

NWI-IBI008-2016-KW1-V Datamining Faculty of Science

Predicting the number of class failures based on environment

*Authors: Teacher:*

Luca Lin, Brian Westerweel Tom Heskes

Date 12 March 2017

**Abstract**

Being able to access to higher education grants financial and health benefits (Baum, Ma, & Payea, 2010). For these reasons, it is important to know how to maximize results during educational career. Alcohol consumption influences certain cognitive skills such as attention bias (Field, Wiers, Christiansen, Fillmore, & Verster, 2010) and alcohol abuse leads to Korsakoff syndrome, a long-term effect of alcohol abuse that leads to amnesia (Butters, 2008). In 2010, 36% of children in fifth and sixth grade and 93% of sixteen year olds have drunk alcohol at least once (Hasselt, 2010). We applied regression analysis on a dataset of students with independent variable alcohol consumption and dependent variable class failures to predict if drinking alcohol could possibly influence class failures and fit the dataset into a decision tree algorithm to see if we could predict the number of class failures based on the dataset.

**0.1 Introduction**

In 2014, the government of the Netherlands raised the minimum drinking age from 16 to 18 to prevent young people from drinking. Alcohol has serious impairing effects on the growing brain (Donald W. Zeigler, et al., 2005). Young drinkers are more susceptible to a wide range of disadvantages, including impairments in brain activity and impairments in memory. Men drink on average more than women but women are more susceptible to the effects of alcohol due to the physiological differences between the genders (Dawnson & Archer, 1992).

In America, the average age at which a person drinks his or her first alcoholic beverage is 12 years (Donald W. Zeigler, et al., 2005), while the legal drinking age is 21. This tells us that a high minimal legal drinking age does not prevent people from drinking at a young age so it is left to the individual to decide.

Society and individual both benefit from higher education (Baum, Ma, & Payea, 2010). In the Global Competitiveness Report, 2016–2017, Switzerland, Singapore and the United States are reported to have the world’s most competitive economies. From the Human Development Report from 2014, we see the United States and Switzerland in the top 20 in the HDI (Human Development index, which is a composite statistic for life expectancy, education and income). The Netherlands is ranked fourth in HDI and fifth in GCI index.

**0.2 The dataset:**

For our dataset, we used attributes for both student-mat.csv (Math course) and student-por.csv (Portuguese language course) datasets. First, we had to import the datasets into python and pre-process them and second, we filtered out the unnecessary attributes for linear regression while keeping all the attributes for the decision tree. The attributes were as follows:

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:   
31 G1 - first period grade (numeric: from 0 to 20)   
31 G2 - second period grade (numeric: from 0 to 20)   
32 G3 - final grade (numeric: from 0 to 20, output target)

1. **Methods**

**1.1 Alcohol consumption and class failures**

For our report, we would like to know how well we can predict the number of class failures based on the alcohol attributes. A multiple regression analysis was done with dependent variable class failures and independent variables Workday alcohol consumption (x1) and Weekend alcohol consumption (x2). Our general null hypothesis is that b(x1) = b(x2) =0 while the individual null hypothesis is that b(x1) =0 and that b(x2) = 0

Furthermore, due to boys and girls having different physical properties and different drinking patterns, we split the dataset based on the attribute gender and applied multiple regression analysis on the split dataset. The null hypothesis is again that b(x1) = b(x2) = 0 while the individual null hypothesis is that b(x1) = 0 and that b(x2)=0

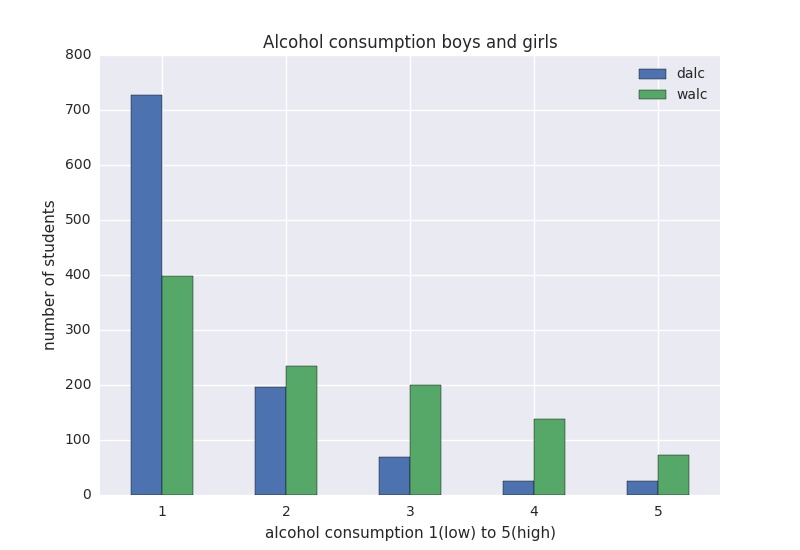
The dataset contains 453 boys and 591 girls. Below are histograms showing the distribution between dalc (workday alcohol consumption) and walc (weekend alcohol consumption). The x-axis represents the rate of alcohol consumption, from 1=low to 5=high:

Figure 1.a

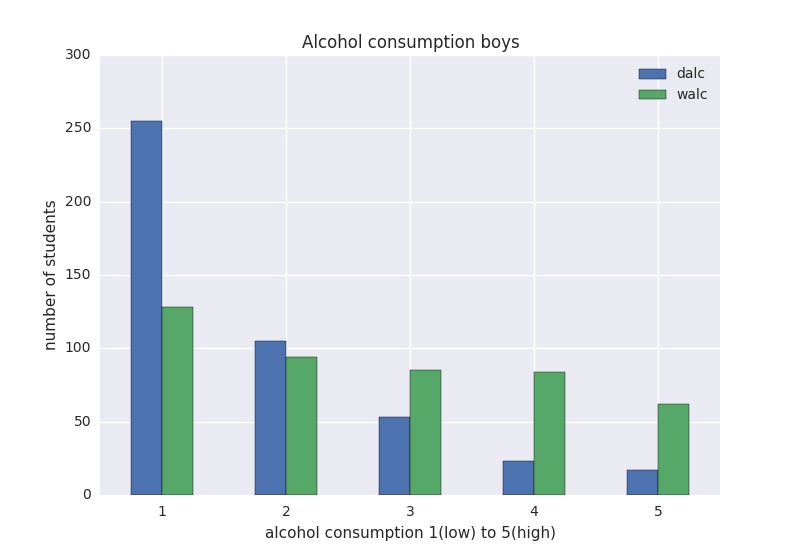


Figure 1.b

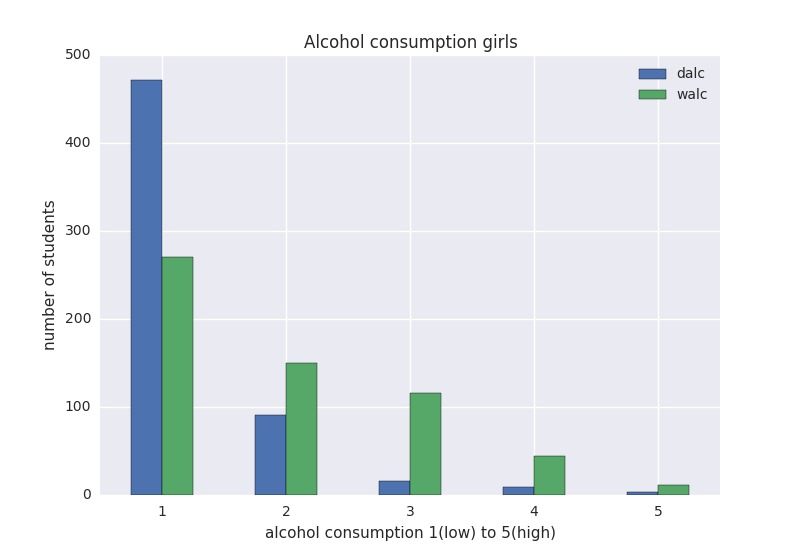


Figure 1.c

**1.2 Decision Tree**

Next, we fit the dataset in a decision tree algorithm that we learned from the Datamining 2016-2017 homework exercises to see if we could predict the number of class failures based on our dataset.

Before the decision tree algorithm could be applied pre-processing was needed. We got the data in the form of two CSV-files. To fit a decision tree based on the whole dataset the files need to be merged. After the merge the String-values needed to be translated to numeric values. This had to be done so the decision tree algorithm can interpret the values. The attribute which we want to use for classification is ‘failures’-column. Since this column is in the dataset it needed to be extracted so that it wouldn’t influence the classification.

After merging, extracting the ‘failures’-column and translating the String-values the data is ready to fit a decision tree upon.

**2.0 Results:**

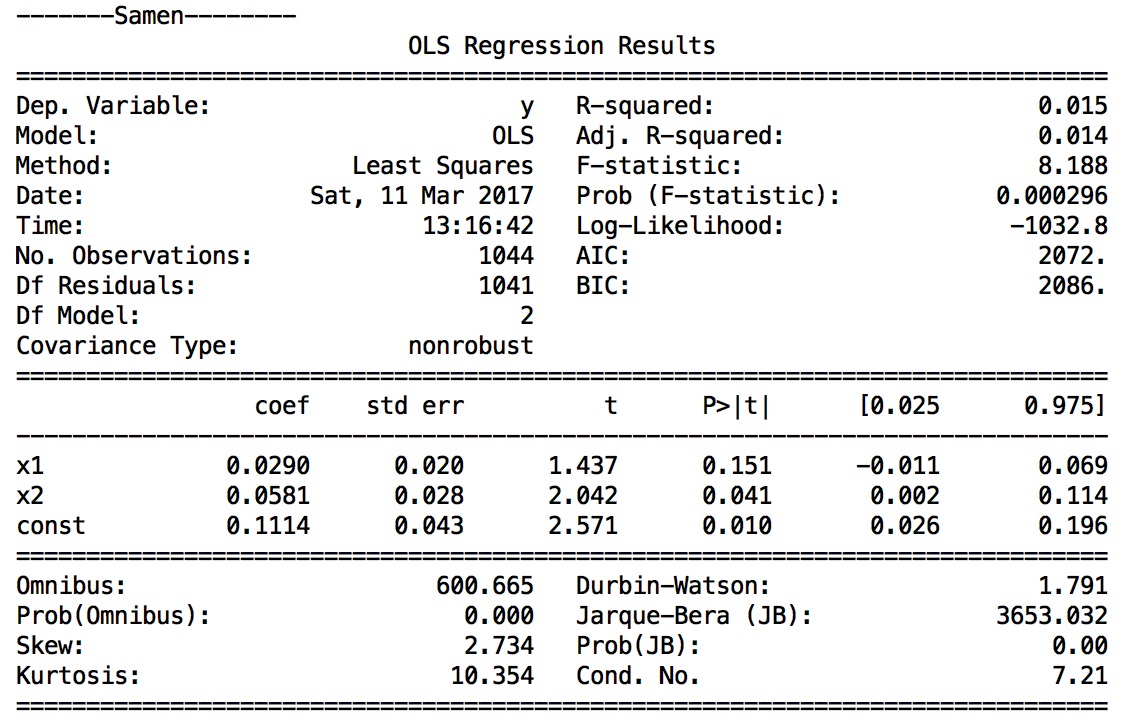
**2.1 Alcohol consumption and class failures:**

Dependent variable: class failures (Quantitative)

Independent variables: x1=Workday alcohol consumption (Quantitative)

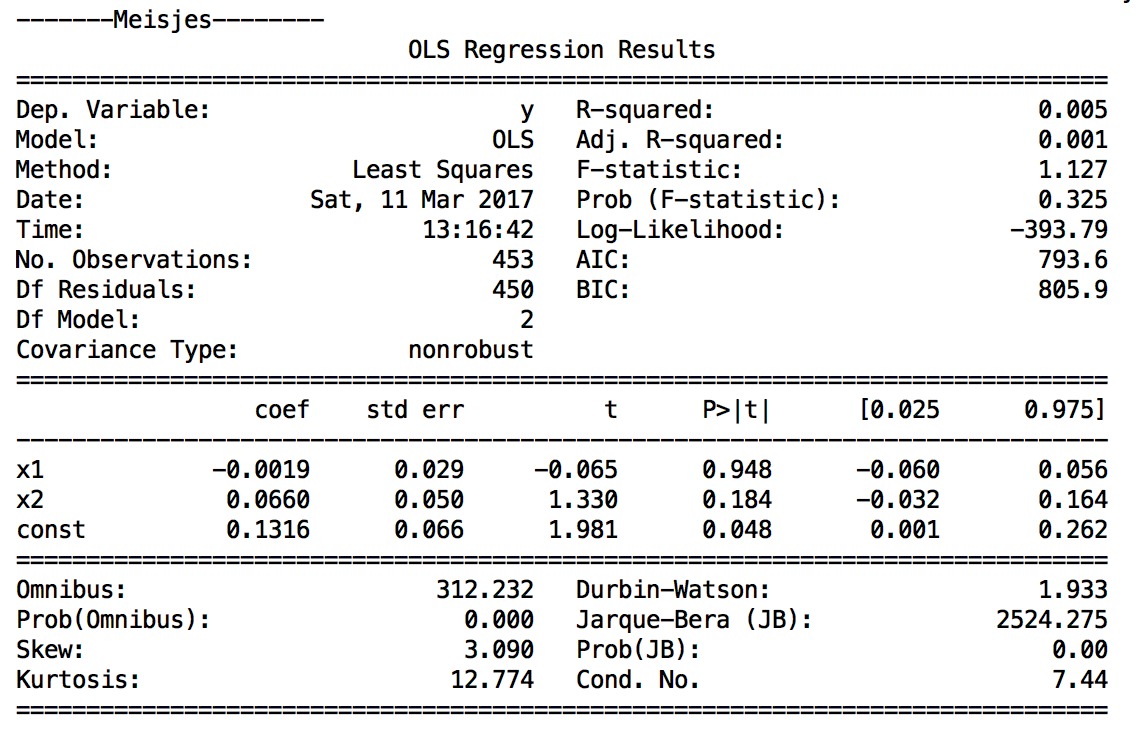
x2=Weekend alcohol consumption(Quantitative)

**Multiple regression analysis for all students:**



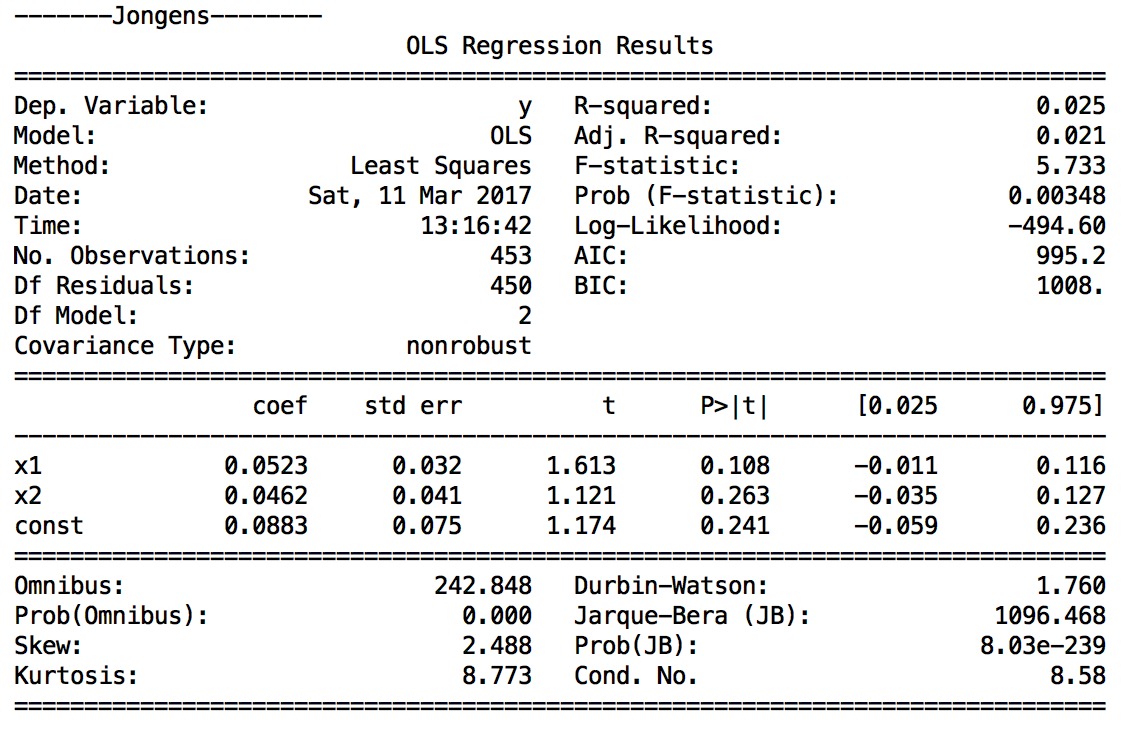
We reject the general null hypothesis (p≈0.000). The R-squared (0.015) for predicting the number of class failures out of the workday alcohol consumption and weekend alcohol consumption is significantly greater than 0. The effect is weak. Furthermore, the individual null hypothesis for x2 is rejected (p≈0.041) because the b weight (0.041) significantly deviates from 0, while we preserve the null hypothesis for x1(p≈0.151), because the b weight does not deviate significantly from zero.

**Multiple regression analysis for girls:**



We preserve the general null hypothesis for girls (p=0.325) because the found R-squared value is not significantly greater than zero.

**Multiple regression analysis for boys:**



We reject the general null hypothesis (p≈0.003). The R-squared (0.025) for predicting the number of class failures out of the workday alcohol consumption and weekend alcohol consumption is significantly greater than 0. The effect is weak. Furthermore, the individual null hypothesis for both x1 and x2 are preserved because the found b weights (p(x1) = 0.108, p(x2) = 0.263) do not significantly deviate from 0.

**2.2 Decision Tree**

We fit the dataset into a decision tree to try to determine the best attributes to split on if we want to minimize the number of class failures. The decision tree was made using the gini index as measure of node impurity.

Below is the decision tree.



Below is a legend for all variables in the tree:

X[0]=school (student school, binary)

X[1]=sex (binary, F=female, M=male)

X[2]= age (numeric from 15 to 22)

X[3]= address (home address, binary, U=urban, R=rural)

X[4]=famsize (binary, less or equal to 3, greater than 3)

X[5]=Pstatus (parents cohabilitation status, binary, together, apart)

X[6]=Mother education (numeric, 0=none, 1=primary education, 2=5th to 9th grade, 3=secondary education, 4=higher education)

X[7]=Father education (numeric, 0=none, 1=primary education, 2=5th to 9th grade, 3=secondary education, 4=higher education)

X[8]=Mother job (nominal, teacher, healthcare, civil services, at home, other)

X[9]=Father job (nominal, teacher, healthcare, civil services, at home, other)

X[10]=reason to choose this school (nominal, close to home, school reputation, course preference, other

X[11]=guardian (nominal: mother, father, other)

X[12]=traveltime (numberic, 1=<15min, 2=15-30min, 3=30-60min, 4=>1hour)

X[13]=studytime (weekly study time, numeric, 1=<15min, 2=15-30min, 3=30-60min, 4=>1hour)

X[14]=schoolsup (extra educational support, binary, yes,no)

X[15]=famsup (family support, binary, yes,no)

X[16]=paid (extra paid classes, binary, yes, no)

X[17]=activities (extra curricular activities, binary, yes, no)

X[18]=nursery (attended nursery school, binary, yes, no)

X[19]=higher (wants to take higher education, binary, yes, no)

X[20]=internet(internet access at home, binary, yes, no)

X[21]=romantic(with a romantic relationship, binary, yes, no)

X[22]=famrel(quality of family relationships (numeric, 1 for low to 5 for excellent)

X[23]=freetime (free time after school, numeric, 1 very low to 5 very high)

X[24]=goout (going out with friends, 1 very low, 5 very high)

X[25]=Dalc (workday alcohol consumption, numeric, 1 very low, 5 very high)

X[26]=Walc (weekend alcohol consumption, numeric, 1 very low, 5 very high)

X[27]=health (current health status, numeric, 1 very bad, 5 very good)

X[28]=absences (number of school abseces, numeric from 0 to 93)

X[29]=G1 (first period grade, from 0 to 20)

X[30]=G2 (second period grade, from 0 to 20)

X[31]=G3 (final grade, numeric, from 0 to 20)

**3.0 Conclusions and discussion**

**3.1 Conclusions for alcohol consumption and class failures**

There are multiple explanations for the results. In general, students who drink more are significantly more likely to fail classes than students who drink less or do not drink at all based on the dataset. The results can also be explained by an interfering variable like studytime. A study by D.E.Ukpong & I.N.George in 2012 shows that students who study for more hours perform better than those who study for short amounts of time.

We would like to mention that even though we could have included the variable studytime into the multiple regression analysis, we decided not to, as we did not plan that in our project proposal and it was not what we wanted to know.

Furthermore, drinking on workdays seems to have insignificant effect on class failures while drinking on weekends does have a significant effect on class failures. Primary explanation is that drinking during workdays has no effect on the number of class failures while drinking on the weekend does have effect on number of class failures.

However, non-students who are on their way of becoming adults drink more during the weekend than on workdays (Cathy Lau Barraco, Abby L.Braitman, Ashley N. Linden Carmichael, & Amy L. Stamates, 2016). From our dataset, we can see that this is also the case for students (fig 1.a). An alternative explanation for the results is thus that students who do not drink or restrain themselves from drinking during workdays but still fail classes, actually make it up by drinking more on the weekends. However, this is not yet researched.

We also found an insignificant effect of alcohol consumption for girls. The primary explanation for this is that alcohol does not influence failed classes for girls. An alternative explanation is that girls drink less than men (Figure 1.b shows that around 250 boys have low alcohol consumption during workdays and around 125 have low alcohol consumption in weekends, while girls have around 475 low alcohol consumers during the workweek and 275 low alcohol consumers during weekend in figure 1.c. Likewise, if we compare the other end of the extreme, boys are more heavy drinkers than girls).

The effect of alcohol consumption on failed classes was for boys significant. There are again, multiple explanations possible. Primary explanation is that alcohol consumption influences class failures. Alternatively, boys who drink more could be studying less. We could and should have included study time in our linear regression but we did not and it is a point in which we could improve on in the future.

**3.2 Decision tree**

From the decision tree, we would like to predict who fails classes and who passes based on the dataset attributes. Note that some of the attributes are predetermined, like who the guardians are and age and some attributes are not used at all in the tree because the algorithm split based on Gini index, a measure of node impurity. We decided to include all the attributes in the decision tree, as we did not know whether the attributes would influence the grown decision tree.

From our list of best attributes, we see that students who have higher grades in their first period, tend to do better overall. The decision tree starts splitting by looking at first period grades. If we follow the path to the leaf with least amount of class failures and lowest gini index, the decision tree splits based on the following attributes: age, number of absences, grades in first period again, want to take higher education, grades in second period, guardian, weekend alcohol consumption, grades in second period in that order. Doing this, we end up in the leaf with the lowest gini index and highest samples (303 students).

Our conclusion is that the best way to determine whether someone will pass most of their classes is to look at what they have achieved during the first period. If we look at the first split, we can immediately see that of the 745 students who have a first period grade higher than 9.5 (maximum is 20), 686 of them pass all their classes versus the 298 of which 174 pass all their classes. Furthermore, if we split based on age we find that younger students pass more classes (712 students younger than 18.5, of which 669 passed all their classes vs 33 students older than 18.5 of which only 17 passed all their classes) and that older people (18.5+) who have low grades the first period, also fail more later (of the 36 students who have low first period grade, 3 pass all their classes).

Wanting to take a higher education also seems to be relevant. Of the 534 people who have first semester grades higher than 10.5, are younger than 18.5, have less than 11.5 absences, 519 of those pass all their classes.

For future improvements, we would like to reduce the size of the tree by pre-pruning or post-pruning the nodes that are not needed to prevent overfitting. This would most likely improve the decision tree by generalizing it more and thus prevent overfitting. Now, attributes like first period grade are used multiple times and the tree sometimes splits the dataset in a group of 2 persons and a group of 522. This seems redundant and we would like to prevent this from happening in the future.

# Bibliography (Schwab & Sala-i-Martín, 2016-2017)

Bank, W. (2017). *Gross national income per capita 2015, Atlas method and PPP.* World Development Indicators database.

Baum, S., Ma, J., & Payea, K. (2010). *Education Pays, 2010: The Benefits of Higher Education for Individuals and Society. Trends in Higher Education Series .*

Butters, N. (2008). *Alcoholic korsakoff's syndrome: Some unresolved issues concerning etiology, neuropathology, and cognitive deficits .* Journal of Clinical and Experimental Neuropsychology .

Cathy Lau Barraco, P., Abby L.Braitman, P., Ashley N. Linden Carmichael, M., & Amy L. Stamates, B. (2016). *Differences in Weekday versus Weekend Drinking among Nonstudent Emerging Adults.* Exp Clin Psychopharmacol.

D.E.Ukpong, & I.N.George. (2012). *Length of study time behaviour and academic achievement of social studies education students in the university of Uyo.* Canadian Center of Science and Education.

Dawnson, D. A., & Archer, L. (1992). *Gender differences in alcohol consumption:effects of measurement.* British Journal of Addiction.

Donald W. Zeigler, P., Claire C. Wang, M., Richard A. Yoast, P., Barry D. Dickinson, P., Mary Anne McCaffree, M., Carolyn B. Robinowitz, M., & Melvyn L. Sterling, M. (2005). *The neurocognitive effects of alcohol on adolescents and college students.* for the Council on Scientific Affairs, American Medical Association.

Field, M., Wiers, R. W., Christiansen, P., Fillmore, M. T., & Verster, J. C. (2010). *Acute Alcohol Effects on Inhibitory Control and Implicit Cognition: Implications for Loss of Control Over Drinking.*

Field, Wiers, Christiansen, Fillmore, & Verster. (2010). *acute alcohol effects on inhibitory control and implicit cognition: implications for loss of control over drinking.*

Hasselt, N. v. (2010). *Preventie van schadelijk alcoholgebruik en drugsgebruik onder jongeren.* Utrecht: Trimbos Instituut.

Schwab, P. K., & Sala-i-Martín, P. X. (2016-2017). *The Global Competitiveness Report.*

1617 Data Mining (NWI-IBI008-2016-KW1-V) *homework assignment 3*, Radboud University Nijmegen