Educational Data Mining Project, Fall 2018

Walter Genchi
Andrei-Daniel Comănescu

December 12, 2018

Abstract

This paper analyses the relationship between students' overall academic performance and their personal attributes. A series of statistical techniques are used to provide useful insights regarding student behaviour.

Contents

1 Methodology

1.1 Data

The used data comes from the *Student Performance Dataset* hosted by *UCI Machine Learning Repository*[2]. This dataset is part of a 2008 research paper[1] published by *P. Cortez* and *A. Silva.* In this paper, the researchers collected the grades, the number of absences and the answers to some questions, that revealed 29 socio-demographic variables from the pupils of two distinct Portuguese high schools that attended the classes of Mathematics and Portuguese.

The dataset of the students that took the Mathematics class is composed of 395 students, while the dataset containing the information pertaining to the Portuguese class contains 649 entries. The number of students that took both classes is that of 320. The data sets, original ones and the combined ones that stem from them, and the description of the variables is available on the project's Github repository[1].

1.2 Context

This project aims to understand the relationship between high school grades, which act as an universal metric in measuring academic performance, and the personal attributes of a student. These attributes are of multiple kinds including aspects such as: social life, habits, amount of time used for studying, family relationships, marital status of the parents, educational levels of the parents and so on. The latent variables which cannot be directly measured, e.g. student behaviour, also have a significant impact on the student's overall performance.

1.3 Research Questions

The research question of this project are as follows:

- Are there any patterns or correlations which clearly emerge when performing statistical analysis from a descriptive (e.g. data visualisation) and inferential (e.g. hypothesis testing) point of view?
- Can we discover within the high school student sample some clusters that reveal interesting insights into the student behaviour?
- Which are the undesirable high school student behaviours that lead them to perform poorly in an educational setting?

1.4 Approach

A series of tools for descriptive statistics, e.g. scatter plots, and inferential statistics, such as ANOVA-test, are used to answer the first research question. The answer to the second research question is provided with the aid of clustering techniques, e.g. K-Means algorithm. The third research question is tackled via latent variable discovery and dimensionality reduction techniques such as Principal Component Analysis (PCA) and Factor Analysis (FA). Moreover, Lasso Techniques and Decision Trees are used to analyse the relationship between student's final grade and personal attributes.

2 Results

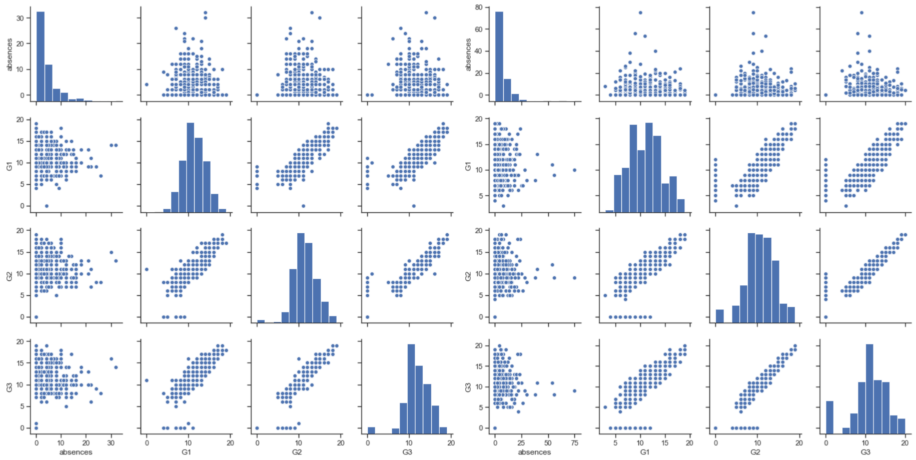
2.1 Descriptive Statistics

The datasets were visualised based on their variable types. A full set of visualisations can be observed in the Jupyter Notebooks available in the repository[2].

The datasets were visualised with the help of scatter plots when comparing the quantitative variables to the quantitative ones, with box plots and violin

[1] https://github.com/andrei-comanescu/edm-2018-hy/tree/master/data

[2] *Visualizing the Combined data set.ipynb*, *Visualizing the Mathematics class data set.ipynb* and *Visualizing the Portuguese class data set.ipynb* from https://github.com/andrei-comanescu/edm-2018-hy/tree/master/src

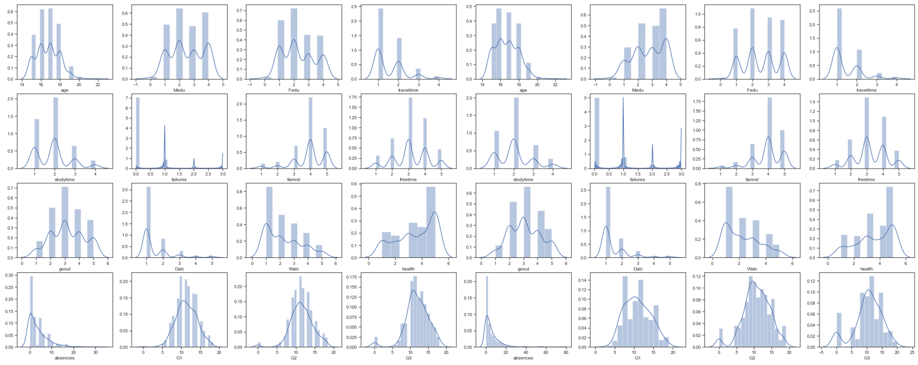


(a) Portuguese class

(b) Mathematics class

Figure 1: Pairplot of grades and absences in the two classes.

plots[3] when comparing factor variables to quantitative ones and with mosaic plots when visualising factor variables vs. quantitative ones. The marginal distribution of some variables was also observed. Some violin plots will be displayed in the coming sections, thus in this section we will only present a couple of scatter plots and distribution plots, which have been selected based on personal curiosity or interesting patterns observed in the data.



(a) Portuguese class

(b) Mathematics class

Figure 2: Marginal distributions of selected variables in the two classes.

[3] A variation of box plots which allows to better visualise the number of statistical units in each plotted percentile.

2.2 Results, Discussion and Interpretation

The input to principal component analysis (PCA), factor analysis (FA) and clustering methods was a set of 11 quantitative variables (out of 29). In the following analysis we excluded:

- the 17 factor variables, since their information was redundant and could not be naturally handled by PCA.

- variables *age* and *traveltime* were excluded since their information was redundant w.r.t. *failures* and *Medu/Fedu*. In fact, it is reasonable to assume that the more classes you failed in the past, the older you are. Moreover, *traveltime* was practically influenced by *Medu/Fedu*, since highly-educated parents are likely to live in urban areas and thus close to the schools.

- variables *G1*, *G2*, *G3*, since they are the "target" variables, i.e. they are the effect (not the cause) of student behaviour.

Besides these a series of models were trained with the help of the Lasso technique to either predict if a student passes or fails a course, or the final grade of a given student. Decisions trees were also constructed to extract a series of patterns that would better explain why a student passed of failed a class.

2.2.1 Principal Component Analysis (PCA)

The objective of PCA was to summarize the set of the 11 chosen variables in fewer dimensions, e.g. 2 principal components.

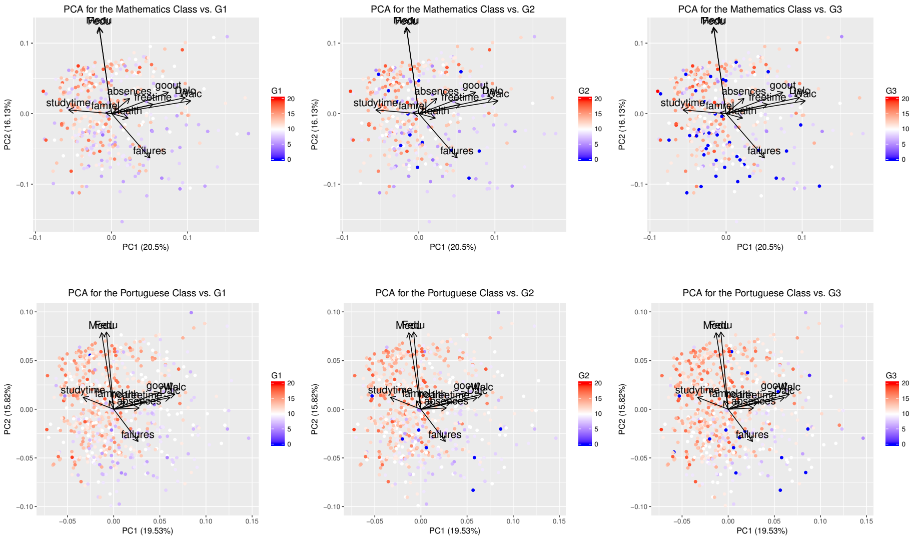


Figure 3: Biplots for Mathematics and Portuguese classes vs. Academic Periods

The results in both Portuguese and Mathematics class shows that the first 2 principal components (after scaling the data) can explain together only 36% of the total variance. In other words, data do not show any strong linear relationship (the only which can be captured by PCA). Another visual interpretation is that students are spread as a “cloud” of points in the 11-dimensional space, i.e. the 11 variables are almost “orthogonal”.

Figure 3 (also known as biplot) shows the the first 2 principal components scores for the Mathematics and Portuguese students. The arrows represent the projected coordinate system for the original variables, i.e. the orthogonal projection of a student onto the “failures” vector will provide the “failure score” for that student. The colour of each student represents the grade obtained for each period, where students coloured in white are those who passed the exam with the minimum (grade=10).

Both Portuguese and Mathematics class show that students who failed school periods are mostly located in the low region (i.e. have negative PC2 score) and have failed many classes in the past (“failure” vector is almost parallel to y-axis). Among the students who failed, some have studied a lot and have good family relationships (especially in the Mathematics class), but most of them have spent their time going out, drinking alcohol and skipping school (especially in the Portuguese class). Regarding students who passed school periods, Mother and Father’s education together with time of studying seem to strongly influence student’s performance (*Medu* and *Fedu* vectors are almost parallel to y-axis), especially in the Portuguese class. Having said that, students who are good in Mathematics seem to be less influenced by these variables.

By looking in sequence at the plots for each school period, a time trend visually emerges in the distribution of grades: students who did not pass the first period, got same or worse results in the following two periods.

2.2.2 Factor Analysis (FA)

After having looked at possible summary variables with PCA, our focus has later shifted on modelling variables’ variability with some latent variables. The technique used is called Exploratory Factor Analysis (EFA), which has more underlying assumptions than PCA, e.g. data distribution and continuous input (ordinal variables has been discretized), which are overall satisfied in our dataset.

The number of factors (4 for Mathematics, 5 for Portuguese) has been decided according to the Chi-Square test provided in the R output. The used rotation function was the default “varimax”, which tries to maximize the variance between the squared loadings of a factor. This type of rotation helps to easily identify which factors has the most influence on which variables, i.e. for each variable the factor loadings are either small or large. In Figure 4 the factor loadings are represented with their absolute values, where blue and red colours mean they have positive or negative sign respectively.

Starting with the Mathematics class, Factor1 essentially includes the students who drink alcohol, go out with friends, have a lot of free time and do not study a lot. This factor seems to include the subset of dimensions already considered with PCA for students who failed classes in the past. Factor2 highlights the group of students with highly-educated parents and few past failures: this reminds of the “red group”, i.e. students with high grades, at the top of biplots. Factor3 seems to naively represent only the students who failed in the past. Factor4 finds a common pattern between students who have a lot of free time spent with family and friends, and not on alcohol consumption.

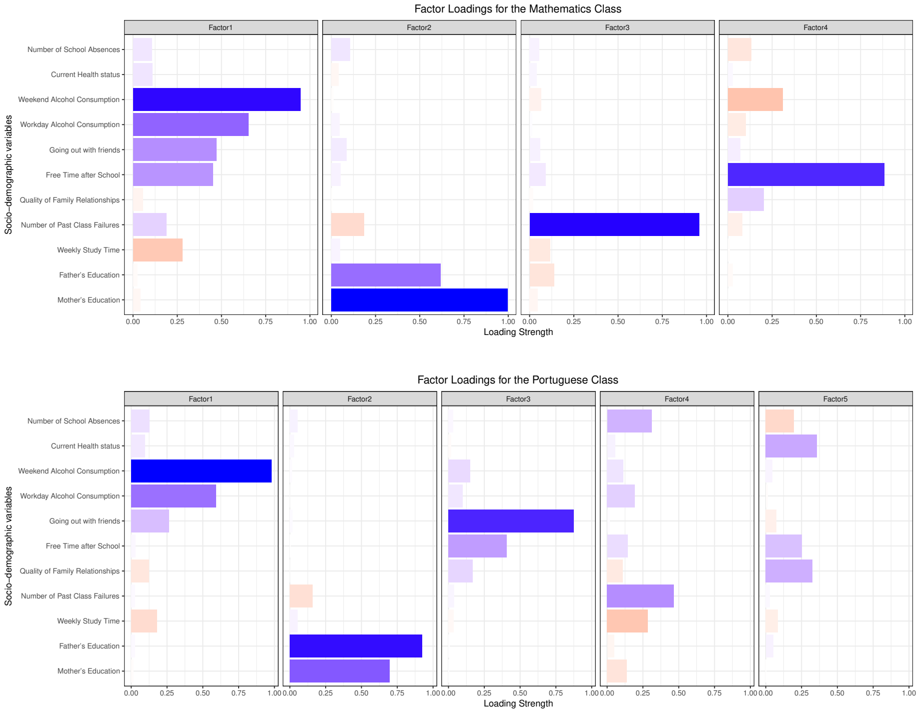


Figure 4: Factor loadings for the chosen 11 socio-demograhic variables.

In the Portuguese Class, Factor1 and Factor2 are really close to the Factor1 and Factor2 found in the Mathematics course respectively. Factor3 can be somehow related to Factor4 found in Maths class, with the only difference in the alcohol consumption. Factor4 and Factor5 are difficult to be interpreted, since all the weights are very small in absolute value.

2.2.3 Clustering Methods

Besides the factor loadings, FA also provides factor scores, which have been used to perform clustering. In Figure 5 the box plots show the distribution of the final grades (G3) among the K-Means clusters (after some tries, the best numbers are $K = 4$ for Math, $K = 5$ for Portuguese) obtained using original variables (left) and factor scores (right). The plot for the Mathematics class shows clearly how the clusters are more meaningful by providing as input the factor scores (4 columns for Mathematics, 5 columns for Portuguese), instead of the original variables (11 columns). ANOVA-test confirms that there is a statistically significant difference between the means of the clusters obtained using factor scores (in Math $F = 9.685$, $p = .0019$; in Portuguese $F = 15.42$, $p < 0.01$). However, such difference between clusters is validated by Kruskal-Wallis test [4] for Portuguese ($X^2(4) = 24.387$, $p < .01$) and not for Math ($X^2(3) =$

[4] Kruskal-Wallis test is the ANOVA non-parametric equivalent, used when ANOVA assumptions are not satisfied. Its null hypothesis is that the samples originated from the same distribution, which is slightly different from ANOVA null hypothesis.

5.2632, $p = .1535$).

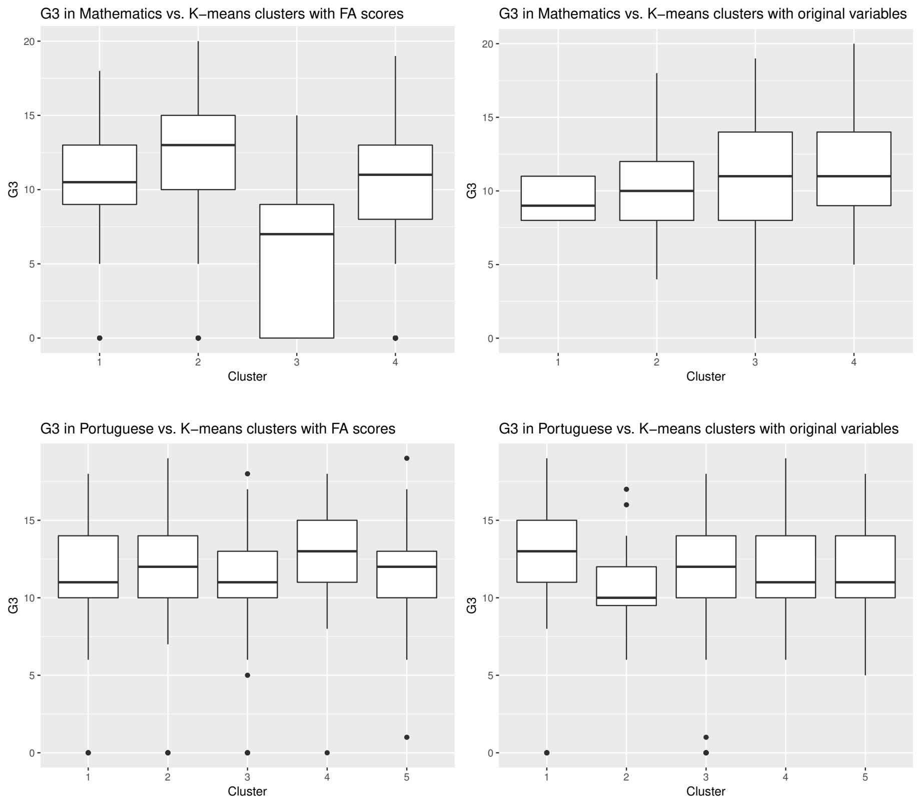


Figure 5: Distribution of the final grades among the K-Means clusters.

The cluster-analysis has been completed using other two clustering methods. Figure 6 shows the dendrogram obtained via hierarchical clustering using euclidean distance and 3 different linkage criteria (single, complete and average). By looking at the dendrogram obtained with complete linkage, the number of clusters seems to be the same as the number of factors.

Figure 7 shows the clusters obtained with K-Medoids algorithm (K=4 for Math, K=5 for Portuguese), which cannot easily separate students (clusters are highly overlapping).

Overall, the cluster analysis did not provide successful results, due to weak relationships between the variables. However, the combination of FA with K-Means provided useful insights on student's performance on the final period.

2.2.4 Lasso Technique

A series of features (e.g, absences, G1, G2, health) and the target variable (G3, the final grade) were used to train a series of Lasso models. These models were then used to predict either the final grade of a student, given that student's features, or if the student failed or passed the final period (two types of models

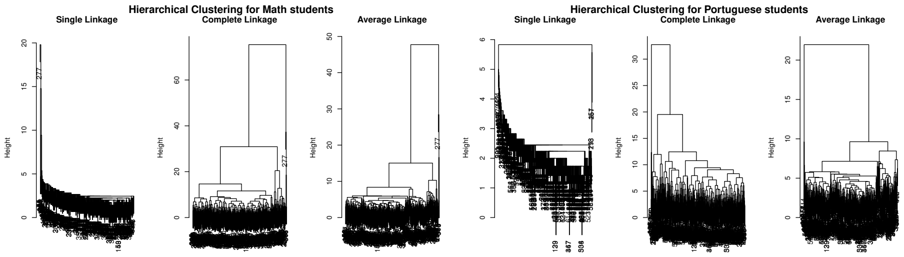


Figure 6: Dendrograms obtained using Hierarchical Clustering for both classes.

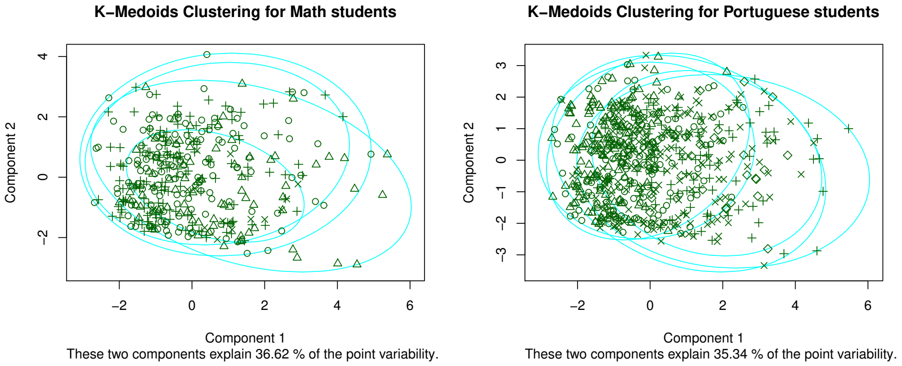


Figure 7: Clusters obtained using K-Medoids algorithm.

were trained, one for each kind of predictions). The used features for training varied for trained model to trained model. All the models and the results of the predictions can be seen in the repository[5].

A series of interesting remarks can be drawn when analysing the predictive capabilities of the models based on their training features. If we train the models to predict the final grade using as training data only qualitative (with the exception of *age*) and ordinal (with the exception of *traveltime*) variables we notice that the overall most relevant features are represented by the previous grades G1 and G2, which are then followed by the number of absences and by the family relationship of the student. When we predict both the final Portuguese and the final Mathematics grade then we can state the following (the examples from the parentheses stem form *Lasso technique on the data sets*).:

- If we can use only one grade as a feature in our prediction, then having the second Mathematics grade is more relevant (e.g., model 13, $\alpha = 0.1$, test data accuracy is 0.75 / 1) than having the second Portuguese grade

[5] https://github.com/andrei-comanescu/edm-2018-hy/tree/master/src , where the Jupyter Notebooks are *Lasso techniques on the data sets.ipynb*, *Lasso techniques also using factor variables.ipynb* and *Pass or Fail predictions via Lasso.ipynb*

(e.g. in model 12, $alpha = 0.1$, the accuracy rate on the test data is 0.41 / 1).

- If we have both Portuguese grades and the second Mathematics grade we don't have any significant prediction improvement (e.g., for the ninth model, $\alpha = 0.1$, the accuracy rate on the test data is 0.87 / 1) compared to the case in which we only have the second Mathematics grade and any of the two previous Portuguese grades (e.g, in model six we have the second Mathematics grade and the second Portuguese grade, the accuracy rate on the test data for $\alpha = 0.1$ is 0.85 / 1).

- If we have only the Mathematics grades the overall predictive score declines (e.g. in model 11 the accuracy rate on the test data, for $\alpha = 0.1$, is around 0.66 / 1) but not as much as in the scenario in which we only have the Portuguese grades (e.g., in model 10 for $\alpha = 0.1$ the accuracy rate for the test data is around 0.31 / 1).

- Without any of the previous grades the predictive power of the trained model is really low (e.g, for the fifth model the accuracy rate on the test data with $\alpha = 0.1$ is 0.07 / 1).

The previous conclusions hold true even when we include multiple factor variables in our set of features. Regardless of the used set of features, the overall grade prediction success rate is quite low as long as we don't include any of the previous grades. We finally tried to convert the grades from the scale of 0 to 20 to a binary one, in which we either marked the students as either failed (G3 $<$ 10) or passed (G3 $>=$ 10). Unfortunately even in that scenario the predictive power of the model depended quite heavily on the passed or failed status of the previous grades.

2.2.5 Decision Trees

We tried to see the relationship between failed / passed courses and other variables via decision trees. These were built for either subsets of the Portuguese data set variables or for subsets of the Mathematics data set variables[6]. The resulting decision trees with the explanation mentioning of the variables that were used in their case can also be found in the repository[7]. Some interesting things were noticed.

- In the case of the Portuguese data set these can be resumed as: if the education of the father is quite high or the family relationship is quite good then the student is most likely to pass the Portuguese class (e.g. if the father holds higher education that just one student out of 122 failed).

- Looking at the Mathematics decision trees we can see some trends: if the student hasn't previously failed (which is highly likely given the age of the students) then it is more likely to pass the course if the goal is to pursue higher education. Like in the case of the Portuguese class if the family relationship is good, then it is most likely that the student shall not fail

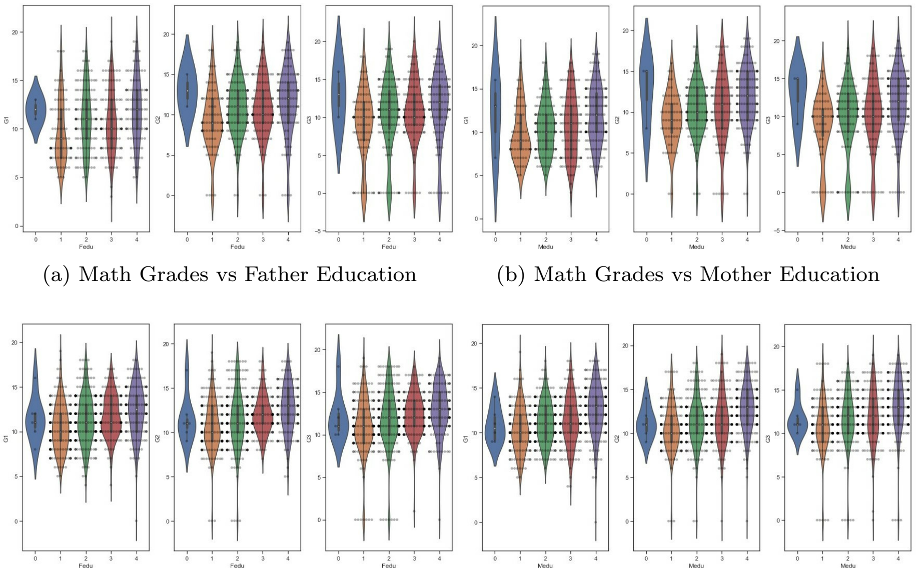
[6] the code is available here https://github.com/andrei-comanescu/edm-2018-hy/tree/master/src in the file *Decision Trees.ipynb*

[7] https://github.com/andrei-comanescu/edm-2018-hy/tree/master/decision%20tree%20visualisations

the class (e.g. if the family relationship is excellent then 34 students out of 34 pass the Mathematics class; if the family relationship is just good then 7 students out of 93 fail the Mathematics class).

For both classes the higher the level of education of the parents (especially the father) the less likely was for the student to fail the class. Students with good family relationships and / or an interest to pursuit higher education were more likely to pass the classes.

Similar results can be seen when testing whether the mean grades are the same among students with different parents' education (see Figure 8). ANOVA-test[8] rejects the null hypothesis for all grades and for both classes. Finally, in the Portuguese class the average number of absences is statistically significantly lower for students who want to take higher educations ($t = 3.04, p < .01$) and for students whose parents live together ($t = 2.49, p < .05$).



(a) Math Grades vs Father Education

(b) Math Grades vs Mother Education

(c) Portuguese Grades vs Father Education

(d) Portuguese Grades vs Mother Education

Figure 8: Grades vs. Parents Education.

References

[1] P. Cortez and A. Silva. Using Data Mining to Predict Secondary School Student Performance. In A. Brito and J. Teixeira Eds., Proceedings of 5th FUture BUsiness TEChnology Conference (FUBUTEC 2008) pp. 5-12, Porto, Portugal, April, 2008, EUROSIS, ISBN 978-9077381-39-7.

[8] The statistical units with *G2* and *G3* equal to 0 have been removed (38 and 16 students for the Mathematics and Portuguese class respectively) for these steps, in order to meet the assumptions of the test.

[2] https://archive.ics.uci.edu/ml/datasets/student+performance, accessed December 2018

A Variable explanation

The quantitative variables are:

- age - student's age (integer: from 15 to 22)
- absences - number of school absences (integer: from 0 to 93)
- G1 - first period grade (integer: from 0 to 20)
- G2 - second period grade (integer: from 0 to 20)
- G3 - final grade (integer: from 0 to 20)

The ordinal variables:

- Medu - mother's education (integer: 0 - none, 1 - primary education (4th grade), 2 5th to 9th grade, 3 secondary education or 4 higher education)
- Fedu - father's education (integer: 0 - none, 1 - primary education (4th grade), 2 5th to 9th grade, 3 secondary education or 4 higher education)
- traveltime - home to school travel time (integer: 1 - < 15 min., 2 - 15 to 30 min., 3 - 30 min. to 1 hour, or 4 - > 1 hour)
- studytime - weekly study time (integer: 1 - < 2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - > 10 hours)
- failures - number of past class failures (integer: n if 1 <= n < 3, else 4)
- famrel - quality of family relationships (integer: from 1 - very bad to 5 - excellent)
- freetime - free time after school (integer: from 1 - very low to 5 - very high)
- goout - going out with friends (integer: from 1 - very low to 5 - very high)
- Dalc - workday alcohol consumption (integer: from 1 - very low to 5 - very high)
- Walc - weekend alcohol consumption (integer: from 1 - very low to 5 - very high)
- health - current health status (integer: from 1 - very bad to 5 - very good)

The factor variables are:

- school - student's school (Factor with 2 levels: "GP" - Gabriel Pereira or "MS" - Mousinho da Silveira)
- sex - student's sex (Factor with 2 levels: "F" - female or "M" - male)
- address - student's home address type (Factor with2 levels: "U" - urban or "R" - rural)
- famsize - family size (Factor with 2 levels: "LE3" - less or equal to 3 or "GT3" - greater than 3)

- Pstatus - parent's cohabitation status (Factor with 2 levels: "T" - living together or "A" - apart)
- Mjob - mother's job (Factor with 5 levels: "teacher", "health" care related, civil "services" (e.g. administrative or police), "at_home" or "other")
- Fjob - father's job (Factor with 5 levels: "teacher", "health" care related, civil "services" (e.g. administrative or police), "at_home" or "other")
- reason - reason to choose this school (Factor with 4 levels: close to "home", school "reputation", "course" preference or "other")
- guardian - student's guardian (Factor with 3 levels: "mother", "father" or "other")
- schoolsup - extra educational support (Factor with 2 levels: yes or no)
- famsup - family educational support (Factor with 2 levels: yes or no)
- paid - extra paid classes within the course subject (Math or Portuguese) (Factor with 2 levels: yes or no)
- activities - extra-curricular activities (Factor with 2 levels: yes or no)
- nursery - attended nursery school (Factor with 2 levels: yes or no)
- higher - wants to take higher education (Factor with 2 levels: yes or no)
- internet - Internet access at home (Factor with 2 levels: yes or no)
- romantic - with a romantic relationship (Factor with 2 levels: yes or no)

For the combined files the additional features are:

- _x - information related to the Portuguese class
- _y - information related to the mathematics class
- classes - 0 if the student took both classes, 1 if the student took only the mathematics class and 2 if the student took only the Portuguese class