Student Failure Rates and Alcohol Consumption

Raymond Peter David

5/29/2020

"A research into student failures and its potential underlying factors"

Table Of Contents

Introduction

In this project, we will use the data found on kaggle titled "Student Alcohol Consumption" that is created by UCI Machine Learning.This dataset is obtained from a survey of students taking math and portugeese lessons, which we will use to predict and test our hypothesis.There is a combined total of 1044 observation with 34 different variables. We will conduct an eploritary data analysis and create a model to predict the number of student failures in relation to alcohol and other variables present in the dataset.

To get an idea of what each variables meant, you can check the footnotes provided[^1].

1(a).Objectives

The objective of this project is to conduct a research to evaluate student performance in school based on the relationship between daily alchol consumption and student failures. Also, we will try to explore which variables are closely related to each other and also find out which variables have the biggest influence on student failures.

1(b).Hypothesis

*Daily alcohol consumption will not be the most influential determining factor of students failure.

*Parents Cohabitation (living apart) will be a strong factor for students failure.

*Travel time will not influence student failures but reason to go to school will.

1(c).Preprocessing

```
##Install and Load packages
if (!require(tidyverse)) install.packages('tidyverse')
```

```
## Loading required package: tidyverse
```

```
## -- Attaching packages ---------------------------------------------------------------------- tidyv
```

```
## v ggplot2 3.3.0     v purrr   0.3.4
## v tibble  3.0.1     v dplyr   0.8.5
## v tidyr   1.0.3     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.5.0
```

```
## -- Conflicts ------------------------------------------------------------------------ tidyverse_c
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(tidyverse)
if (!require(randomForest)) install.packages('randomForest')
```

```
## Loading required package: randomForest
```

```
## randomForest 4.6-14
```

```
## Type rfNews() to see new features/changes/bug fixes.
```

```
##
## Attaching package: 'randomForest'
```

```
## The following object is masked from 'package:dplyr':
##
##     combine

## The following object is masked from 'package:ggplot2':
##
##     margin
```

```
library(randomForest)
if (!require(kknn)) install.packages('kknn')
```

```
## Loading required package: kknn
```

```
library(kknn)
if (!require(plyr)) install.packages('plyr')
```

```
## Loading required package: plyr

## ------------------------------------------------------------------------------

## You have loaded plyr after dplyr - this is likely to cause problems.
## If you need functions from both plyr and dplyr, please load plyr first, then dplyr:
## library(plyr); library(dplyr)

## ------------------------------------------------------------------------------

##
## Attaching package: 'plyr'

## The following objects are masked from 'package:dplyr':
##
##     arrange, count, desc, failwith, id, mutate, rename, summarise,
##     summarize

## The following object is masked from 'package:purrr':
##
##     compact
```

```
library(plyr) # to combine dataset
if (!require(dplyr)) install.packages('dplyr')
library(dplyr)
if (!require(caret)) install.packages('caret')
```

```
## Loading required package: caret

## Loading required package: lattice

##
## Attaching package: 'caret'
```

```
## The following object is masked from 'package:kknn':
##
##     contr.dummy

## The following object is masked from 'package:purrr':
##
##     lift
```

```
library(caret) # to evaluate model
if (!require(Boruta)) install.packages('Boruta')
```

```
## Loading required package: Boruta
```

```
library(Boruta) #variable selection to improve model
```

```
d1 <- read_csv("~/Downloads/datasets_251_561_student-mat.csv")
```

```
## Parsed with column specification:
## cols(
##   .default = col_character(),
##   age = col_double(),
##   Medu = col_double(),
##   Fedu = col_double(),
##   traveltime = col_double(),
##   studytime = col_double(),
##   failures = col_double(),
##   famrel = col_double(),
##   freetime = col_double(),
##   goout = col_double(),
##   Dalc = col_double(),
##   Walc = col_double(),
##   health = col_double(),
##   absences = col_double(),
##   G1 = col_double(),
##   G2 = col_double(),
##   G3 = col_double()
## )
```

```
## See spec(...) for full column specifications.
```

```
#add new column class Math
d1<-data.frame(Class="Math",d1)
d2 <- read_csv("~/Downloads/datasets_251_561_student-por.csv")
```

```
## Parsed with column specification:
## cols(
##   .default = col_character(),
##   age = col_double(),
##   Medu = col_double(),
##   Fedu = col_double(),
##   traveltime = col_double(),
```

```
##   studytime = col_double(),
##   failures = col_double(),
##   famrel = col_double(),
##   freetime = col_double(),
##   goout = col_double(),
##   Dalc = col_double(),
##   Walc = col_double(),
##   health = col_double(),
##   absences = col_double(),
##   G1 = col_double(),
##   G2 = col_double(),
##   G3 = col_double()
## )
## See spec(...) for full column specifications.
```

```
#add new column class Math
d2<-data.frame(Class="Port",d2)
#BIND
d3<-rbind.fill(d1,d2)
```

Since the number of failures are reported with numbers from 0-4, we will make it simpler by annotaing the number 1 to resemble students that fails and 0 for students that don't fail.

```
#create binary data, FAIL or NOT (dichotomic approach)
d3$failures<-ifelse(d3$failures==0,0,1)%>%
  as.factor
```

1(d).Intial data exploration #Checking for missing values One of the most important step before doing anything else is to check for missing values.If missing values are present, it should be removed. In this dataset, we see that there is no missing values.

```
sum(is.na(d3))
```

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

```
#Check data structure
str(d3)
```

```
## 'data.frame':    1044 obs. of  34 variables:
##  $ Class     : chr  "Math" "Math" "Math" "Math" ...
##  $ school    : chr  "GP" "GP" "GP" "GP" ...
##  $ sex       : chr  "F" "F" "F" "F" ...
##  $ age       : num  18 17 15 15 16 16 16 17 15 15 ...
##  $ address   : chr  "U" "U" "U" "U" ...
##  $ famsize   : chr  "GT3" "GT3" "LE3" "GT3" ...
##  $ Pstatus   : chr  "A" "T" "T" "T" ...
##  $ Medu      : num  4 1 1 4 3 4 2 4 3 3 ...
##  $ Fedu      : num  4 1 1 2 3 3 2 4 2 4 ...
##  $ Mjob      : chr  "at_home" "at_home" "at_home" "health" ...
##  $ Fjob      : chr  "teacher" "other" "other" "services" ...
##  $ reason    : chr  "course" "course" "other" "home" ...
##  $ guardian  : chr  "mother" "father" "mother" "mother" ...
##  $ traveltime: num  2 1 1 1 1 1 1 2 1 1 ...
##  $ studytime : num  2 2 2 3 2 2 2 2 2 2 ...
##  $ failures  : Factor w/ 2 levels "0","1": 1 1 2 1 1 1 1 1 1 1 ...
##  $ schoolsup : chr  "yes" "no" "yes" "no" ...
##  $ famsup    : chr  "no" "yes" "no" "yes" ...
##  $ paid      : chr  "no" "no" "yes" "yes" ...
##  $ activities: chr  "no" "no" "no" "yes" ...
##  $ nursery   : chr  "yes" "no" "yes" "yes" ...
##  $ higher    : chr  "yes" "yes" "yes" "yes" ...
##  $ internet  : chr  "no" "yes" "yes" "yes" ...
##  $ romantic  : chr  "no" "no" "no" "yes" ...
##  $ famrel    : num  4 5 4 3 4 5 4 4 4 5 ...
##  $ freetime  : num  3 3 3 2 3 4 4 1 2 5 ...
##  $ goout     : num  4 3 2 2 2 2 4 4 2 1 ...
##  $ Dalc      : num  1 1 2 1 1 1 1 1 1 1 ...
##  $ Walc      : num  1 1 3 1 2 2 1 1 1 1 ...
##  $ health    : num  3 3 3 5 5 5 3 1 1 5 ...
##  $ absences  : num  6 4 10 2 4 10 0 6 0 0 ...
##  $ G1        : num  5 5 7 15 6 15 12 6 16 14 ...
##  $ G2        : num  6 5 8 14 10 15 12 5 18 15 ...
##  $ G3        : num  6 6 10 15 10 15 11 6 19 15 ...
```

```r
#Check variable names
names(d3)
```

```
##  [1] "Class"      "school"     "sex"        "age"        "address"
##  [6] "famsize"    "Pstatus"    "Medu"       "Fedu"       "Mjob"
## [11] "Fjob"       "reason"     "guardian"   "traveltime" "studytime"
## [16] "failures"   "schoolsup"  "famsup"     "paid"       "activities"
## [21] "nursery"    "higher"     "internet"   "romantic"   "famrel"
## [26] "freetime"   "goout"      "Dalc"       "Walc"       "health"
## [31] "absences"   "G1"         "G2"         "G3"
```

```r
#Top 6 data
head(d3)
```

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

#Method and Analysis

2(a).Brief overview of model used

To reach our objectives, we will tackle 2 different approaches: random forests and K nearest neighbor.

#Random forest To know how random forest works, we should first understand the concept of decision trees. Decision trees are used to create prediction over an event or to test the characteristic that we want. Random forest works by combining the predictions of large numbers of individual decision trees. Generally, the higher the amount of trees, the more accurate the model will be. Although some of the individual trees may be incorrect, there will be more correct trees that still produces good results.To make it simpler, random forest

is used to create a prediction that is more accurate than random guessing[^2].However, it is important to note that we should try to find the best number of trees which will be suitable for our dataset to be accurate.

#K nearest neighbor K nearest neighbor (KNN), predicts value of datapoints according to the assigned value(K).It is generally conducted to compare the resemblance of our predictions to the training data that we decided in advanced[^3].

Based on these two models, we can't really decide on the best or better model. Both the KNN and random forest has their own approach which may work better for one type of activity,say classification, but not as good for another. However, in general, random forest is more widely used due to the time consuming activity of the KNN model.

2(b).RandomForest model

```
set.seed(123,sample.kind = "Rounding")
```

```
## Warning in set.seed(123, sample.kind = "Rounding"): non-uniform 'Rounding'
## sampler used
```

```
# if using R 3.5 or earlier, use `set.seed(123)` instead
rf0 = randomForest(failures ~., # The model
                   data=d3, # The dataset
                   mtry=2) # Hyperparameter mtry
#Which variable make the model became better?
##The bor function will help us decide which variable to use, so that our model is maximised. If the va
set.seed(123,sample.kind = "Rounding")
```

```
## Warning in set.seed(123, sample.kind = "Rounding"): non-uniform 'Rounding'
## sampler used
```

```
bor = bor = Boruta(failures ~., data=d3)
```

```
bor$finalDecision[which(bor$finalDecision=="confirmed")]
```

```
## factor(0)
## Levels: Tentative Confirmed Rejected
```

```
#CONFIRMED VARIABLES
|   Class  |    age   |   Medu   |   Fedu   |   Mjob   | guardian |
|----------|----------|----------|----------|----------|----------|
|confirmed |confirmed |confirmed |confirmed |confirmed |confirmed |
| studytime|  paid    |  higher  |  famrel  |   Dalc   | absences |
|----------|----------|----------|----------|----------|----------|
|confirmed |confirmed |confirmed |confirmed |confirmed |confirmed |
|    G1    |    G2    |    G3    |
|----------|----------|----------|
|confirmed |confirmed |confirmed |
```

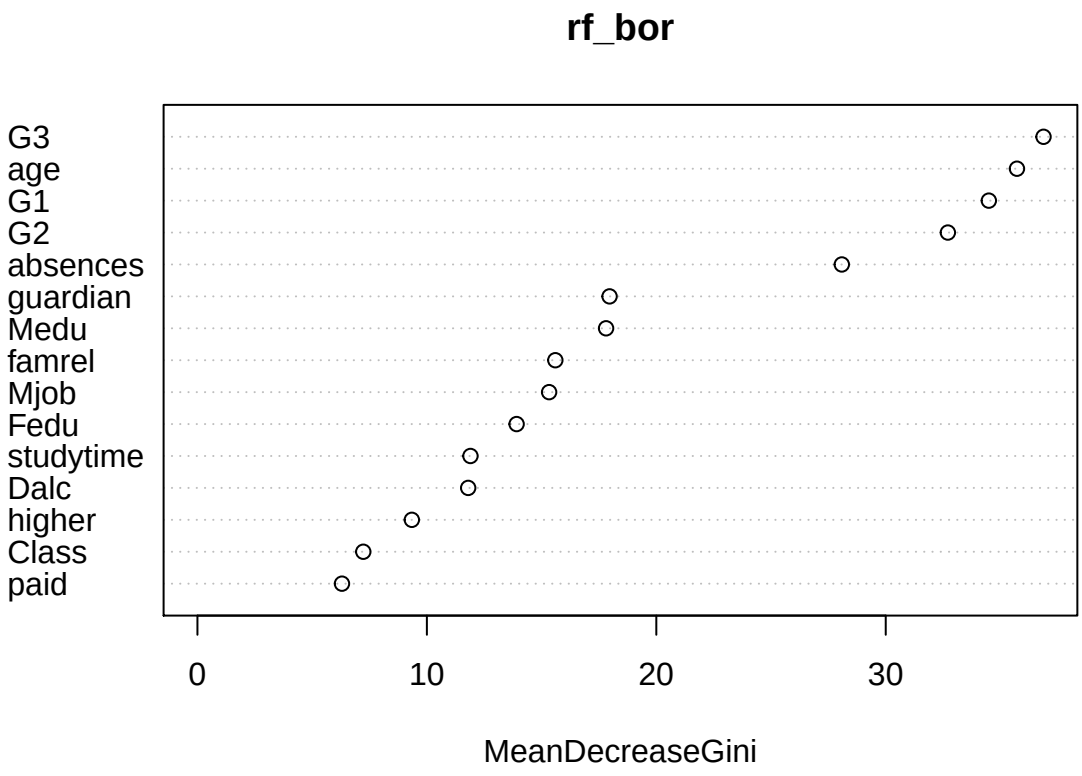
```
#Rejected Variables
|  school  |    sex   |  address | famsize  | Pstatus  |
|----------|----------|----------|----------|----------|
| rejected | rejected | rejected | rejected | rejected |
|  Fjob    |  reason  |traveltime| schoolsup|  famsup  |
|----------|----------|----------|----------|----------|
| rejected | rejected | rejected | rejected | rejected |
|   goout  |    Walc  |  health  |
|----------|----------|----------|
| rejected | rejected | rejected |
```

```r
set.seed(123,sample.kind = "Rounding")
```

```
## Warning in set.seed(123, sample.kind = "Rounding"): non-uniform 'Rounding'
## sampler used
```

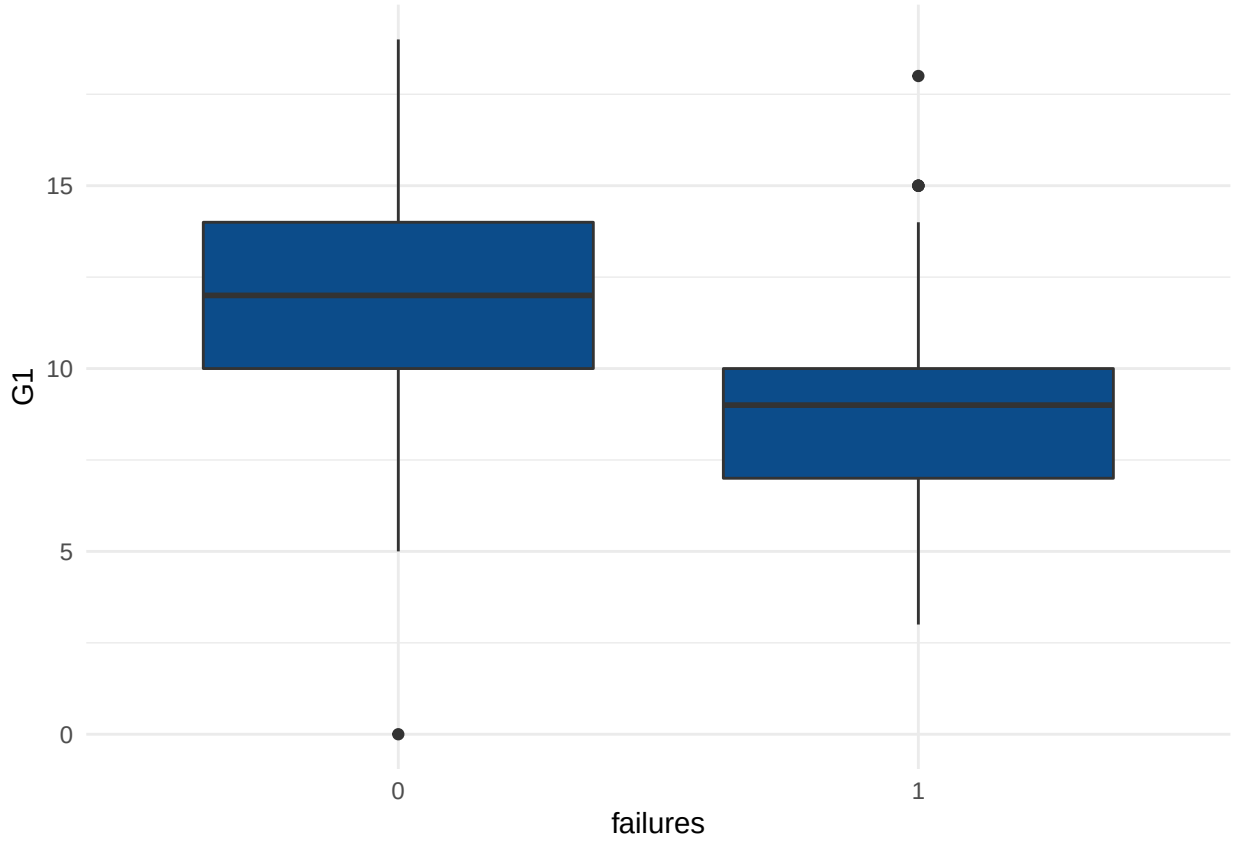
```r
# Using the variables selected by Boruta algorithm
rf_bor <- randomForest(failures ~ Class + age + Medu + Fedu + Mjob + guardian + studytime + paid + high
                       data = d3)
```

```r
# Checking the model
varImpPlot(rf_bor)
```

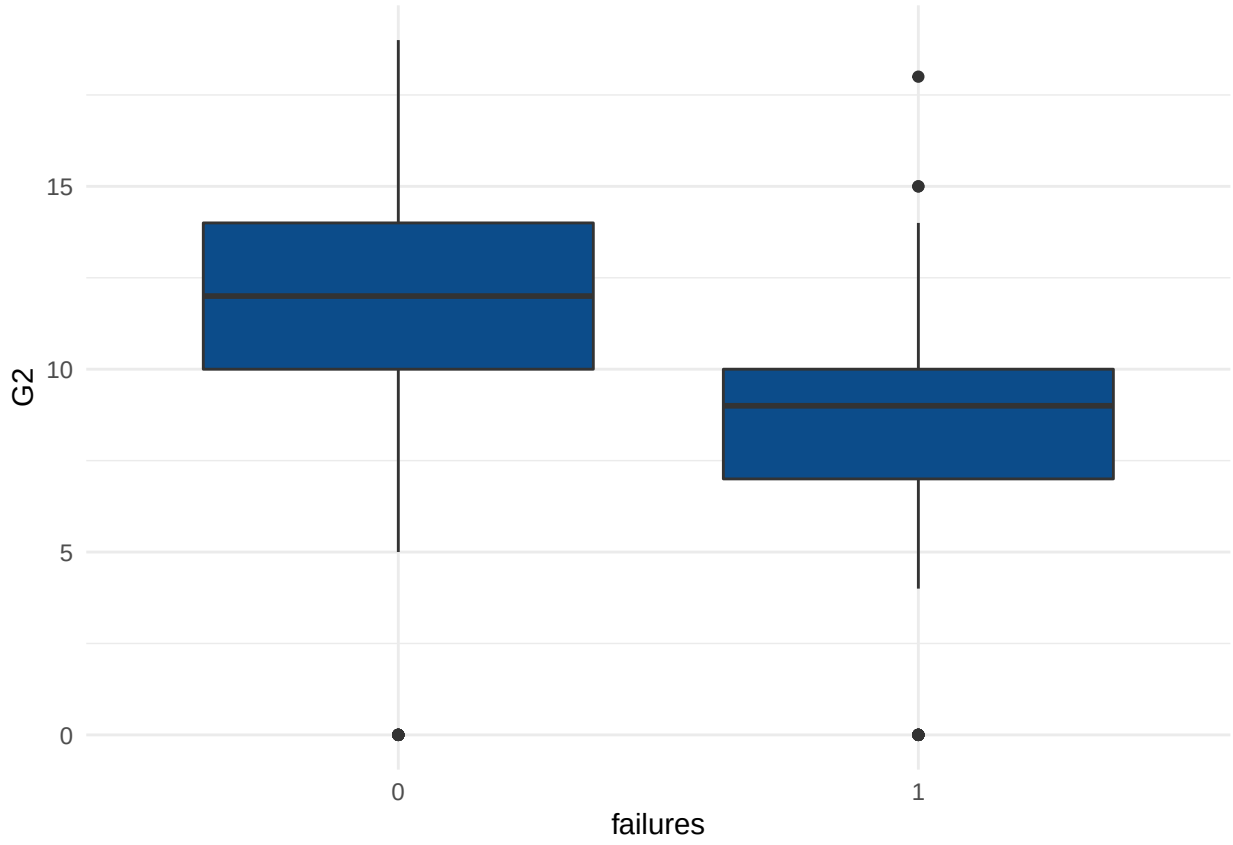


Looking at the plot above, we can see how G3,G1,and G2 are in the top 5 most important variables to determine failures. Lets dig into the data deeper.

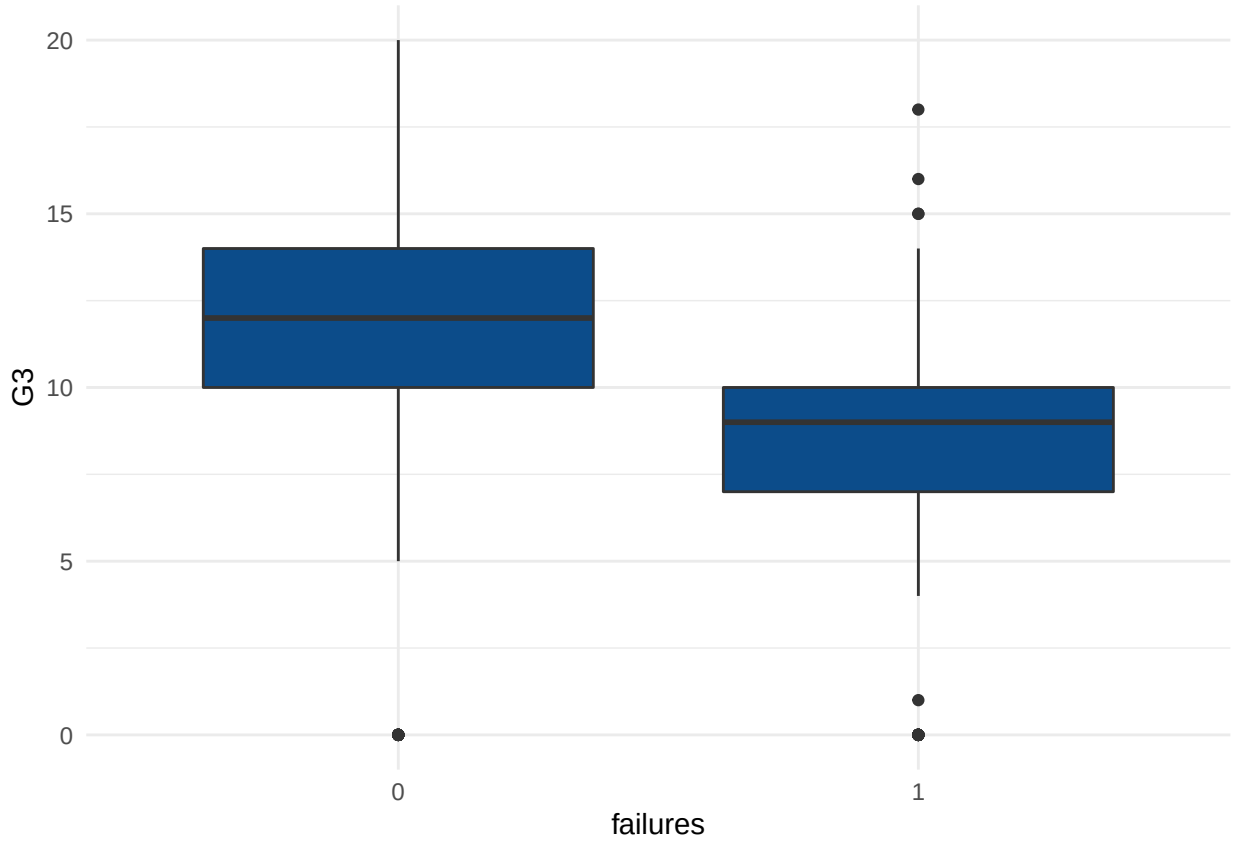
```r
#Boxplot of G1,G2,G3 against failures
ggplot(d3) +
aes(x = failures, y = G1) +
geom_boxplot(fill = "#0c4c8a") +
theme_minimal()
```



```
ggplot(d3) +
aes(x = failures, y = G2) +
geom_boxplot(fill = "#0c4c8a") +
theme_minimal()
```



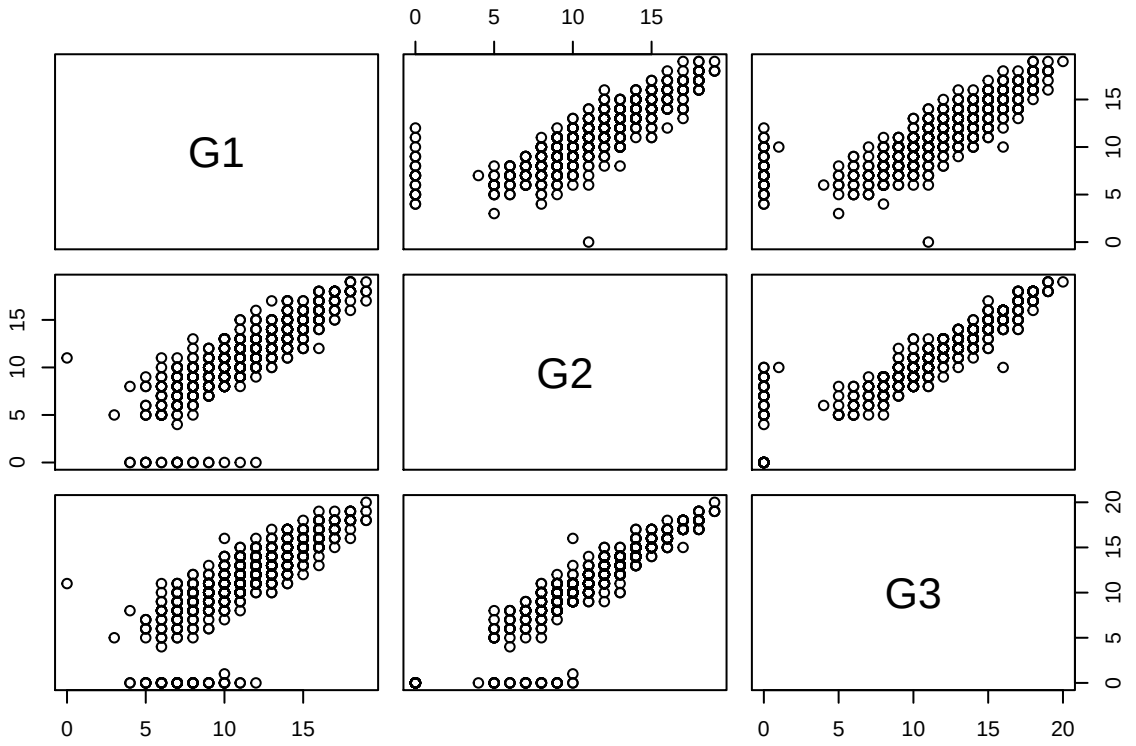
```
ggplot(d3) +
aes(x = failures, y = G3) +
geom_boxplot(fill = "#0c4c8a") +
theme_minimal()
```



G1:first period grade G2:second period grade G3:final grade Looking at these 3 variables and the boxplot, we don't get anything other than the obvious. The higher the score (G1,G2,G3), the lesser the number of failures. The lower the score, the higher number of failures occurs.

Auxiliar Plots

```
# Checking correlation between G1, G2, G3
pairs(d3[,c("G1","G2","G3")])
```



Based on the auxiliar plot, we can see that G1, G2, and G3 are highly correlated.Therefore, it might be a good idea to just combine these variables into one for the sake of clarity and observation.

```
# Joining the 3 variables by average grade
d4 <- mutate(d3, avg_Grade = (d3$G1 + d3$G2 + d3$G3)/3)
d4 <- select(d4, !c(G1,G2,G3))
```

After joining these variables, we will try to see if these changes the accuracy of our model.

```
# In this case, its nothing better...
set.seed(123,sample.kind = "Rounding")
```

```
## Warning in set.seed(123, sample.kind = "Rounding"): non-uniform 'Rounding'
## sampler used
```

```
rf_bor2 <- randomForest(failures ~ Class + age + Medu + Fedu + Mjob + guardian + studytime + paid + high
                        data = d4)
```

```
#checking for error rate:
rf_bor
```

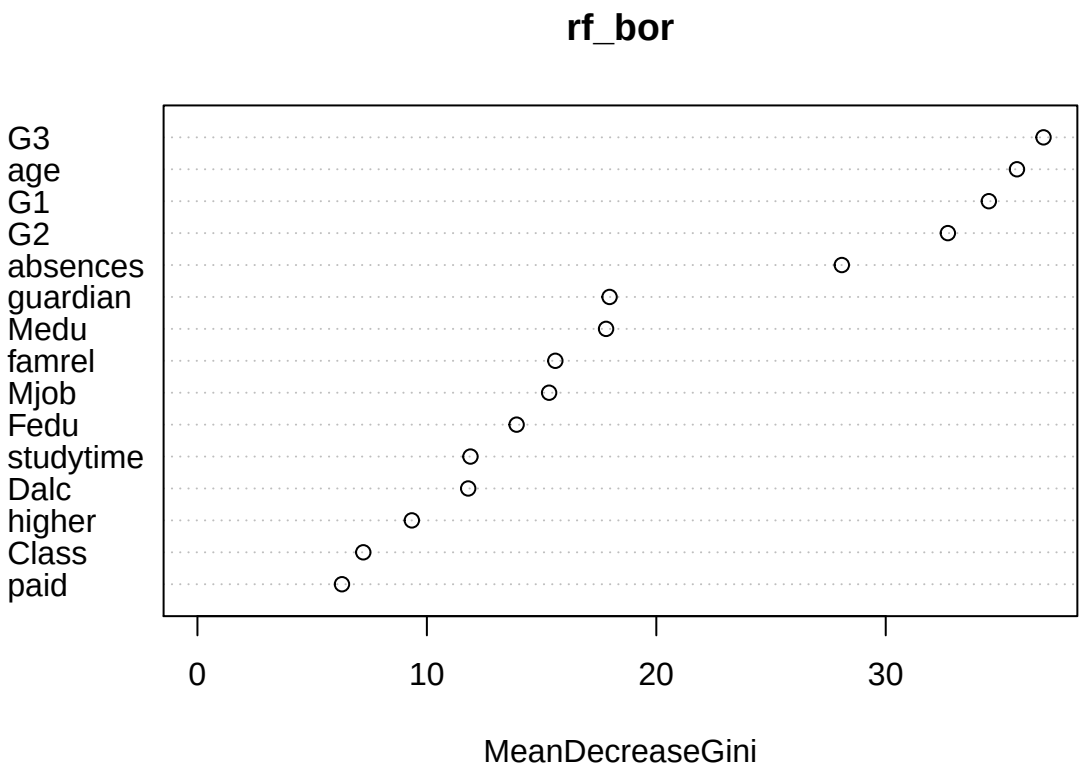
```
##
## Call:
##  randomForest(formula = failures ~ Class + age + Medu + Fedu +      Mjob + guardian + studytime + pa
##                Type of random forest: classification
##                      Number of trees: 500
## No. of variables tried at each split: 3
##
##         OOB estimate of  error rate: 13.89%
## Confusion matrix:
##     0  1 class.error
## 0 824 37  0.04297329
## 1 108 75  0.59016393
```

```
rf_bor2
```

```
##
## Call:
##  randomForest(formula = failures ~ Class + age + Medu + Fedu +      Mjob + guardian + studytime + pa
##                Type of random forest: classification
##                      Number of trees: 500
## No. of variables tried at each split: 3
##
##         OOB estimate of  error rate: 14.46%
## Confusion matrix:
##     0  1 class.error
## 0 821 40  0.04645761
## 1 111 72  0.60655738
```

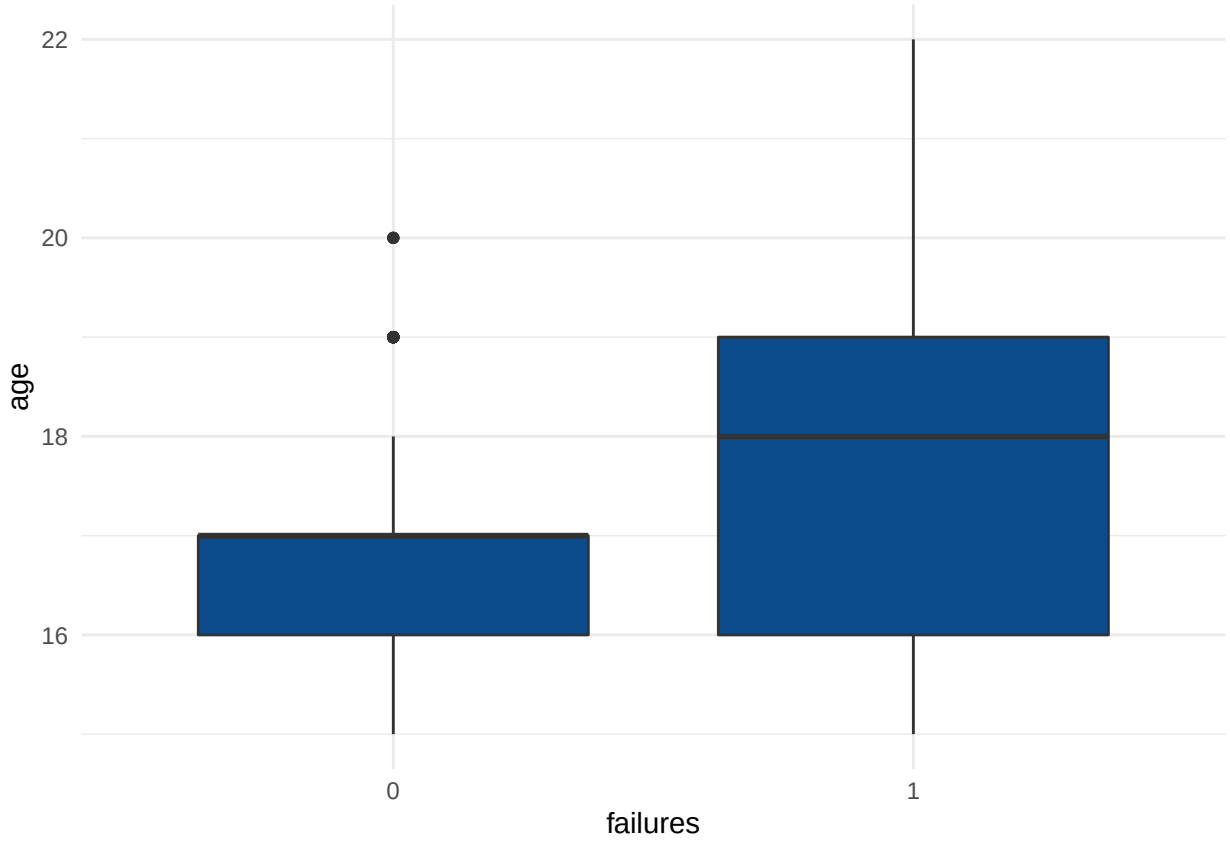
Looking back at the rf_bor plot, we can get more useful insights than just looking at the average grades.

```
varImpPlot(rf_bor)
```



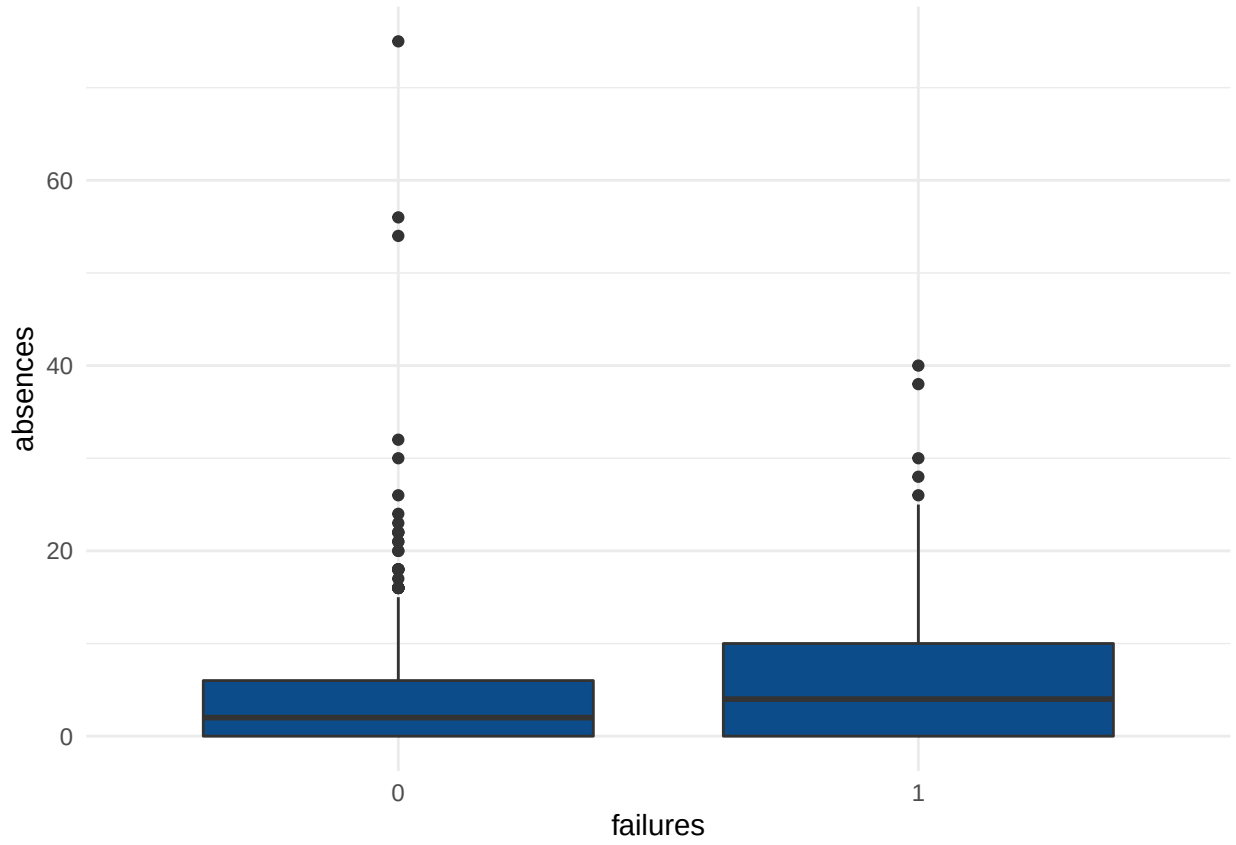
Ignoring G1,G2,and G3, we can see that the best variable that determines failure is age. From this we can deduced that as the age of the students increase, so is the number of failures.This makes sense since as the student grow older, they are more vulnerable to all kinds of different things. But, does this imply that age is more important than alcohol to determine children failure? For example, the students might reached their legal age which will increase the chance of consuming alcohol. And as we can see, daily alcohol intake(Dalc) is also one of the determinant of student failures.However, it is wrong to say that this is all caused by alcohol consumption since as the students grow older, they will more likely to be in a relationship and party more, which may also cause an increase in failure rates. However, this data can be useful for the school since if age is really the most important factor, then the school must apply some strategies to deal with this problem such as adding more extra classes for the older students.

```
ggplot(d3) +
 aes(x = failures, y = age) +
 geom_boxplot(fill = "#0c4c8a") +
 theme_minimal()
```



Next,we can see that absences are on a higher rank than alcohol consumption. The higher the number of absences, the higher the number of failures. This is easily explained because once a student is absent from class, he or she has a lot to catch up and as the number of absents increased, the workload will pile up. Hence, it might cause the students to be behind of his or her classmates and also increase the chances of failing.Therefore, the school must really emphasize on absences and discouraged students from missing school. Parents must also discourage their children from skipping school intentionally without any proper excuse.

```
ggplot(d3) +
 aes(x = failures, y = absences) +
 geom_boxplot(fill = "#0c4c8a") +
 theme_minimal()
```

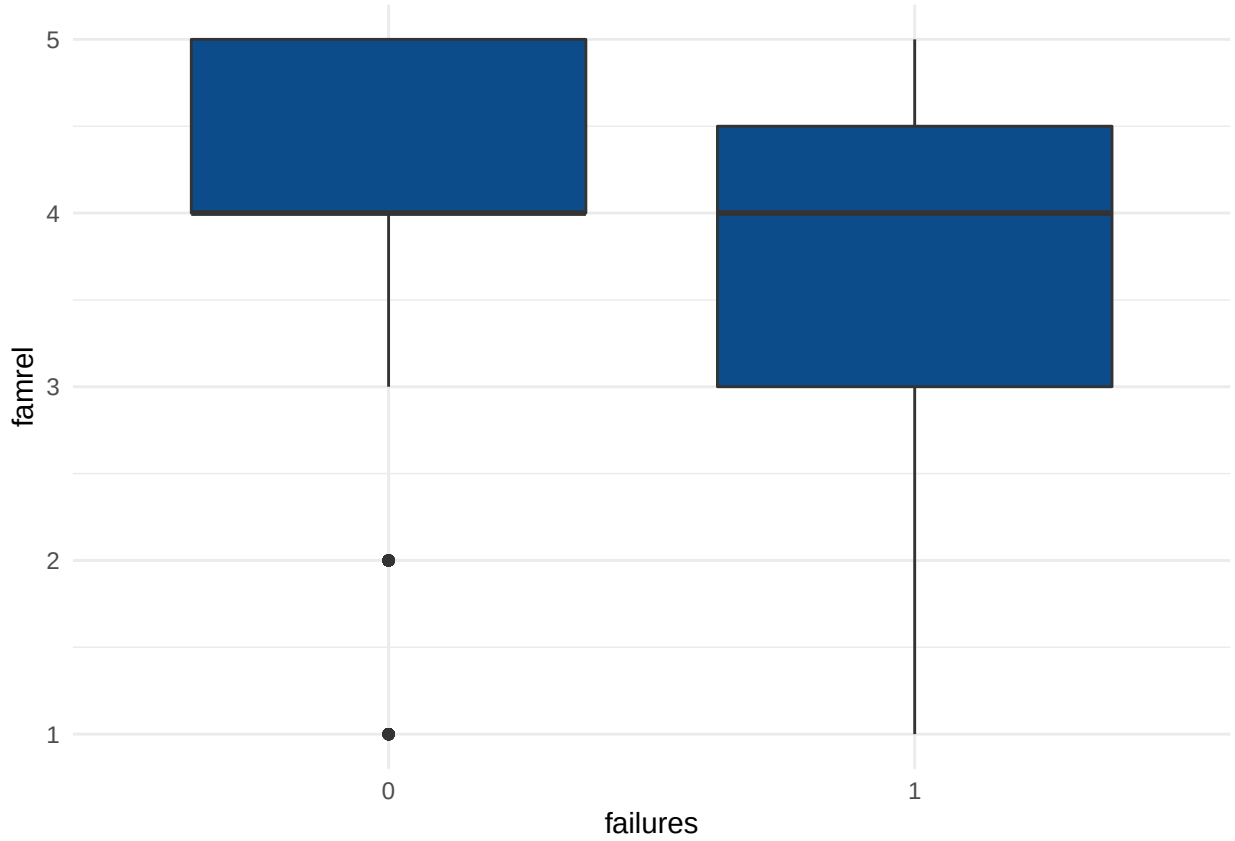


The guardian of the student is also ranked higher than alcohol consumption.When a guardian is present, the student will be supervised and therefore tend to be more obedient to the rules and regulation. Therefore, the students will be more likely to study for an exam and the chances of them to drink alcohol will be lessen by a huge extent. Therefore from this observation, we can't really determine if alcohol really is damaging the students grade.

Another factor that are in the top 5 ignoring G1,G2,G3, is family relationship. This data shows us that a bad family relationship will increase the likelihood of failures. This is an important information since it can be used in fields such as psychology to ensure that the students relationship with his or her family members is good since it might hurt their academic performances. Also, this information tells us that parents must also work hard in maintaining a good relationship with their children so that it won't take a toll on their academic performances.

Surprisingly daily alcohol consumption is the bottom five in the plot. This shows that alcohol does not really influence children failure.Factors such as age and absences plays a bigger role than daily alcohol intake. However,we can't directly eliminate and ignore this potential relationship. Based on the plot, we can see that daily alcohol intake is just above study time. This means that we should also be worried because alcohol's influence on student failures is just as much as study time if not more.However, this is still an open claim and needs further backing from multiple surverys to ensure that our data really backs this. Also, remembering from the 'confirmed' and 'rejected' table shown previously, we can see that weekend alcohol intake is not included in this list because generally people tend to party more and study less on weekends. Therefore, its influence on school failures is too small to take into account.

```
ggplot(d3) +
 aes(x = failures, y = famrel) +
 geom_boxplot(fill = "#0c4c8a") +
 theme_minimal()
```



As shown previously, Pstatus which stand for parents cohabituation is surprisingly not inside the confirmed list. This shows us that factors such as age, absences, and even alcohol has more impact than whether or not the parents are living together or not.In the other hand, paid which stands for extra classes,is inside the 'confirmed' list.Hence, this shows us the importance of having extra classes on the performance of students in school. This is sometimes underestimated since student and even parents tend to look down on the importance of extra classes. Looking at this, students must consider joining extra classes although they might have to spend a little bit of money.

#Evaluating RF Here we will set the training and teating dataset.70% will be on the training set and 30% will be on the testing set.

```
set.seed(123,sample.kind = "Rounding")
```

```
## Warning in set.seed(123, sample.kind = "Rounding"): non-uniform 'Rounding'
## sampler used
```

```
idTrain = createDataPartition(d3$failures,
                              p = 0.7,
                              list = FALSE)

train = d3[idTrain,] # 70%
test = d3[-idTrain,] # 30%
```

Using the same variables selected by Boruta algorithm, we will try to create our final prediction'rf_final'.

```r
set.seed(123,sample.kind = "Rounding")
```

```
## Warning in set.seed(123, sample.kind = "Rounding"): non-uniform 'Rounding'
## sampler used
```

```r
rf_final <- randomForest(failures ~ Class + age + Medu + Fedu + Mjob + guardian + studytime + paid + hig
```

We will then try to evaluate our final model by comparing failures to our predicted value 'rf_final' and test it on our testing set.Also we will then print our evaluation 'rf_eval' to see the summary and performance of the model.

```r
rf_eval <- confusionMatrix(test$failures,
                           predict(rf_final,
                                   test))

rf_eval
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   0   1
##          0 246  12
##          1  33  21
##
##                Accuracy : 0.8558
##                  95% CI : (0.8118, 0.8928)
##     No Information Rate : 0.8942
##     P-Value [Acc > NIR] : 0.986552
##
##                   Kappa : 0.4046
##
##  Mcnemar's Test P-Value : 0.002869
##
##             Sensitivity : 0.8817
##             Specificity : 0.6364
##          Pos Pred Value : 0.9535
##          Neg Pred Value : 0.3889
##              Prevalence : 0.8942
##          Detection Rate : 0.7885
##    Detection Prevalence : 0.8269
##       Balanced Accuracy : 0.7590
##
##        'Positive' Class : 0
##
```

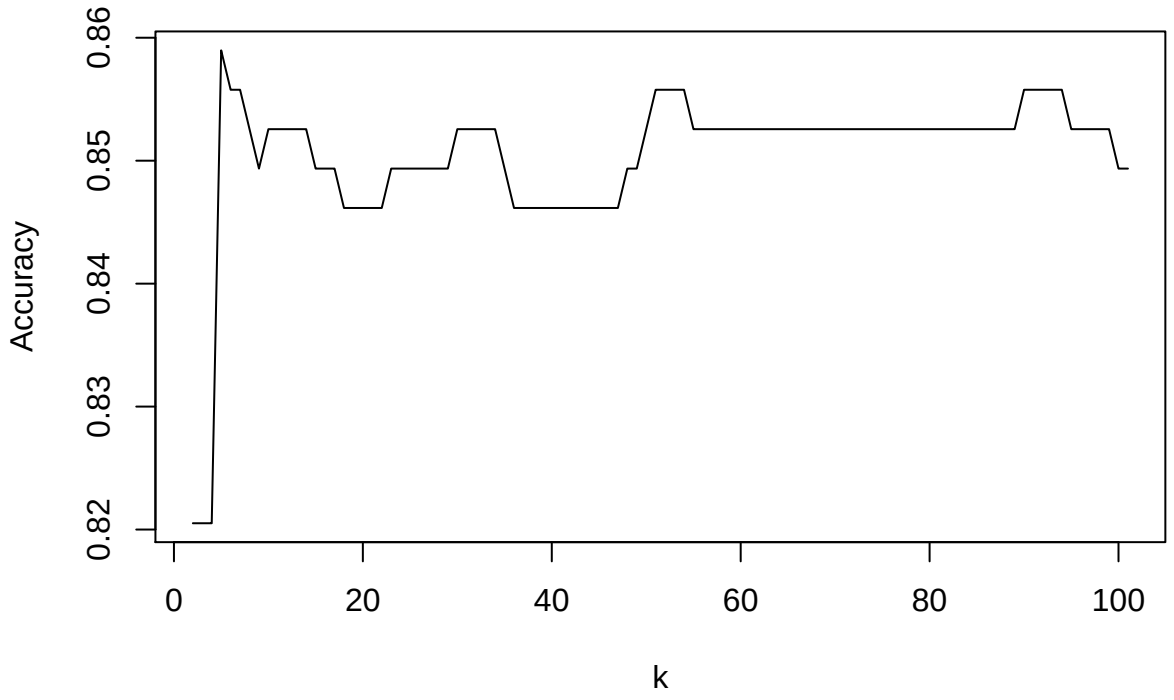
2(c).Knn model #2nd Model:KNN

We will first create a list to compare k and accuracy and use the same variables decided by the boruta algorithm that we also use in the random forest model.

```
results = list(k = rep(0,100), Accur = rep(0,100))

for (i in 2:101){
  my_kknn <- kknn(failures ~ Class + age + Medu + Fedu + Mjob + guardian + studytime + paid + higher + i

  kknn_eval <- confusionMatrix(test$failures, my_kknn$fitted.values)

  results$k[i] = i
  results$Accur[i] = kknn_eval$overall[1]

}
```

Next, we will try to find the value of 'K' by generating a plot that visualizes the distribution of the values of 'K' and find the best value. This is an important step because we always want to improve our model in any way possible.Randomly selecting a number will not produce the most effective model that we can. This is important especially knowing the fact that we can't produce a 100% accurate model, but we can try to maximise it to be as close as possible. Genereally, a lower value of K will lower the error rate, however, it is not the same for validation error. That is why we must pick the best value of k. There are different ways to get the value of k and this is just one of them.Here we will use cross validation where we will try to predict the best value of K from the training dataset. The general formula of K is K=sqrt(N) where N is the number of samples in the training dataset.

```
# K vs Accuracy
plot(results$k[-1], results$Accur[-1], type="l",
     xlab="k", ylab="Accuracy")
```



```
best.K = results$k[which.max(results$Accur)]
```

Next, we will then input the best value of k,which is encoded in 'best.k' to evaluate our final knn model 'best_kknn'.

```
best_kknn <- kknn(failures ~ Class + age + Medu + Fedu + Mjob + guardian + studytime + paid + higher + 

(best_kknn_eval <- confusionMatrix(test$failures, best_kknn$fitted.values))
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   0   1
##          0 250   8
##          1  36  18
##
##                Accuracy : 0.859
##                  95% CI : (0.8153, 0.8956)
##     No Information Rate : 0.9167
##     P-Value [Acc > NIR] : 0.9998
##
##                   Kappa : 0.3803
##
##  Mcnemar's Test P-Value : 4.693e-05
##
##             Sensitivity : 0.8741
##             Specificity : 0.6923
##          Pos Pred Value : 0.9690
##          Neg Pred Value : 0.3333
##              Prevalence : 0.9167
##          Detection Rate : 0.8013
##    Detection Prevalence : 0.8269
##       Balanced Accuracy : 0.7832
##
##        'Positive' Class : 0
##
```

#Results

```
rf_eval #random forest results
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   0   1
##          0 246  12
##          1  33  21
##
##                Accuracy : 0.8558
##                  95% CI : (0.8118, 0.8928)
##     No Information Rate : 0.8942
##     P-Value [Acc > NIR] : 0.986552
```

```
##
##                   Kappa : 0.4046
##
##  Mcnemar's Test P-Value : 0.002869
##
##             Sensitivity : 0.8817
##             Specificity : 0.6364
##          Pos Pred Value : 0.9535
##          Neg Pred Value : 0.3889
##              Prevalence : 0.8942
##          Detection Rate : 0.7885
##    Detection Prevalence : 0.8269
##       Balanced Accuracy : 0.7590
##
##        'Positive' Class : 0
##
```

```
best_kknn_eval #knn results
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   0   1
##          0 250   8
##          1  36  18
##
##                Accuracy : 0.859
##                  95% CI : (0.8153, 0.8956)
##     No Information Rate : 0.9167
##     P-Value [Acc > NIR] : 0.9998
##
##                   Kappa : 0.3803
##
##  Mcnemar's Test P-Value : 4.693e-05
##
##             Sensitivity : 0.8741
##             Specificity : 0.6923
##          Pos Pred Value : 0.9690
##          Neg Pred Value : 0.3333
##              Prevalence : 0.9167
##          Detection Rate : 0.8013
##    Detection Prevalence : 0.8269
##       Balanced Accuracy : 0.7832
##
##        'Positive' Class : 0
##
```

3(a).Random forest results Accuracy : 0.8558
Sensitivity : 0.8817
Specificity : 0.6364

3(b).KNN results Accuracy : 0.859
Sensitivity : 0.8741
Specificity : 0.6923

3(c).Summary of results

Accuracy,sensitivity, and specificity are three important variables to look at when observing binary outcomes. To really see this, we need to first understand the idea of confusion matrix which we conduct with the confusionMatrix formula in R (as shown earlier).To visualize and simpify how the confusion matrix works, below is a table to describe it.

Output	Test Indicator		
		No	Yes
No		True Negative	False Positive
Yes		False Negative	True Positive

The test indicator is the predictions that we make with the model. In this case, 'Yes' means that the students do fail. "No" means that the students pass.The outcome is the actual event or what actually happens. True Negative and True Positive are generally what we want since this means that both the prediction and the actual events matches(Both "No" or Both "Yes"). However, False Positive and False Negative are what we want to avoid since it means that either our test predicts that the student fail when they dont or they predict that the student fail when they pass.Accuracy is a measure of the True Positive and True Negative of the confusion matrix. When the model predicts mostly True Positive and True Negative, then the model will be more accurate.The formula of accuracy is

Accuracy = (TN+TP)/(TN+TP+FN+FP)

As we can see above, the KNN model is slightly better in producing accurate prediction by 0.004 or 0.4%.But, measuring the effectiveness of the model as a whole, it is not enough to conclude based on accuracy only.

Sensitivity is defined as the proportion of positives that is predicted to be positives.In our case, the higher the sensitivity,the less likely for it to predict students failing when they don't.The formula os sensitivity is as follows:

Sensitivity = TP/(TP + FN).

In the other hand, specificity is the proportion of negatives that is predicted as negatives.The higher the sensitivity, the less likely for the result to predict students failing when they actually pass.The formula is given by:

Specificity = TN/(TN + FP)

Although the data shows that KNN is slightly better in accuracy than the random forest model, we can see that the random forest model has a higher percentage of sensitivity. However, Knn has a higher level of specificity. Taking into account all of these factors and acknowledging the purpose of this project, we can decide that the random forest is the better model due to the higher level of sensitivity. Since we are trying to predict the number of failures(Positive), it is better for us to look at sensitivity which measures the level of positives while minimizing the level of false negative.

```
#Conclusion
```

4(a).Summary of findings According to the data and results we get, both the KNN and the randomforest model shows very similar level of accuracy. However, we can agree that the random forest model is better due to the higher level of sensitivity compared to the KNN algorithm. Also, from our predictions we can see that alcohol is not the strongest predictor of students failure. Although it is true that daily alcohol average is somewhat related to students failures, other categories such as age, absences,and family relationship has greater influences on student failures.This is congruent with our hypothesis which predicts that there will be other more important factors than alcohol. However, factors such as age and guardian can undermine the effect of alcohol mainly because as the students grow older, they might start to drink alcohol as they pass their legal age which may not be accurately reflected in our data. Also, the presence of a guardian will reduce the chances of underage drinking or alcohol consumption in general. So, there might be bigger influences of alcohol that we need to further research.

The results shows that parent cohabituation is indeed not a good determinant of students failures.Hence, this proves that the second hypothesis is unsubstantiated. This is an interesting thing since parents status(divorced or together) might seem to be an important factor that might influence the student psychologically. However, it is not backed by the data that we try to analyze. Maybe, parents cohabituation might influence daily alcohol consumption more than it influences student failures, but that is another case of research.

Another thing that we found is that travel time is not related to students failures as expected from our hypothesis. However, our hypothesis is wrong due to the fact that reason to go to school is not that important to student failures like what we expected. A student's reason to go to school is the foundation of their motivation.Lack of reason to go to school will most likely make the students 'lazier' and hence lower their overall performance. However, this is not the case.

4(b).Potential impact

The findings on the report will be beneficial to parents and headmasters since this report can give them an understanding that as students age, they are more likely to receive lower grades. Hence, both parents and headmasters can try to minimize this by adjusting and modifying their methods and plans. For example, the school might want to encourage students to take extra classes since as supported by the data, there is a correlation between extra class and failures too. Also, parents might want to try to figure out why their children's score is diminishing.

The findings might also help in encouraging schools and universities to require a guardian for students below 18,especially when living in other countries alone. This might help them to perform better academically and also maintain a healthier lifestyle free from alcohol.

Parents should also know the importance of their role in their children's academic performance since mothers and fathers education is a pretty strong determinant of student failures. This implies that parents should interfere, if needed, to set up tutors or teach their children when applicable.

This information might also be beneficial for students since it is shown that studytime is indeed important. Therefore, a student must not underestimate the impact of study time to their final output grades. This is somewhat backed by the correlation coefficient between failures and study time which shows a negative relationship.

```
## [1] -0.1497324
```

Also, increasing the final grades (G3) is really important to reduce failures. Hence, the hours spent on studying is really important.

```
## [1] -0.400151
```

4(c).Limitations

#Random forest[^4]

The random forest model is complex due to the number of trees that varies. Also, it reuqires a lot of time to train the dataset.

#KNN[^5]

The limitation of KNN is that it is inefficient because the whole training set is processed for every prediction.

In general, both our model's limitation is that we cant produce a model with 100% accuracy.There will still be a false positives or false negatives.Also, we have to make several models with various levels of training and testing dataset in oder to find the model that generates the best accuracy.

4(d).Future work

After analyzing the results in this model, we can build on the insights gained by trying to find if these relationships and results exists in other similar datasets. We should also work with larger datasets to prove if our model is good enough since the dataset used in this project is not so big. Therefore, further research must be conducted to make our findings and reasonings more established. Some of the research that we can build up on is to investigate why mothers job influences failures, but fathers job does not. Another research might be predicting the number of alcohol consumption based on factors such as family relationship, parents cohabituation and weekend alcohol consumption.

#5.Credits(Citation)

5(a). Dataset owner and conrtibutors Dataset source:

```
Citation:
P. Cortez and A. Silva. Using Data Mining to Predict Secondary School Student Performance. In A. Brito a

Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, C

BiBTeX citation:
@misc{Dua:2019 ,
author = "Dua, Dheeru and Graff, Casey",
year = "2017",
title = "{UCI} Machine Learning Repository",
url = "http://archive.ics.uci.edu/ml",
institution = "University of California, Irvine, School of Information and Computer Sciences" }

MLA Citation:

Cortez, P, and A Silva. "Student Alcohol Consumption." Kaggle, UCI Machine Learning, 19 Oct. 2016, www.k

Kumar, Naresh. "Advantages and Disadvantages of Random Forest Algorithm in Machine Learning." Advantages

MLNerds. "How Does KNN Algorithm Work ? What Are the Advantages and Disadvantages of KNN ?" Ace the Data

Tutorialspoint. "KNN Algorithm - Finding Nearest Neighbors." Tutorialspoint, www.tutorialspoint.com/mach

Yiu, Tony. "Understanding Random Forest." Understanding Random Forest, Towards Data Science, 14 Aug. 201

[^1]:https://www.kaggle.com/uciml/student-alcohol-consumption
[^2]:https://towardsdatascience.com/understanding-random-forest-58381e0602d2
[^3]:https://www.tutorialspoint.com/machine_learning_with_python/machine_learning_with_python_knn_algori
[^4]:http://theprofessionalspoint.blogspot.com/2019/02/advantages-and-disadvantages-of-random.html
[^5]:https://machinelearninginterview.com/topics/machine-learning/how-does-knn-algorithm-work-what-are-t
```

```r
print("Operating System:")
```

```
## [1] "Operating System:"
```

```r
version
```

```
##                _
## platform       x86_64-apple-darwin17.0
## arch           x86_64
## os             darwin17.0
## system         x86_64, darwin17.0
## status
## major          4
## minor          0.0
## year           2020
## month          04
## day            24
## svn rev        78286
## language       R
## version.string R version 4.0.0 (2020-04-24)
## nickname       Arbor Day
```

####END####