Havardx:Capstone CYO

Rumbidzai Pamacheche

11 April 2019

1. *Introduction*

Academic achievement is commonly measured through examinations or continuous assessments but there is no general consensus on how it is best evaluated, or which aspects are most important (Ward et al, 1996). It has been said that academic performance is dependent on external factors: **Academic background factors**, **Behavioral factors**, and **Demographic factors**.

Te data set used in the report is extensive and provides us with 33 factors ranging from Academic, Behavioral and Demographic factors. Cortez and Silva (2008), modeled this data set under binary/five-level classification and regression tasks. An important note is that the target attribute G3 has a strong correlation with attributes G2 and G1. This occurs because G3 is the final year grade (issued at the 3rd period), while G1 and G2 correspond to the 1st and 2nd period grades.

For the prediction of students performance (G3), we will: 1) Analyze how some of the factors in above mentioned categories contribute individually to student's academic performance as well as finding correlation between factors. 2) Building a predictive model using algorithm to predict a student's performance based on these factors.3) Lastly, evaluate performance of the model developed.

2. *Methods*

Summary of data set: There are 395 observations and 33 rows. There are no empty cells in the data set.

```
##   school sex age address famsize Pstatus Medu Fedu     Mjob     Fjob
## 1     GP   F  18       U     GT3       A    4    4  at_home  teacher
## 2     GP   F  17       U     GT3       T    1    1  at_home    other
## 3     GP   F  15       U     LE3       T    1    1  at_home    other
## 4     GP   F  15       U     GT3       T    4    2   health services
## 5     GP   F  16       U     GT3       T    3    3    other    other
## 6     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        3       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
## 1         no     yes    yes       no       no      4        3     4    1
## 2         no      no    yes      yes       no      5        3     3    1
## 3         no     yes    yes      yes       no      4        3     2    2
## 4        yes     yes    yes      yes      yes      3        2     2    1
## 5         no     yes    yes       no       no      4        3     2    1
## 6        yes     yes    yes      yes       no      5        4     2    1
##   Walc health absences G1 G2 G3
## 1    1      3        6  5  6  6
## 2    1      3        4  5  5  6
## 3    3      3       10  7  8 10
## 4    1      5        2 15 14 15
## 5    2      5        4  6 10 10
## 6    2      5       10 15 15 15
```

```
## [1] 395  33
```

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

2.1. *Data Cleaning*

```
## '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 ...
```

```
## [1] FALSE
```

Our data is in tidy format and our categorical variables do not need a change in datatype. Additionally, there are no missing values in the data set.

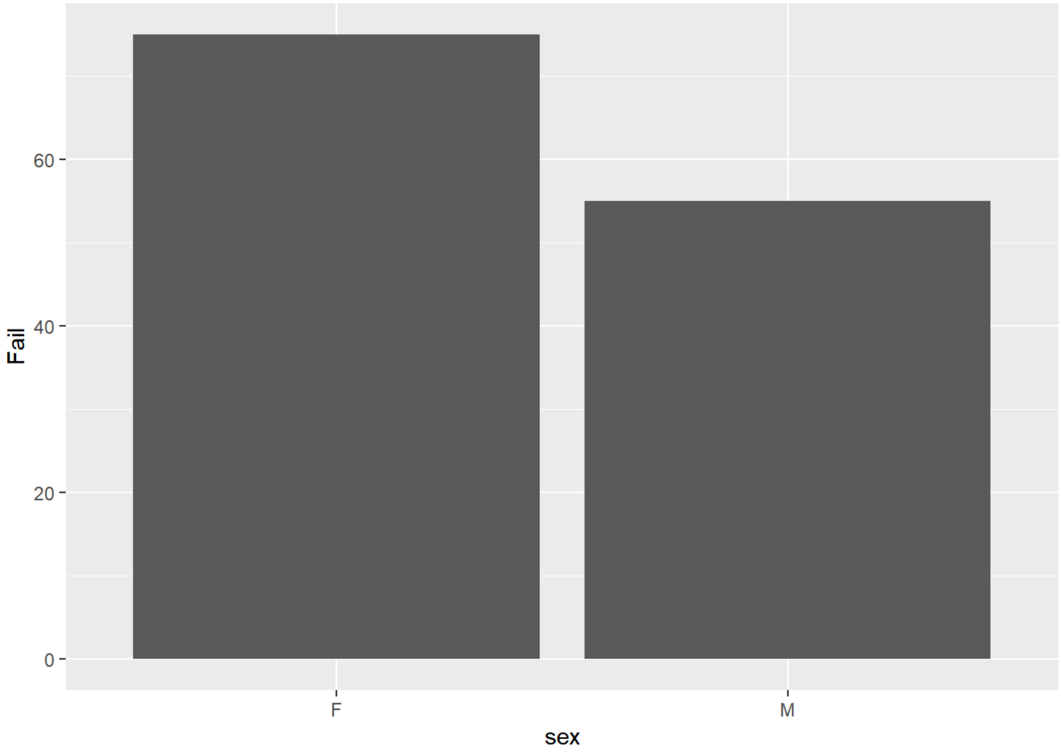
2.2. *Explore and Visualize the data set*

In this section we will explore and visually represent the data set. We will also learn more about the relationships between predictors and student performance.

The number of male vs female that passed and the exam

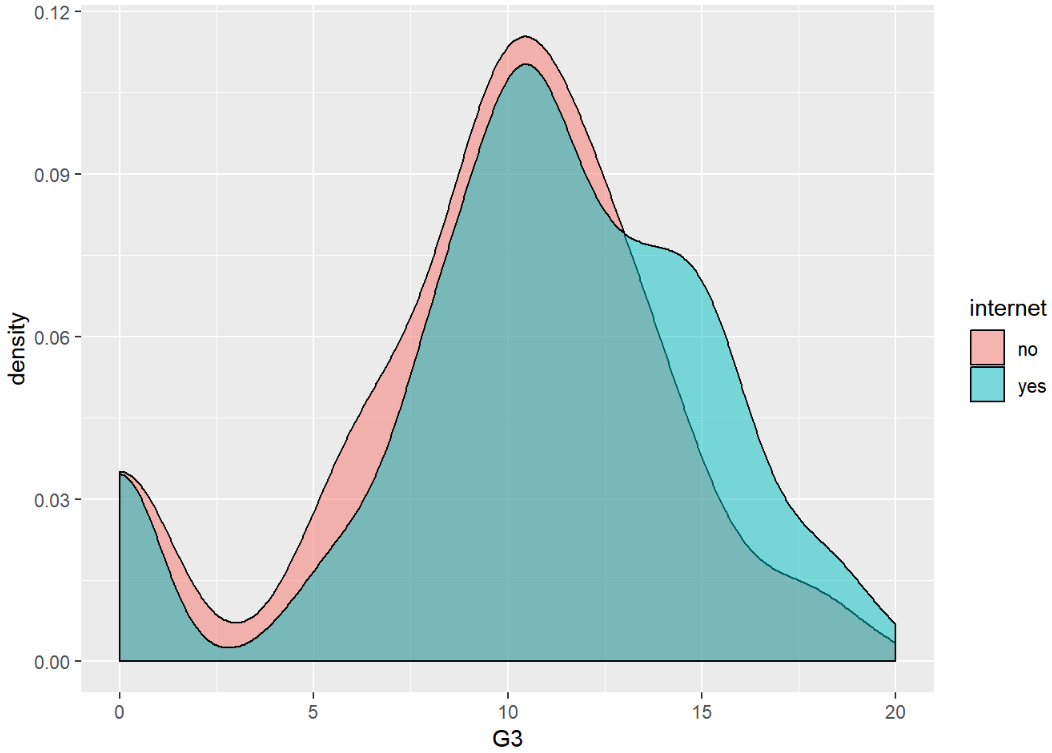
```
math_performance_gender<-student_math_performance%>%
    mutate(pass=ifelse(G3>=10,1,0), fail= ifelse(G3<10,1,0))%>%
    filter(sex=="F"|sex=="M")%>%
    group_by(sex)%>%
    summarise(Pass=sum(pass),
              Fail=sum(fail))

#Graphical representation:
math_performance_gender%>%
  ggplot(aes(x=sex,y=Fail))+
  geom_bar(stat="identity")
```

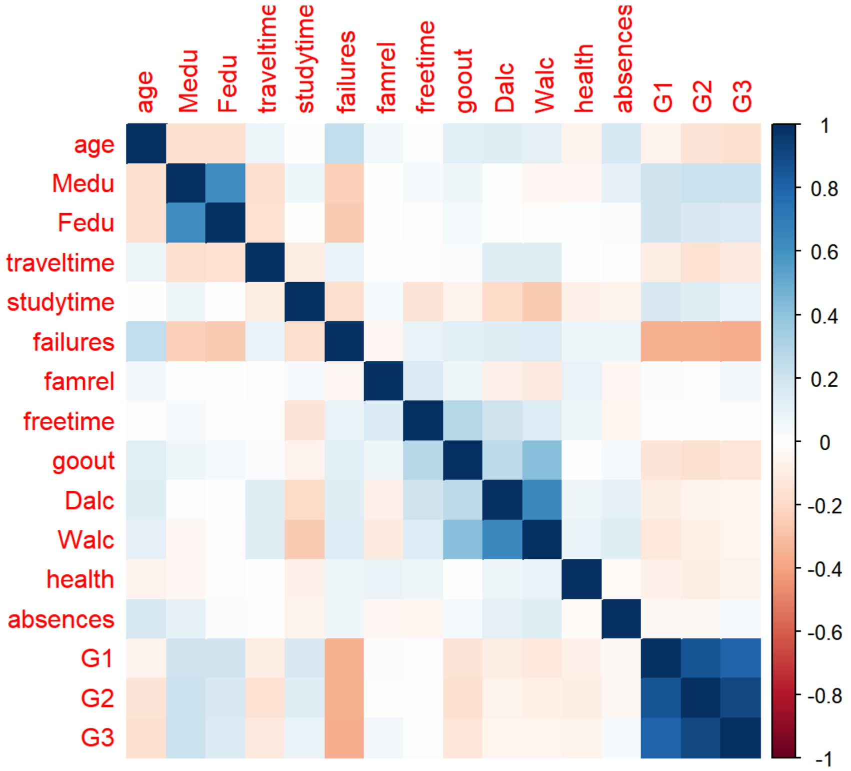


We can see that more female students failed than male students, there was a small discrepancy between the genders.

The graph below shows the *relationship between access to internet and the performance of the students* . It tells us that lack of internet access does impact a students average grade.



2.2.1. Correlation We will use correlation plots to explore the data and see if there are any significant information. We will only do this for the numeric data.

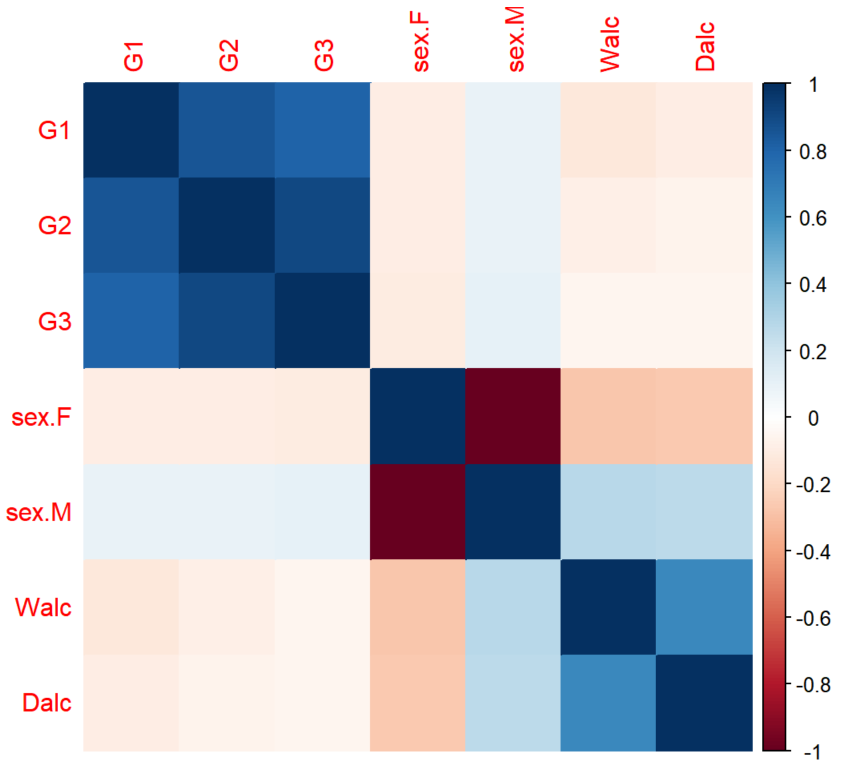


There is a clear high correlation between G1,G2 and G3, which makes sense. G1: first period grade, G2: second period grade, G3: final grade. This means that performing students do well each period, and poor performing students do worse for each period.

Additionally, a high grade value has a negative correlation with past failure. Mother and Father education levels are positively correlated, meaning the higher level of parental education level the higher grade a student will obtain.

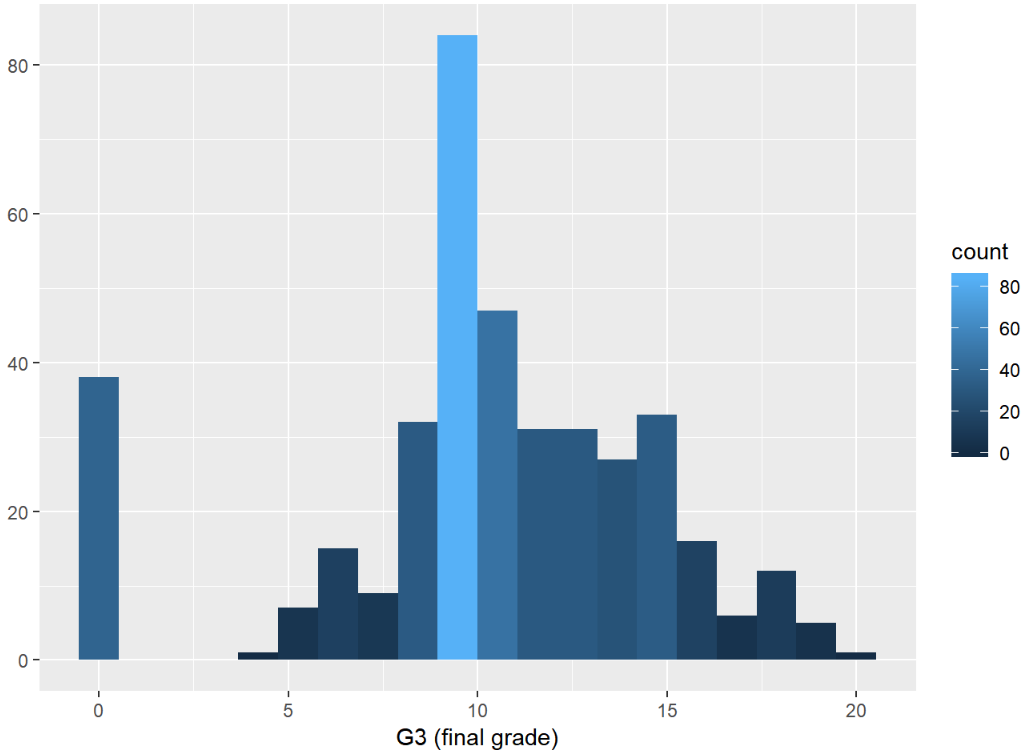
We can also look at the correlation between alcohol consumption, gender and performance of the students.

```
dmy <- dummyVars("~.", data=student_math_performance)
newdata <- data.frame(predict(dmy, newdata=student_math_performance))
correl1 <-cor(newdata[,c("G1", "G2", "G3","sex.F","sex.M","Walc","Dalc")])
corrplot::corrplot(correl1, method = 'color')
```



We also find that gender has a lower correlation with respect to performance as compared to alcohol consumption. We find that both Weekend and Workday Alcohol Consumption is highly negatively correlated with the grades.

2.2.2. G3 Now lets look at the variable that we want to predict *G3*.



There is quite a number of students that obtain a *0*.In addition, there is quite a high mean occurrence.

2.3. *Modelling*

Now that we have a basic understanding of our data set. After some strenuous model fitting attempts, a linear regression model was the most appropriate to use for this report. Take note: output for alternative models not provided for presentation purposes.

2.3.1. Linear Regression Model

```
##
## Call:
## lm(formula = G3 ~ ., data = train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.7681 -0.6423  0.2294  1.0691  4.5942
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -1.329568   2.474569  -0.537 0.591574
## schoolMS          0.838581   0.470545   1.782 0.076016 .
## sexM              0.034883   0.275586   0.127 0.899382
## age              -0.214994   0.119579  -1.798 0.073472 .
## addressU          0.067190   0.326035   0.206 0.836905
## famsizeLE3       -0.111068   0.283228  -0.392 0.695302
## PstatusT         -0.153653   0.401679  -0.383 0.702417
## Medu              0.279949   0.171111   1.636 0.103164
## Fedu             -0.221275   0.151103  -1.464 0.144422
## Mjobhealth        0.002065   0.610532   0.003 0.997304
## Mjobother         0.509947   0.403195   1.265 0.207209
## Mjobservices      0.475476   0.435332   1.092 0.275857
## Mjobteacher       0.285345   0.550640   0.518 0.604802
## Fjobhealth        0.433172   0.774191   0.560 0.576343
## Fjobother        -0.296792   0.577217  -0.514 0.607611
## Fjobservices     -0.311595   0.593628  -0.525 0.600148
## Fjobteacher      -0.321205   0.712695  -0.451 0.652628
## reasonhome       -0.431435   0.319907  -1.349 0.178755
## reasonother       0.159612   0.454480   0.351 0.725755
## reasonreputation -0.051845   0.317894  -0.163 0.870589
## guardianmother    0.267462   0.311371   0.859 0.391226
## guardianother    -0.157335   0.554872  -0.284 0.777003
## traveltime        0.274301   0.197865   1.386 0.166968
## studytime        -0.140650   0.155149  -0.907 0.365577
## failures         -0.185333   0.211040  -0.878 0.380739
## schoolsupyes      0.562716   0.379303   1.484 0.139268
## famsupyes         0.369402   0.268848   1.374 0.170745
## paidyes           0.060643   0.270971   0.224 0.823107
## activitiesyes    -0.286006   0.247519  -1.155 0.249063
## nurseryyes       -0.426858   0.315064  -1.355 0.176774
## higheryes         0.503353   0.677346   0.743 0.458148
## internetyes      -0.097405   0.331040  -0.294 0.768834
## romanticyes      -0.243837   0.268414  -0.908 0.364577
## famrel            0.494479   0.136663   3.618 0.000363 ***
## freetime         -0.139869   0.138297  -1.011 0.312879
## goout             0.078871   0.128794   0.612 0.540879
## Dalc             -0.248633   0.178366  -1.394 0.164651
## Walc              0.221434   0.139178   1.591 0.112950
## health            0.027495   0.095100   0.289 0.772748
## absences          0.060830   0.017915   3.396 0.000804 ***
## G1                0.162966   0.076956   2.118 0.035255 *
## G2                0.994677   0.066450  14.969  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.868 on 235 degrees of freedom
## Multiple R-squared:  0.8616, Adjusted R-squared:  0.8374
## F-statistic: 35.68 on 41 and 235 DF,  p-value: < 2.2e-16
```

3. *Results*

There were some residual values that have a value less than -5. It seems the model predicted that poor performing students would get a negative test results. In order to improve our model performance we will replace all negative values with *0*. So we should test our model on the testing set, and create a data set of actual and predicted results to check the model performance.

```
## [1]  0.00000 20.25837
```

```
## [1]  0.00000 20.25837
## [1] 0.7779023
```

This model performance obtains a R2 = 0.7779023, meaning that our model can explain about 78% variance in our test data.

4. *Conclusion*

The goal for this report was to predict G3 the final grade for students maths performance based on 33 predictors. As we explored the data set we did find some interesting things such as, more female students had a high failure rate for the final exam compared to the male students. Access to internet showed a deteriorating effect on the performance of the students as their average grades increased compared to those without access to internet.

Regarding correlation, we did find that there was a high correlation between G1,G2 and G3, which was expected. Mother and Father education levels were positively correlated, which means that the higher level of parental education level the higher grade a student will obtain. Additionally, both Weekend and Workday Alcohol Consumption is highly negatively correlated with the grades.

After attempting various model fittings, a linear regression model was the most appropriate to use for this report. The model obtained an R2 of *0.7779023*, meaning that the model can explain about 78% variance in our test data. Although, linear regression models are generally too rigid to be useful, in this case it worked rather well.

Recommendations

Alternatively, we could have focused on distinguishing above average students from students in difficulty, and only used a single measure (G1) of a student’s performance should be enough to classify them in one of two categories.

References

Ward, A. et al.(1996). “Achievement and Ability Tests - Definition of the Domain”, Educational Measurement, 2, University Press of America, pp. 2-5, ISBN 978-0-7618-0385-0 Cortez, P., & Silva, A. M. G. (2008). Using Data Mining to Predict Secondary School Student Performance. In A. Brito, & J. Teixeira (Eds.), Proceedings of 5th Annual Future Business Technology Conference, Porto, 5-12.