Family, Gender, Alcohol, and Academic Performance: What're the Connections?

Nguyen Pham, Yen Nguyen

Denison University, DA 101 - 01

Dr. Wang Zhe

May 4th, 2023

I. Introduction

1.1. Abstract

In 2016, the National Institute of Health reported that 26% of 8th graders, 47% of 10th graders, and 64% of 12th graders have all had experience consuming alcoholic drinks. This finding indicates an accelerating trend in underage alcohol use among school students, which has become a growing concern among the public regarding students' academic performance. However, it is common for people to believe female students will perform better than male students, and parents' education will also affect students' performance. To uncover these stereotypes on students' academic performance, our research aims to use statistical analysis and visualization from a comprehensive dataset collected from a survey of students in math and Portuguese language courses at Gabriel Pereira and Mousinho da Silveira secondary school.

1.2. Introduction

This project uses statistical analysis to explore the stereotypes surrounding high school student's academic performance in Portuguese language courses. The dataset was obtained from UCI Machine Learning Repository containing detailed information on students at Gabriel Pereira and Mousinho da Silveira secondary school. It has a total of 33 variables and 649 observations.

In this research, our independent variables are Sex, parents' education(parent_edu), weekly alcohol consumption (alc), and our dependent variable is Final grade (G3).

Research Question: Do family background, alcohol consumption, and gender affect high school students' grades?

Null Hypothesis (H0): There is no significant influence of parents' education, gender, and alcohol consumption on students' academic performance

Alternative Hypothesis (Ha): There is a significant influence of parents' education, gender, and alcohol consumption on students' academic performance

1.3. Objectives

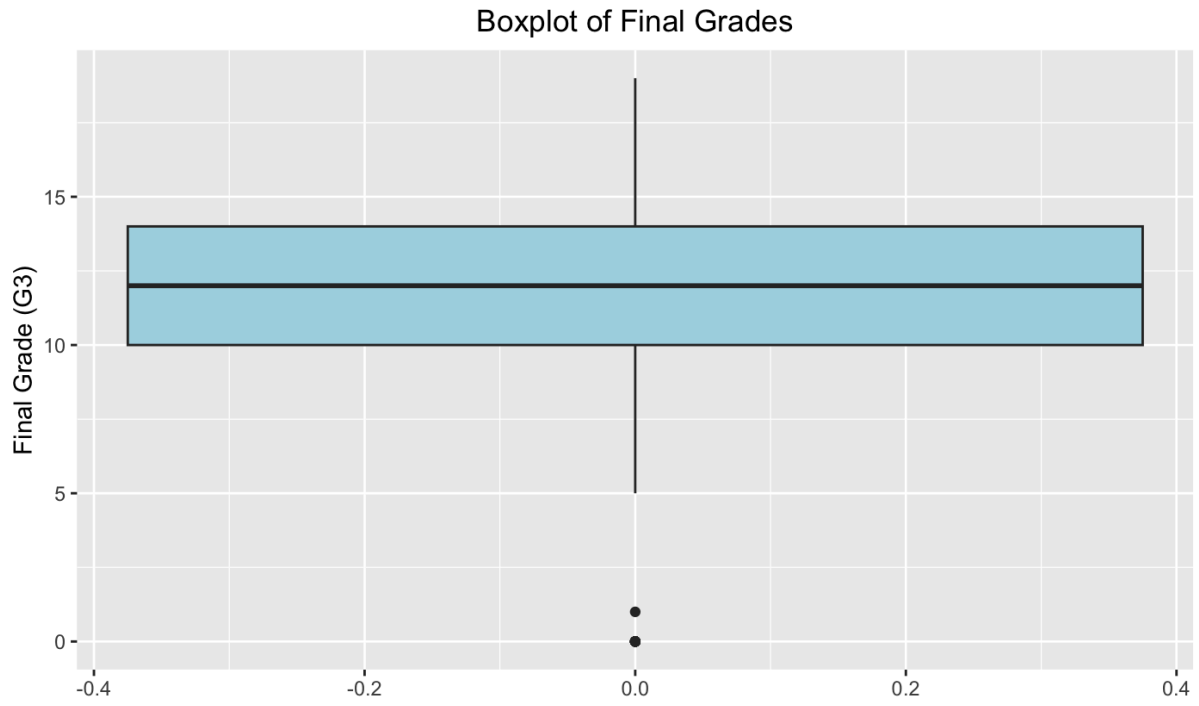
By doing this statistical research, we want to explore the relationship between gender, family background, alcohol consumption, and academic performance. We will use a t.test and linear regression to examine how these factors impact students' final grades in Portuguese language courses and identify risk elements for poor academic performance. From the results, we can inform interventions, policies, and prevention efforts aimed at supporting students' academic achievement.

II. Ethical Consideration

When conducting this study, it is important to consider several ethical considerations such as informed consent, confidentiality and privacy, and minimizing harm. As the dataset contained information about grades, family status, genders, and drinking habits, it is crucial to ensure that participants are fully aware of the study and willing to provide the answers. Also, because of this personal information, participants' privacy and confidentiality should be protected from misuse or privacy threats. Finally, researchers should ensure the harm to participants is minimized by avoiding sensitive and triggering questions and providing resources for counseling services if their mental well-being is affected. By considering these ethical principles, researchers can conduct their study in a responsible and respectful manner that contributes to students' academic success.

III. Data Preparation

3.1. Data Cleaning



The boxplot shows there are some outliers in the grades. Therefore, we have to remove the outliers which are those with grades lower than 5 to make the distribution normal and improve the accuracy of finding the factors affecting the final grades.

3.2. Data Wrangling

From the dataset, we will focus on the variables such as sex, Weekend alcohol consumption(Dalc), Workday alcohol consumption(Walc), Students' Final grade(G3), Father's education(Fedu), and Mother's education(Medu). The table below is the summary of our variable lists.

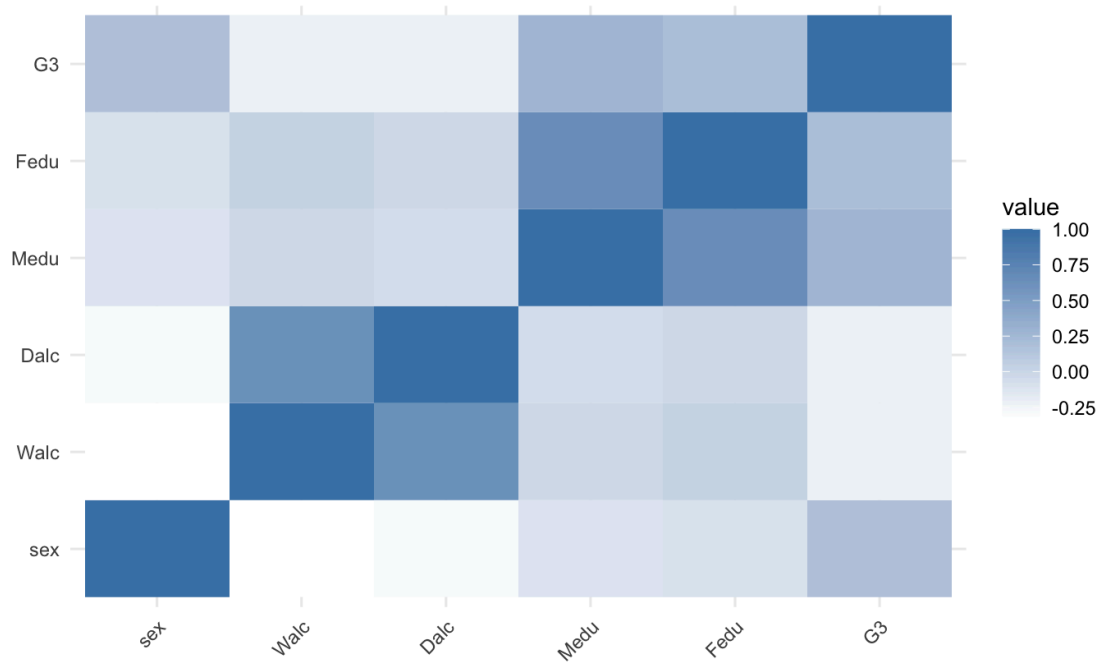
Variable Name	Description	Data Type	Range	Percentage of Missing Data
Sex	Students' gender	Nominal	M/F	0%
Walc	Weekend alcohol consumption	Integer	1 - very low to 5 - very high	0%
Dalc	Workday alcohol consumption	Integer	1 - very low to 5 - very high	0%
G3	Students' Final grade	Integer	0 - 20	0%
Fedu	Father's education	Integer	0 - none, 1 - primary education (4th grade), 2 – 5th to 9th grade, 3 – secondary education or 4 – higher education	0%
Medu	Mother's education	Integer	0 - none, 1 - primary education (4th grade), 2 – 5th to 9th grade, 3 – secondary education or 4 – higher education	0%

Since there are a small number of students whose parents do not live together and the dataset does not provide information on whether they live with their mother or father, we have decided to focus our analysis on students whose parents' cohabitation status is "together" (Pstatus = T). This subset of data is likely to provide more accurate and meaningful insights into the factors affecting student performance. This subset of data comprises approximately 90% of the students in the dataset.

IV. Data Exploration

4.1. Heatmap

We will use a heatmap to explore the relationship between different factors and their impact on final grades in our dataset. This visualization will provide us with a clear overview of the correlation between the variables and help us identify any strong positive or negative relationships between them.



From the heatmap, we observed a strong positive correlation between weekday and weekend alcohol consumption as well as a high positive correlation between the education level of a student's mother and father.

Therefore, we would like to use composite variables (one for parental education, and another for alcohol consumption) to combine these two pairs of variables in order to reduce multicollinearity in our regression model and improve its accuracy. The parental education

composite variable ("parent_edu") and the weekly alcohol consumption composite variable ("alc") are calculated using the following formulas:

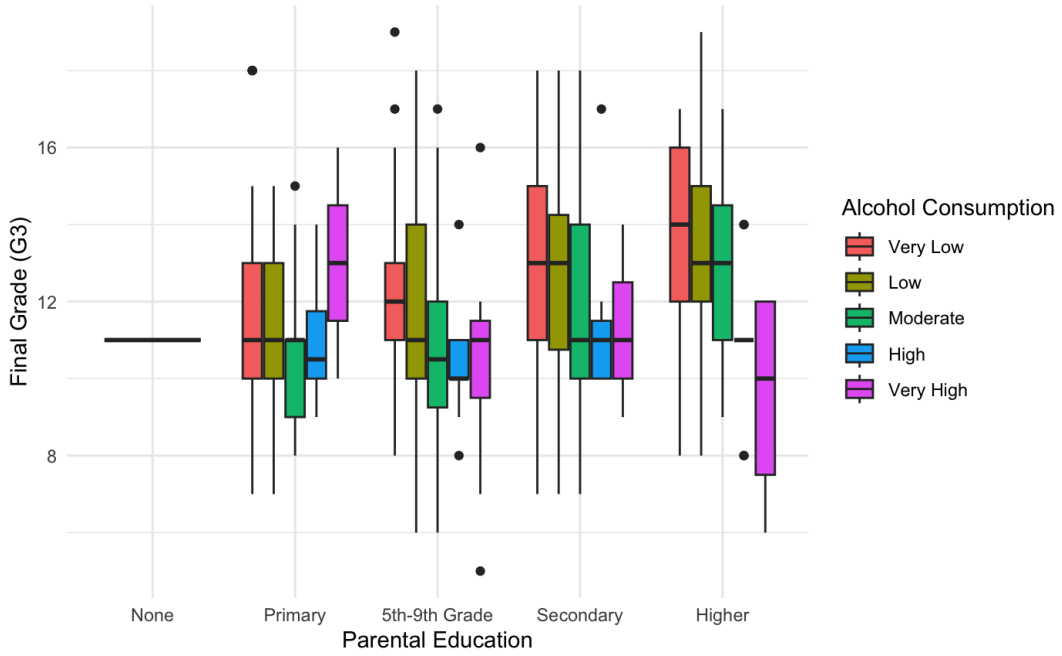
$$alc = (Walc*2 + Dalc*5)/7$$

$$parent\_edu = (Medu + Fedu)/2$$

The results are rounded to the nearest number and turned into categorical values with different factors to be used in the regression model.

4.2. Boxplots

We will generate boxplots to visualize the distribution of final grades (G3) across different levels of parental education and alcohol consumption. This allows us to compare the distributions of final grades between groups and identify any potential relationships or patterns between the variables in our dataset.



We can observe that as the level of alcohol consumption increases, the range of final grades decreases, indicating a negative correlation between alcohol consumption and

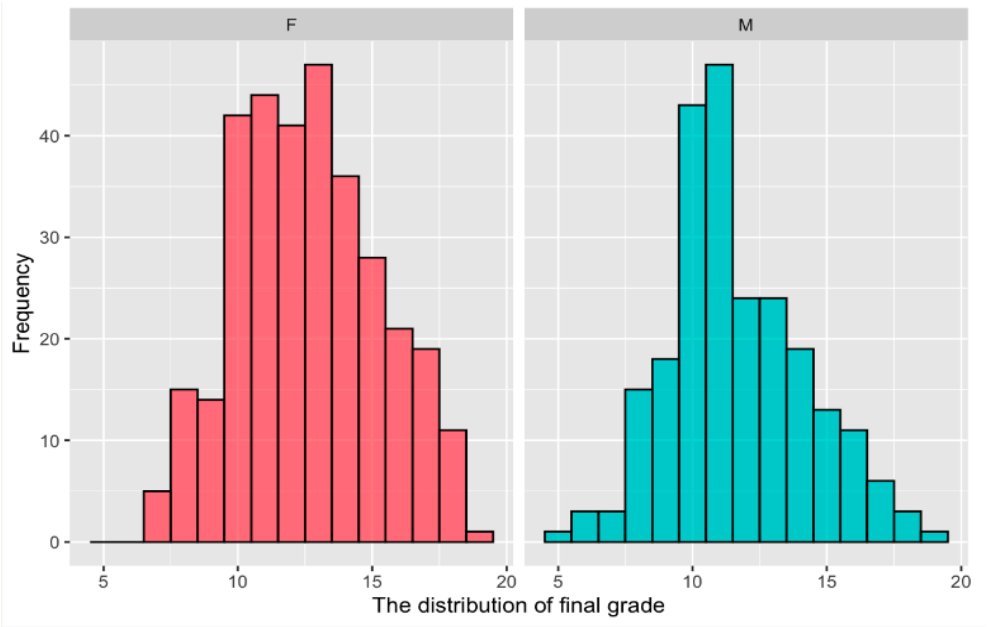
academic performance. On the other hand, the boxplots for parental education show that students with parents having higher education tend to have higher average grades in Portuguese. To gain a deeper understanding of this pattern and to analyze the relationship between these variables, further statistical analysis is required.

V. Statistical Analysis and Interpretation

4. 1. T-test to compare two sex groups

4.1.1. Check for assumptions

Before performing a t-test to compare the mean Portuguese grades between two sex groups, we first check the assumption of normality on the data of each group. We do so by using side-by-side histograms.



Looking at the above histograms, we can see that the data for both groups are now normally distributed and the assumption of normality is met.

4.1.2. Results

```
        Welch Two Sample t-test

data:  G3 by sex
t = 4.4278, df = 504.99, p-value = 1.168e-05
alternative hypothesis: true difference in means between
group F and group M is not equal to 0
95 percent confidence interval:
 0.5529005 1.4349142
sample estimates:
mean in group F mean in group M
       12.61728        11.62338
```

From the t.test, the p-value is 1.168e-05, which is much smaller than 0.05. This means that the difference between the mean final grade of the two groups is significant. Also, the 95% confidence interval does not include 0 so we can conclude that 95% of the true population mean difference is non-zero. Overall, the mean final grade of the female group is larger than the male group and this difference is statistically significant.

4.2. Multiple linear regression to predict the final grades

4.2.1. Results

We perform a multiple linear regression model to find the relationship among gender, family background, alcohol consumption, and students' final grade. In this model, gender is represented by the dummy variable "sex" which takes a value of 0 for Male and 1 for Female. Family background is represented by the ordinal variable "parent_edu" that shows the 5

education levels of students' parents (0-4). Alcohol consumption is represented by the ordinal variable "alc" with 5 different consumption levels (1-5).

```
lm(formula = G3 ~ sex + alc + parent_edu, data = studentData)

Residuals:
    Min      1Q  Median      3Q     Max
-6.3642 -1.6635 -0.1338  1.5722  6.8662

Coefficients:
                         Estimate Std. Error t value Pr(>|t|)
(Intercept)              10.0922     2.5061   4.027 6.45e-05 ***
sex                       0.9078     0.2293   3.958 8.54e-05 ***
alcLow                   -0.3862     0.2481  -1.557  0.12006
alcModerate              -0.8564     0.3205  -2.672  0.00777 **
alcHigh                  -1.3510     0.4813  -2.807  0.00518 **
alcVery High             -1.3909     0.6064  -2.293  0.02220 *
parent_eduPrimary         1.1054     2.5137   0.440  0.66030
parent_edu5th-9th Grade   1.4278     2.5070   0.570  0.56924
parent_eduSecondary       2.3642     2.5093   0.942  0.34651
parent_eduHigher          3.0143     2.5104   1.201  0.23038
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.496 on 545 degrees of freedom
Multiple R-squared:  0.1368,    Adjusted R-squared:  0.1225
F-statistic: 9.595 on 9 and 545 DF,  p-value: 1.189e-13
```

The reference levels for "alc" and "parent_edu" are "alc1" and "parent_edu0", which are level 1 (very low) for alcohol consumption and parent_edu0 (no education) for parents' educational level. This also means that the coefficients for these levels are not explicitly shown in the model output.

The intercept coefficient is 10.0922, which represents the predicted mean value of G3 when all predictor variables, including sex, alc, and parent_edu, are equal to their reference levels, which are 0, alc1 and parent_edu0. The intercept coefficient of 10.0922, therefore,

represents the predicted mean value of G3 for a male student who consumes alcohol at a very low level and whose parent has no education.

The estimated regression coefficient for “sex” is 0.9078, which suggests that the mean final grade of males is about 0.91 lower than that of females, holding all other predictor variables constant. The p-value is also much lower than 0.05, so the relationship between students’ gender and Portuguese final grade is statistically significant.

The coefficients for “alc3”, “alc4”, “alc5” are statistically significant, as their p-values are less than 0.05. This means that these predictors have a statistically significant relationship with students’ final grades, while the predictors “alc2”, “parent_edu1”, “parent_edu2”, “parent_edu3”, and “parent_edu4” are not significant, as their p-values are greater than 0.05. The negative descending coefficients for “alc2”, “alc3”, “alc4”, and “alc5” suggest that higher levels of alcohol consumption are associated with lower final grades, although the effect is only statistically significant for “alc3”, “alc4”, and “alc5”.

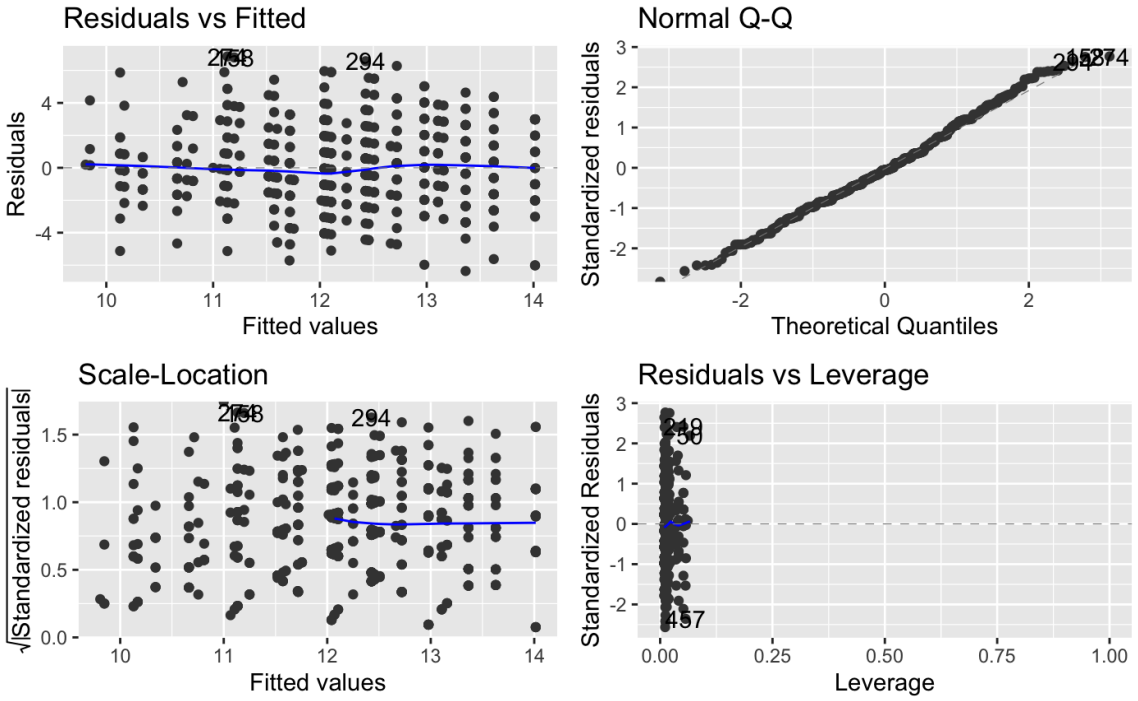
The adjusted R-squared value of 0.1225 indicates that 12.25% of the variance in the final grades can be explained by the predictor variables included in the model. The F-statistic value of 9.595 with a p-value of 1.189e-13 suggests that the overall model is statistically significant in predicting final grades. The residual standard error of 2.496 represents the estimate of the standard deviation of the error term in the regression model, and the residuals show a relatively normal distribution.

In short, “sex” is positively correlated with grades, which means female students have a better academic performance at Portuguese than their male counterparts do. The models also show that alcohol consumption has a significant negative correlation with final grades. On the other hand, the coefficients for all educational levels of students’ parents do not appear to be statistically significant, which suggests that there is not a clear relationship

between parents' education and students' final grades.

4.2.2. Check for assumptions

To check for the assumptions of the above multiple regression model, we generate diagnostic plots.



From the diagnostic plot, the points on the QQ plot follow a relatively straight line, which means that the distribution of the residuals is normal. The Residuals vs Fitted plot shows a random scatter of points with no discernible pattern. Therefore, the assumptions of linearity, homoscedasticity, and normality for this multiple linear regression model are met.

VI. Conclusion

5.1 Summary

From the t-test, female students have significantly higher grades in Portuguese than their male counterparts. Parental education positively affects students' academic success in Portuguese language courses. The higher education of one student's parents is, the higher Portugues grades they are likely to earn. Alcohol consumption is a risk factor for poor academic performance, but its influence is not statistically significant. This result highlights the need for parental involvement, and timely acts to buffer the negative effects of early alcohol on academic success.

5.2 Limitations

As our dataset contains information only about students in Portugal classes at two secondary schools, there are limitations to our research. First, our study may lack diversity and can not be representative of all students in Portugal or other countries to support our conclusion that alcohol consumption negatively affects students' grades. Second, we can not take into account all contextual factors that may impact students' performance such as their study habits, mental and emotional well-being, and socioeconomic status. Because of this, our study will not provide a comprehensive understanding of all factors affecting academic performance. Finally, the dataset was created in 2016, meaning the data may be outdated and may not capture the impact of parental education, alcohol consumption, and gender on academic performance in the long run. Therefore, our findings will potentially be impacted and can not fully capture

5.3 Future works

To address these limitations, researchers can consider collecting more up-to-date data in addition to this existing dataset to provide a more accurate and comprehensive interpretation of students' performance in these secondary schools. Also, from our findings,

we see a need for datasets from other subjects and countries to provide more diverse insights. Therefore, researchers can consider combining data from multiple sources that include information on grades of different subjects and backgrounds to have a more comprehensive dataset. By increasing the sample size, the findings will be more representative and accurate when answering our research question. Finally, as our study does not cover every factor contributing to poor academic outcomes and underage alcohol consumption, researchers can continue to explore the underlying reasons such as social, cultural, psychological, and biological factors. By better understanding the causes of alcohol use and academic failures, interventions can be more effectively targeted and tailored to specific populations.

VII. References

archive.ics.uci.edu. (n.d.). *UCI Machine Learning Repository: Student Performance Data Set*. [online] Available at: http://archive.ics.uci.edu/ml/datasets/Student+Performance.

Palaniappan, S., Kaur, H., Muthusamy, M., and Singh, J. (2017). "Classification of Alcohol Consumption among Secondary School Students." JOIV: International Journal on Informatics Visualization, 1(4-2), 224. https://doi.org/10.30630/joiv.1.4-2.64.

Cortez, P., and Silva, A. (2008). "Using Data Mining to Predict Secondary School Student Performance." Available at: http://www3.dsi.uminho.pt/pcortez/student.pdf.