DATA MINING - Milestone 4 (Interpretation of data & Communication of Insights of data)

Student Alcohol Consumption

Sivanesan Pillai

May 20, 2020

1. Project Introduction

This is a presentation on student alcohol consumption from two group of high school students. This data are obtained from This link here and this data are divided into two parts:

- Math class Students
- Portuguese class students

Research Goals & Objective : To predict the variables that lead to alcohol consumption for these two group of students

2. Data Variables

There are 33 columns. It has been describe in the link, but we'll just paste it here for easier reference: - 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) These grades are related with the course subject, Math or Portuguese: - 1. G1 - first period grade (numeric: from 0 to 20) - 2. G2 - second period grade (numeric: from 0 to 20) - 3. G3 - final grade (numeric: from 0 to 20, output target)

- Additional note: there are several students that belong to both datasets . These students can be identified by searching for identical attributes that characterize each student, as shown in the annexed R file.

3. Libraries

4. Loading data and understanding the dimensionality

```
math_stu <- read.csv("C:/Users/Siva/Documents/Mine/UM/WQD7005/Milestone 4/Student/student-mat.csv")
paste(c("Dimension Math students :", dim(math_stu)), collapse = " ")
```

```
## [1] "Dimension Math students : 395 33"
```

```
port_stu <- read.csv("C:/Users/siva/Documents/Mine/UM/WQD7005/Milestone 4/Student/student-por.csv")
paste(c("Dimension Portuguese students :", dim(port_stu)), collapse = " ")
```

```
## [1] "Dimension Portuguese students : 649 33"
```

5. Looking at the head for each datasets

```
print(head(math_stu))
```

```
##   school sex age address famsize Pstatus Medu Fedu     Mjob     Fjob     reason
## 1     GP   F  18       U     GT3       A    4    4  at_home  teacher     course
## 2     GP   F  17       U     GT3       T    1    1  at_home    other     course
## 3     GP   F  15       U     LE3       T    1    1  at_home    other      other
## 4     GP   F  15       U     GT3       T    4    2   health services       home
## 5     GP   F  16       U     GT3       T    3    3    other    other       home
## 6     GP   M  16       U     LE3       T    4    3 services    other reputation
##   guardian traveltime studytime failures schoolsup famsup paid activities
## 1   mother          2         2        0       yes     no   no         no
## 2   father          1         2        0        no    yes   no         no
## 3   mother          1         2        3       yes     no  yes         no
## 4   mother          1         3        0        no    yes  yes        yes
## 5   father          1         2        0        no    yes  yes         no
## 6   mother          1         2        0        no    yes  yes        yes
##   nursery higher internet romantic famrel freetime goout Dalc Walc health
## 1     yes    yes       no       no      4        3     4    1    1      3
## 2      no    yes      yes       no      5        3     3    1    1      3
## 3     yes    yes      yes       no      4        3     2    2    3      3
## 4     yes    yes      yes      yes      3        2     2    1    1      5
## 5     yes    yes       no       no      4        3     2    1    2      5
## 6     yes    yes      yes       no      5        4     2    1    2      5
##   absences G1 G2 G3
## 1        6  5  6  6
## 2        4  5  5  6
## 3       10  7  8 10
## 4        2 15 14 15
## 5        4  6 10 10
## 6       10 15 15 15
```

```
print(head(port_stu))
```

```
##   school sex age address famsize Pstatus Medu Fedu     Mjob     Fjob     reason
## 1     GP   F  18       U     GT3       A    4    4  at_home  teacher     course
## 2     GP   F  17       U     GT3       T    1    1  at_home    other     course
## 3     GP   F  15       U     LE3       T    1    1  at_home    other      other
## 4     GP   F  15       U     GT3       T    4    2   health services       home
## 5     GP   F  16       U     GT3       T    3    3    other    other       home
## 6     GP   M  16       U     LE3       T    4    3 services    other reputation
##   guardian traveltime studytime failures schoolsup famsup paid activities
## 1   mother          2         2        0       yes     no   no         no
## 2   father          1         2        0        no    yes   no         no
## 3   mother          1         2        0       yes     no   no         no
## 4   mother          1         3        0        no    yes   no        yes
## 5   father          1         2        0        no    yes   no         no
## 6   mother          1         2        0        no    yes   no        yes
##   nursery higher internet romantic famrel freetime goout Dalc Walc health
## 1     yes    yes       no       no      4        3     4    1    1      3
## 2      no    yes      yes       no      5        3     3    1    1      3
## 3     yes    yes      yes       no      4        3     2    2    3      3
## 4     yes    yes      yes      yes      3        2     2    1    1      5
## 5     yes    yes       no       no      4        3     2    1    2      5
## 6     yes    yes      yes       no      5        4     2    1    2      5
##   absences G1 G2 G3
## 1        4  0 11 11
## 2        2  9 11 11
## 3        6 12 13 12
## 4        0 14 14 14
## 5        0 11 13 13
## 6        6 12 12 13
```

6. Labeling for easier tracking

```
math_stu$subject <- "math"
port_stu$subject <- "port"
```

7. Merging datasets

As mentioned in the data desciption, there are some students belong to both group, let us combine and see the total students data.

```
math_port_stu <- rbind(math_stu,port_stu)
nrow_x <- nrow(math_port_stu)#subtotal off mathematic and portuguse datasets

colnames(math_port_stu) <- tolower(colnames(math_port_stu)) #define all column names to lowercase

non_dup_col <- c("school","sex","age","address","famsize","pstatus","medu","fedu","mjob","fjob","reason

dup_row <- duplicated(math_port_stu[,non_dup_col]) #get the duplicated rows
```

```
dup_math_port_stu <- math_port_stu[dup_row,] #duplicated students
```

7. Remove duplicate data values

Using function call “duplicated”, to merged the both mathematics and portugese student’s subject column to “math_port” into temporary datasets, and delete the duplicated data after it is merged.

```
comb_subj <- function(duplicated){

  new_rows <- math_port_stu$school == duplicated$school &
        math_port_stu$sex == duplicated$sex &
        math_port_stu$age == duplicated$age &
        math_port_stu$address == duplicated$address &
        math_port_stu$famsize == duplicated$famsize &
        math_port_stu$pstatus == duplicated$pstatus &
        math_port_stu$medu == duplicated$medu &
        math_port_stu$fedu == duplicated$fedu &
        math_port_stu$mjob == duplicated$mjob &
        math_port_stu$reason == duplicated$reason &
        math_port_stu$nursery == duplicated$nursery &
        math_port_stu$internet == duplicated$internet

        math_port_stu[new_rows,"subject"] <<- "math_port"
}

for(n in 1:nrow(dup_math_port_stu)){
    comb_subj(dup_math_port_stu[n,])
}

dup_math_port_stu <- dup_math_port_stu[!dup_row,]

math_port_stu$subject <- as.factor(math_port_stu$subject)

nrow_y <-nrow(math_port_stu) #subtotal off merged mathematic and portuguse datasets

non_dup_math_port_stu <- nrow_x - nrow_y # number(subtotal of mathematic and portuguse datasets) - numb
```

7. Exploring the data

Starting here we are going to explore the whole data

```
dim(math_port_stu) #dimension
```

```
## [1] 1044   34
```

8. Missing Values

As we can see from below all the columns doesnt have any missing value, except for the column that is class related.

```
nrow(math_port_stu) - sum(complete.cases(math_port_stu)) #finding missing values
```

```
## [1] 0
```

9. Correlation

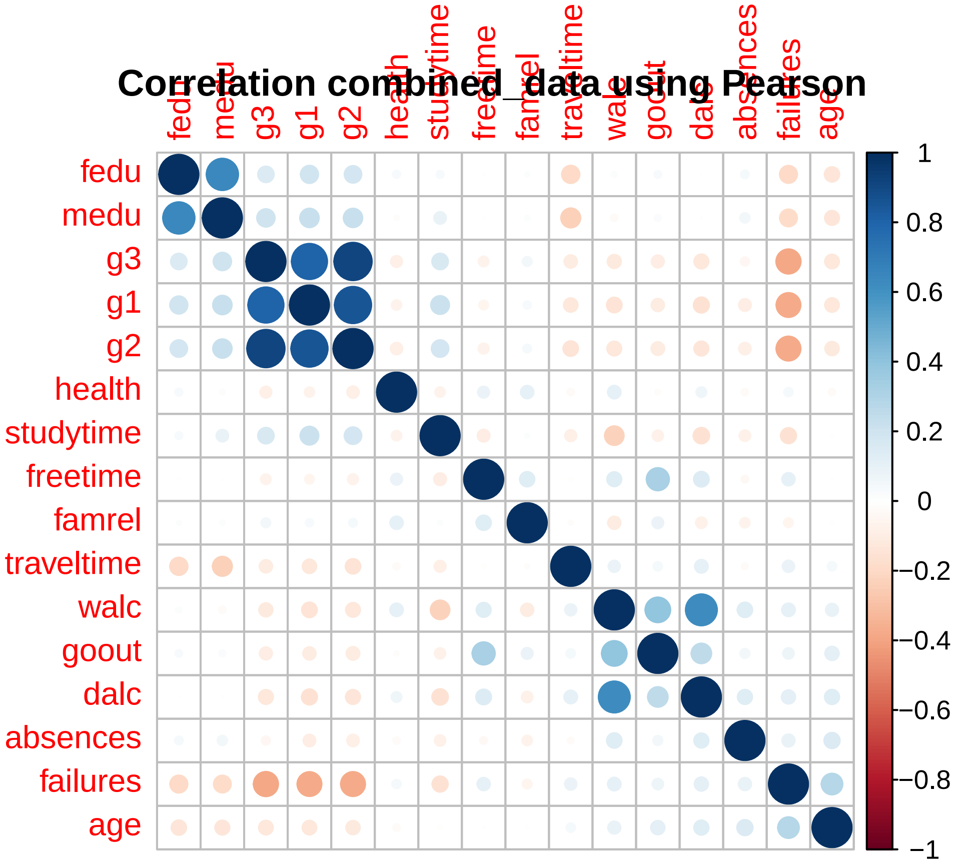
As we can see from below all the columns doesnt have any missing value, except for the column that is class related.

To understand further let's visualize the total correlation table to understand all the numeric variable relationship with each other

```
cor <- cor((math_port_stu %>% keep(is.numeric)), use="pairwise", method="pearson")

# Order the correlations by their strength.
ord <- order(cor[1,])
cor <- cor[ord, ord]
print(cor)
```

```
##                     fedu         medu          g3          g1          g2
## fedu        1.0000000000  0.642063146  0.15979605  0.19589802  0.18263396
## medu        0.6420631457  1.000000000  0.20147169  0.22610060  0.22466175
## g3          0.1597960494  0.201471690  1.00000000  0.80914172  0.91074316
## g1          0.1958980209  0.226100602  0.80914172  1.00000000  0.85873875
## g2          0.1826339619  0.224661748  0.91074316  0.85873875  1.00000000
## health      0.0342882377 -0.013254090 -0.08007864 -0.06047794 -0.08800109
## studytime   0.0334578745  0.090616377  0.16162894  0.21131391  0.18316670
## freetime    0.0021417298  0.001054219 -0.06488968 -0.05198471 -0.06895189
## famrel      0.0130659150  0.015003618  0.05446106  0.03694727  0.04205362
## traveltime -0.1963281605 -0.238180728 -0.10262712 -0.12105330 -0.14016297
## walc        0.0195239342 -0.029330541 -0.11574000 -0.14240140 -0.12811435
## goout       0.0300748764  0.025614278 -0.09787726 -0.10116347 -0.10841089
## dalc       -0.0001648393  0.001515097 -0.12964212 -0.15094254 -0.13157648
## absences    0.0408288855  0.059707676 -0.04567058 -0.09242463 -0.08933169
## failures   -0.1913904210 -0.187769404 -0.38314528 -0.37417487 -0.37717218
## age        -0.1385207614 -0.130196115 -0.12528243 -0.12412125 -0.11947474
##                 health    studytime     freetime       famrel   traveltime
## fedu        0.03428824  0.033457874  0.002141730  0.013065915 -0.196328161
## medu       -0.01325409  0.090616377  0.001054219  0.015003618 -0.238180728
## g3         -0.08007864  0.161628935 -0.064889679  0.054461059 -0.102627118
## g1         -0.06047794  0.211313915 -0.051984712  0.036947274 -0.121053301
## g2         -0.08800109  0.183166702 -0.068951886  0.042053621 -0.140162973
## health      1.00000000 -0.063044459  0.081517225  0.104100776 -0.029001978
## studytime  -0.06304446  1.000000000 -0.094429345  0.012324093 -0.081328016
## freetime    0.08151722 -0.094429345  1.000000000  0.136900650 -0.007402578
## famrel      0.10410078  0.012324093  0.136900650  1.000000000 -0.012577522
## traveltime -0.02900198 -0.081328016 -0.007402578 -0.012577522  1.000000000
## walc        0.10666944 -0.229073148  0.130377028 -0.100663375  0.084292404
## goout      -0.01373623 -0.072940739  0.323555753  0.080619212  0.049739783
## dalc        0.06551534 -0.159664641  0.144979128 -0.076482657  0.109423016
## absences   -0.02747860 -0.075593669 -0.032078736 -0.062170662 -0.022668699
## failures    0.04831102 -0.152023523  0.102678757 -0.053676457  0.087177495
## age        -0.02912927 -0.007870098  0.002645147  0.007161921  0.049215707
##                   walc       goout          dalc    absences    failures
## fedu        0.01952393  0.03007488 -0.0001648393  0.04082889 -0.19139042
## medu       -0.02933054  0.02561428  0.0015150967  0.05970768 -0.18776940
## g3         -0.11574000 -0.09787726 -0.1296421248 -0.04567058 -0.38314528
## g1         -0.14240140 -0.10116347 -0.1509425374 -0.09242463 -0.37417487
## g2         -0.12811435 -0.10841089 -0.1315764840 -0.08933169 -0.37717218
## health      0.10666944 -0.01373623  0.0655153422 -0.02747860  0.04831102
## studytime  -0.22907315 -0.07294074 -0.1596646413 -0.07559367 -0.15202352
## freetime    0.13037703  0.32355575  0.1449791279 -0.03207874  0.10267876
## famrel     -0.10066338  0.08061921 -0.0764826572 -0.06217066 -0.05367646
## traveltime  0.08429240  0.04973978  0.1094230162 -0.02266870  0.08717749
## walc        1.00000000  0.39979373  0.6278138380  0.13970313  0.10743159
## goout       0.39979373  1.00000000  0.2531348291  0.05614214  0.07468331
## dalc        0.62781384  0.25313483  1.0000000000  0.13286713  0.11633579
## absences    0.13970313  0.05614214  0.1328671345  1.00000000  0.09999785
## failures    0.10743159  0.07468331  0.1163357901  0.09999785  1.00000000
## age         0.09829141  0.11851012  0.1334529897  0.15319565  0.28236357
##                     age
## fedu       -0.138520761
## medu       -0.130196115
## g3         -0.125282433
## g1         -0.124121249
## g2         -0.119474744
## health     -0.029129265
## studytime  -0.007870098
## freetime    0.002645147
## famrel      0.007161921
## traveltime  0.049215707
## walc        0.098291406
## goout       0.118510124
## dalc        0.133452990
## absences    0.153195647
## failures    0.282363566
## age         1.000000000
```

```
corrplot(cor, mar=c(0,0,0,0))
title(main="Correlation combined_data using Pearson")
```



From the correlation plot, we can see that:

1) The grades are highly correlated to each other. I think it is safe to say that we can take the average to represent all of the variable.
2) Failures also is correalted to the grades. If we want to try to predict the grades, this failure also can also be as a target since it is highly correlated to the grades. Hence we can remove it for model building.
3) Daily alcohol consumption and weekend alcohol consumption also is highly correlated. Hence, we will take the total of both of these column to represent the total weekly alcohol consumption.

10. Sample Paired T-Test

```
t_test <- t.test(math_port_stu$walc,math_port_stu$dalc,paired = TRUE)
t_test
```

```
##
##  Paired t-test
##
## data:  math_port_stu$walc and math_port_stu$dalc
## t = 25.387, df = 1043, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.7291499 0.8513098
## sample estimates:
## mean of the differences
##               0.7902299
```

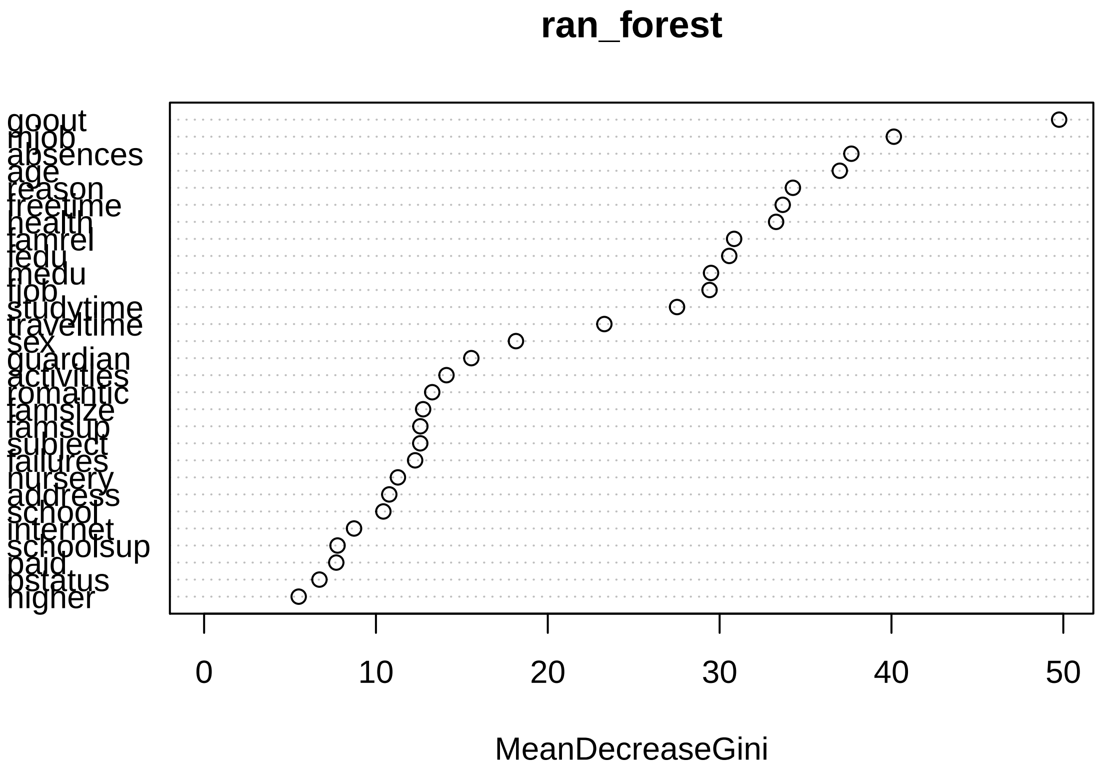
11. Machine Learning

We modelled weekend alcohol consumption (variable walc) and daily alcohol consumption (variable dalc) from the data using Random Forest

```
set.seed(3006) #Simulating random number

math_port_stu_new <- sample(1:nrow(math_port_stu),size = ceiling(0.8*nrow(math_port_stu)),replace = FALS

# i.weekend alcohol consumption (variable walc) - RAINFOREST
math_port_stu_a_train <- math_port_stu[math_port_stu_new,]
math_port_stu_a_test <- math_port_stu[-math_port_stu_new,]

math_port_stu_a_train$walc <- as.factor(math_port_stu_a_train$walc)
math_port_stu_a_test$walc <- as.factor(math_port_stu_a_test$walc)

mtry <- sqrt(ncol(math_port_stu_a_test))
ntree <- 1000
ran_forest <- randomForest(walc~.,data = math_port_stu_a_train[,-c(27,31:33)],mtry=mtry,ntree=ntree)

#Inference
math_port_stu_a_rf <- predict(ran_forest,math_port_stu_a_test)
confMatri_a <- confusionMatrix(math_port_stu_a_rf,math_port_stu_a_test$walc)
print(confMatri_a)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  1  2  3  4  5
##          1 67 15  7  9  0
##          2  4 27  1  0  1
##          3  2  2 32  1  5
##          4  4  2  1 17  4
##          5  0  0  0  0  7
##
## Overall Statistics
##
##                Accuracy : 0.7212
##                  95% CI : (0.6549, 0.7809)
##     No Information Rate : 0.3702
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.6183
##
##  Mcnemar's Test P-Value : NA
##
## Statistics by Class:
##
##                      Class: 1 Class: 2 Class: 3 Class: 4 Class: 5
## Sensitivity            0.8701   0.5870   0.7805  0.62963  0.41176
## Specificity            0.7634   0.9630   0.9401  0.93923  1.00000
## Pos Pred Value         0.6837   0.8182   0.7619  0.60714  1.00000
## Neg Pred Value         0.9091   0.8914   0.9458  0.94444  0.95025
## Prevalence             0.3702   0.2212   0.1971  0.12981  0.08173
## Detection Rate         0.3221   0.1298   0.1538  0.08173  0.03365
## Detection Prevalence   0.4712   0.1587   0.2019  0.13462  0.03365
## Balanced Accuracy      0.8167   0.7750   0.8603  0.78443  0.70588
```

```
varImpPlot(ran_forest,scale=TRUE)
```

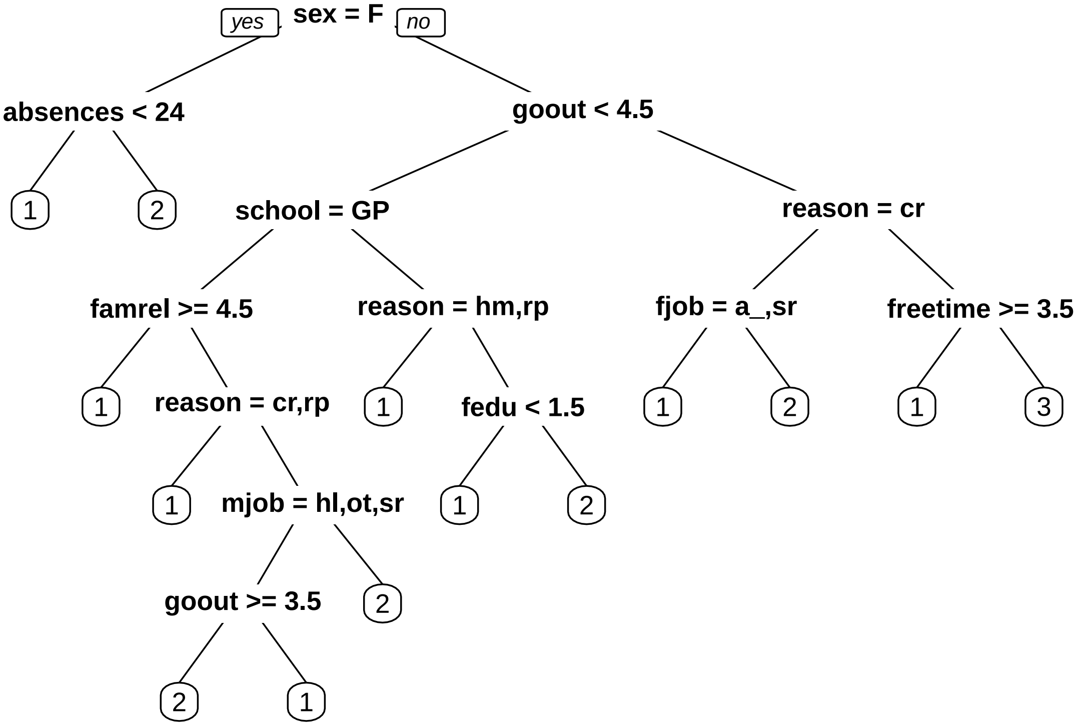


```
# 2.daily alcohol consumption (variable dalc) - DESICION TREE
math_port_stu_b_train <- math_port_stu[math_port_stu_new,]
math_port_stu_b_test <- math_port_stu[-math_port_stu_new,]

math_port_stu_b_train$dalc <- as.factor(math_port_stu_b_train$dalc)
math_port_stu_b_test$dalc <- as.factor(math_port_stu_b_test$dalc)



decision_tree <- rpart(dalc~.,data = math_port_stu_b_train[,-c(28,31:33)])

#Inference
math_port_stu_b_rf <- table(predict(decision_tree,math_port_stu_b_test))
print(math_port_stu_b_rf)
```

```
##
##                   0 0.00429184549356223  0.0135135135135135  0.0150214592274678
##                 105                 114                  17                 114
##  0.0202020202020202  0.0257510729613734  0.0384615384615385  0.0404040404040404
##                  26                 114                   8                  26
##  0.0416666666666667  0.0454545454545455  0.0675675675675676  0.0707070707070707
##                   3                   8                  34                  26
##  0.0769230769230769  0.0909090909090909   0.107142857142857   0.111111111111111
##                   8                   4                   5                   2
##   0.142857142857143   0.143776824034335   0.151515151515152   0.157894736842105
##                  11                 114                  26                   7
##   0.166666666666667   0.181818181818182   0.192307692307692   0.208333333333333
##                  10                   4                   8                   6
##   0.222222222222222   0.285714285714286   0.357142857142857   0.363636363636364
##                   2                  12                   6                   5
##   0.368421052631579   0.428571428571429   0.473684210526316                 0.5
##                   7                   6                   7                   6
##   0.541666666666667   0.636363636363636   0.666666666666667   0.692307692307692
##                   3                   9                   7                   8
##   0.717171717171717                0.75   0.811158798283262   0.851351351351351
##                  26                   5                 114                  17
```

```r
prp(decision_tree,faclen = 2)
```

```
## Warning: Bad 'data' field in model 'call' (expected a data.frame or a matrix).
## To silence this warning:
##     Call prp with roundint=FALSE,
##     or rebuild the rpart model with model=TRUE.
```



12. Conclusion

Through this study I have learned that significant factors that contribute in the indulgence of teenagers in alcoholic activities and affect their academic performance in high schools.

Age, sex and home address was identified as the most significant factors in consuming alcohol by students. At the age of 15-21 they want to prove to their friends and to the society that they are no longer kids.

The home environment is also a primary socialization agent, which affects students' life outside the school, the interest in school as well as the aspirations for the future. Home environment includes parental socio-economic status, parental education background, parental marital status and the quality of family relations, etc. Quality of family relations is one of a key factor.

Students from broken homes suffer psychological effects while in school and this affects their academic performance.

Therefore, family support is a crucial factor in determining the grades and hence, the future of children and should be kept in mind. It is also identified that those who consume high quantity of alcohol have faced more failures in their life than those who consume less quantity.

So, it can be safely concluded that alcohol consumption leads to more failures in life. And therefore, alcohol should be avoided in order to succeed in life and it is high time to handle all these which can be done with the effort of school, family and students themselves