval smax arem  aval originalSupport maxtime support minlen maxlen
        0.1    0.1    1 none FALSE            TRUE       5    0.05      2     10
 target  ext
  rules TRUE

Algorithmic control:
 filter tree heap memopt load sort verbose
    0.1 TRUE TRUE  FALSE TRUE    2    TRUE

Absolute minimum support count: 19

set item appearances ...[3 item(s)] done [0.00s].
set transactions ...[122 item(s), 395 transaction(s)] done [0.02s].
sorting and recoding items ... [107 item(s)] done [0.00s].
creating transaction tree ... done [0.00s].
checking subsets of size 1 2 3 4 5 6 7 8 done [6.30s].
writing ... [3 rule(s)] done [0.33s].
creating S4 object  ... done [0.23s].
Warning message:
In apriori(alcohol, parameter = list(supp = 0.05, conf = 0.1, minlen = 2),  :
  Mining stopped (time limit reached). Only patterns up to a length of 8 returned!
> alchigh
set of 3 rules
> 
```

Figure 9: Apriori Output for Medium Daily Alcohol Intake in the RHS

Figure 10 displays the 3 generated rules, which can be visualized in figure 11.

```
> # view the 3 generated rules by lift
> inspect(head(sort(rules.pruned1, by="lift")))
    lhs                        rhs            support    confidence coverage
[1] {sex=M,higher=yes}      => {Dalc=medium} 0.05063291 0.1169591  0.4329114
[2] {sex=M,schoolsup=no}    => {Dalc=medium} 0.05063291 0.1162791  0.4354430
[3] {sex=M}                 => {Dalc=medium} 0.05316456 0.1122995  0.4734177
    lift     count
[1] 1.776878 20
[2] 1.766547 20
[3] 1.706088 21
```

Figure 10 : 3 Rules Generated for Medium Daily Alcohol Intake in RHS

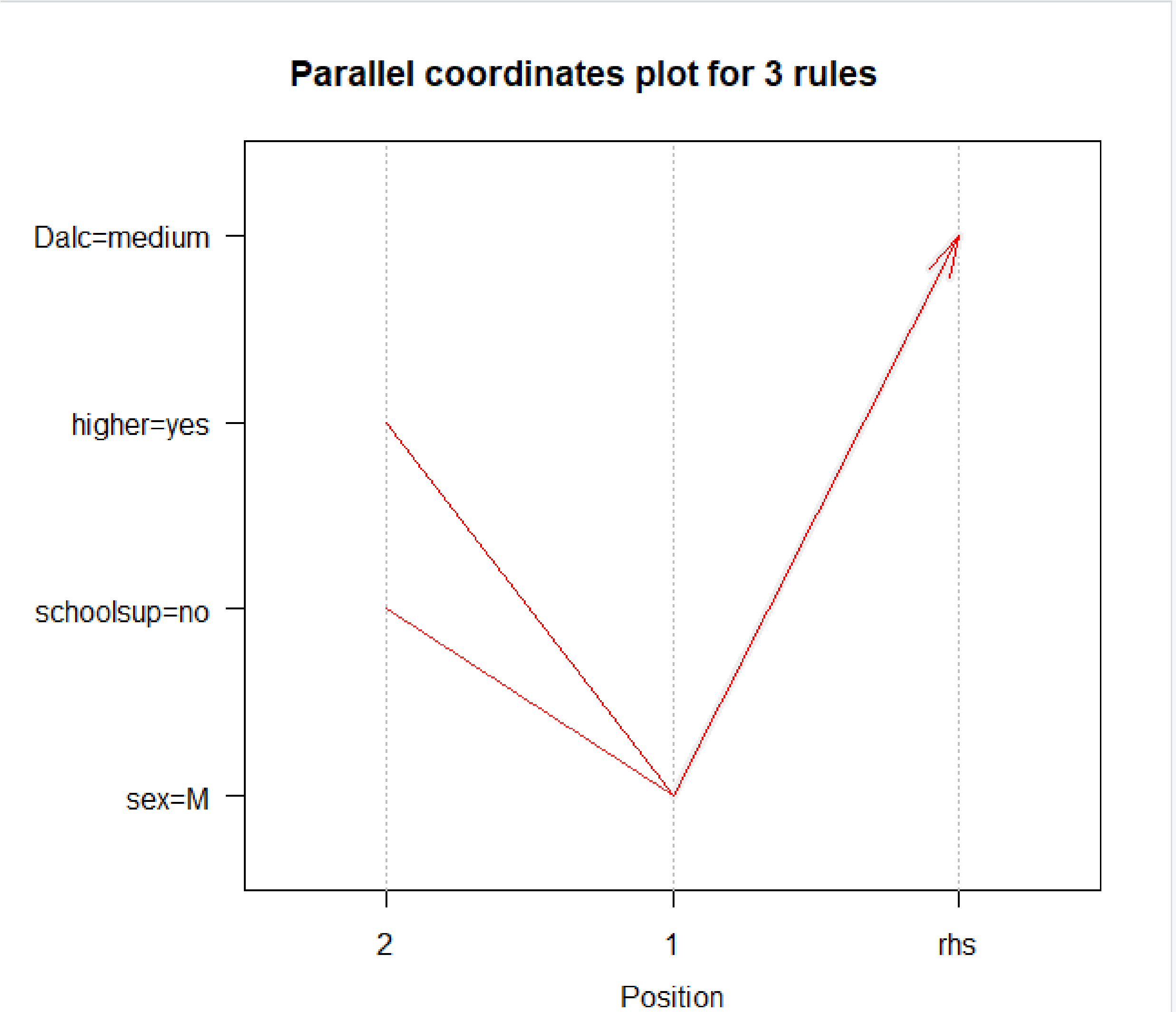

Figure 11: Parallel Coordinates Plot for the 3 rules With Dalc in RHS

The rules indicates that moderate daily alcohol consumption is positively correlated with male students with no school sup.

Conclusion

Association rules are useful in identifying frequent patterns in the data. They can scan a large database and provide many "if this, then that" rules. The method was efficient in answering our research question; however, multiple parameter must be tested to generate the desired rules. We derived from the analysis that students with low alcohol consumption are mostly single females who have internet. The data set provided in this analysis can be analyzed in various ways. Similar analysis could be done to evaluate the student's academic performance based on other attributes in the data. Or how alcohol may influence a student GPA or class attendance.

References

Centers for Disease Control and Prevention (CDC). Alcohol-Related Disease Impact (ARDI). Atlanta, GA: CDC.

https://archive.ics.uci.edu/ml/datasets/STUDENT+ALCOHOL+CONSUMPTION

Appendix

Attributes for both student-mat.csv (Math course) and student-por.csv (Portuguese language course) datasets:

1. school - student's school (binary: 'GP' - Gabriel Pereira or 'MS' - Mousinho da Silveira)
2. sex - student's sex (binary: 'F' - female or 'M' - male)
3. age - student's age (numeric: from 15 to 22)
4. address - student's home address type (binary: 'U' - urban or 'R' - rural)
5. famsize - family size (binary: 'LE3' - less or equal to 3 or 'GT3' - greater than 3)
6. Pstatus - parent's cohabitation status (binary: 'T' - living together or 'A' - apart)
7. Medu - mother's education (numeric: 0 - none, 1 - primary education (4th grade), 2 – 5th to 9th grade, 3 – secondary education or 4 – higher education)
8. Fedu - father's education (numeric: 0 - none, 1 - primary education (4th grade), 2 – 5th to 9th grade, 3 – secondary education or 4 – higher education)
9. Mjob - mother's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at_home' or 'other')
10. Fjob - father's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at_home' or 'other')
11. reason - reason to choose this school (nominal: close to 'home', school 'reputation', 'course' preference or 'other')
12. guardian - student's guardian (nominal: 'mother', 'father' or 'other')
13. traveltime - home to school travel time (numeric: 1 - <15 min., 2 - 15 to 30 min., 3 - 30 min. to 1 hour, or 4 - >1 hour)

14. studytime - weekly study time (numeric: 1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - >10 hours)
15. failures - number of past class failures (numeric: n if 1<=n<3, else 4)
16. schoolsup - extra educational support (binary: yes or no)
17. famsup - family educational support (binary: yes or no)
18. paid - extra paid classes within the course subject (Math or Portuguese) (binary: yes or no)
19. activities - extra-curricular activities (binary: yes or no)
20. nursery - attended nursery school (binary: yes or no)
21. higher - wants to take higher education (binary: yes or no)
22. internet - Internet access at home (binary: yes or no)
23. romantic - with a romantic relationship (binary: yes or no)
24. famrel - quality of family relationships (numeric: from 1 - very bad to 5 - excellent)
25. freetime - free time after school (numeric: from 1 - very low to 5 - very high)
26. goout - going out with friends (numeric: from 1 - very low to 5 - very high)
27. Dalc - workday alcohol consumption (numeric: from 1 - very low to 5 - very high)
28. Walc - weekend alcohol consumption (numeric: from 1 - very low to 5 - very high)
29. health - current health status (numeric: from 1 - very bad to 5 - very good)
30. absences - number of school absences (numeric: from 0 to 93)

**Copied directly from https://www.kaggle.com/uciml/student-alcohol-consumption/home

R Code for the analysis

```
#load libraries

library("arules")

library("arulesViz")


#Load the Student Alcohol consumption data

alcohol<-read.csv(file="student-mat.csv", head=TRUE, sep=";", as.is = FALSE)


#run the Str and summary commands to acquaint ourself with the data

str(alcohol)

summary(alcohol)


#graph and review daily alcohol consumption


hist(alcohol$Dalc)


#Discretize numerical variables and add labels for supervised learning

alcohol$age<-factor(alcohol$age)

alcohol$Medu <-factor(alcohol$Medu)

alcohol$Fedu <-factor(alcohol$Fedu)

alcohol$traveltime <-factor(alcohol$traveltime)

alcohol$studytime <-factor(alcohol$studytime, labels = c("<2hrs", "2 to 5 hrs","5
to 10 hrs","over 10 hrs"))
```

```
alcohol$famrel <-factor(alcohol$famrel, labels = c("very bad", "bad","fair","good","very good"))
alcohol$failures <-factor(alcohol$failures)
alcohol$freetime <-factor(alcohol$freetime, labels = c("very low", "low","medium","high","very high"))
alcohol$goout <-factor(alcohol$goout, labels = c("very low", "low","medium","high","very high"))
alcohol$Dalc <-factor(alcohol$Dalc, labels = c("very low", "low","medium","high","very high"))
alcohol$Walc <-factor(alcohol$Walc, labels = c("very low", "low","medium","high","very high"))
alcohol$health <-factor(alcohol$health, labels = c("very bad", "bad","fair","good","very good"))
alcohol$G1<-discretize(alcohol$G1, method="interval", breaks=5)
alcohol$G2<-discretize(alcohol$G2, method="interval", breaks=5)
alcohol$G3<-discretize(alcohol$G3, method="interval", breaks=5)
alcohol$absences<-discretize(alcohol$absences, method="fixed", breaks=c(0, 5, 10, 15, 75))

# Structure and summary of the discretized data
str(alcohol)
summary(alcohol)
```

```
#Run apriori method rules to get rules for low daily alcohol use
alclow<-apriori(alcohol, parameter= list(supp=0.35, conf=0.8, minlen=2),
appearance=list(rhs=c("Dalc=very low", "Dalc=low"), default="lhs"))
alclow

#remove the redundant rules and display the remaining rules
rules.sorted <- sort(alclow, by="lift")
inspect(rules.sorted)
subset.matrix <- is.subset(rules.sorted, rules.sorted)
subset.matrix[lower.tri(subset.matrix, diag=T)] <- F
redundant <- colSums(subset.matrix, na.rm=T) >= 1
which(redundant)
rules.pruned <- rules.sorted[!redundant]
inspect(rules.pruned)
summary(rules.pruned)

# preview the top 5 rules by lift
inspect(head(sort(rules.pruned, by="lift")),n=5)

# reduce the number of rules by changing the min-supp to 0.3
alclow<-apriori(alcohol, parameter= list(supp=0.35, conf=0.8, minlen=2),
appearance=list(rhs=c("Dalc=very low", "Dalc=low"), default="lhs"))
```

```
alclow


#Graph the data

plot(rules.pruned, method="paracoord", control=list(reorder=TRUE))



#Run apriori method rules to get rules for high daily alcohol use

alchigh<-apriori(alcohol, parameter= list(supp=0.05, conf=0.1, minlen=2),
appearance=list(rhs=c("Dalc=very high", "Dalc=high","Dalc=medium"),
default="lhs"))

alchigh



# view the 3 generated rules by lift

inspect(head(sort(rules.pruned1, by="lift")))




#Graph the data

plot(rules.pruned1, method="paracoord", control=list(reorder=TRUE))


# End Script
```