Danmarks Tekniske Universitet

DTU

02450 Introduction to Machine Learning & Data Mining

Report #1

Data: Feature extraction and visualisation

Authors

Dmitrij Mordasov - s203350
Mario Cesar Rodríguez - s203359

September 28, 2022

1 Dataset Description

The title of our dataset is *Student Alcohol Consumption*[1], which contains for a set of secondary school students their socioeconomic attributes along with their academic performance (and yes, their alcohol consumption). There are actually two datasets that are available, for both Maths and Portuguese students. Obtained from a survey, conducted due to the alarming low passing rate in particular group of students - in order to identify attributes that could be correlated to the student's final grades. Whilst completely unrelated to the field of study of the writers of this report, it is their hope that their common sense and personal experience will help them hypothesise about the insights gathered from the data.

In the original study [1], before fitting the models, some preprocessing was performed for their neural network (NN) and support vector machine (SVM) models. The nominal variables were transformed using *1-of-C* encoding and all attributes were standardized to a zero mean and one standard deviation. As a baseline comparison, a naive predictor was also tested. The authors of the paper suspected that the first period (G1) and second period (G2) grades had a high impact on the model's ability to predict a student's final grade (G3). Therefore, three input configurations were tested for a regression model. Configuration A was the ideal scenario, where both G1 and G2 were used to predict the final grade. Configuration B is the same as configuration A, but without G2. Lastly, configuration C is the same as configuration B, but without G2. The results showed that the A input configuration achieves the best results. The predictive performance decreases when the second period grade is not known (B) and the worst results are obtained when no student scores are used (C).

The results indicate that using only the last available evaluation is the best option for the Mathematics classification goals and Portuguese regression under the A input configuration. Lastly, the nonlinear function methods (NN and SVM) were outperformed by the tree based ones.

The authors of this report were inspired by the dataset to launch an investigation into how they, as parents, could structure their families and their relationships with their future children in a way that lead to (or is at least correlated with) their better academic performance. Thus, a collection of socio-economic factors that the authors deemed that parents may have a degree of control over was selected (13) from the larger attribute pool (33) of the dataset.

During the classification task the goal will be to classify a data point, such as classifying whether a particular type of socioeconomic background would result in a student either passing or failing the class.

Whilst previous student's performance (G1 and G2), an attribute that improves the model's ability to predict the final grade in a subject could be used, in this paper the focus is only

[1] kaggle.com/uciml/student-alcohol-consumption [Accessed on 05/09/2021]

on attributes which parents have an effect on, namely:

Investigated attributes	**Discrete / Continuous**	**Types of attributes**
address - Student's home address	Discrete	Nominal
Pstatus - Parent's cohabitation status	Discrete	Nominal
Medu - Mother's education	Discrete	Ordinal
Fedu - Father's education	Discrete	Ordinal
Mjob - Mother's job	Discrete	Nominal
Fjob - Father's job	Discrete	Nominal
famsup - Family educational support	Discrete	Nominal
paid - Extra classes within course subject	Discrete	Nominal
internet - Internet access at home	Discrete	Nominal
nursery - attended nursery school	Discrete	Nominal
famrel - Quality of family relationships	Discrete	Ordinal
Dalc - Workday alcohol consumption	Discrete	Ordinal
Walc - Weekend alcohol consumption	Discrete	Ordinal
Grade - Final student grade (G3)	Discrete	Ratio

Table 1: Attributes (reasonably) within parental control

The final grade (G3) is used as the final outcome in this report, due to its significance on the student's life (as the previous grades G1 and G2 were not deciding whether the student has passed or failed the exam in the end).

Categorical variables such as sex (M/F), parents' jobs and more were one-hot encoded (equivalent to 1-of-C encoding) into nominal boolean variables. The parent's education level was left untouched, although encoding could be considered for it as well in the next report. Based on a method in [1], the exam scores were transformed into a boolean pass/fail label, based on the criterion of achieving at least 10 out of 20 points. This can also be used for classification model training purposes in the next report.

2 Dataset Attributes

Investigated attributes description		Counts	Mean	Std
Address	Urban / Rural	452 / 197	-	-
Pstatus	Living together / Apart	569 / 80	-	-
Medu	Mother's education level[2]	-	2.51	1.13
Fedu	Father's education level[2]	-	2.31	1.10
Mjob	Home/health/services/teacher/other	135/48/136/72/258	-	-
Fjob	Home/health/services/teacher/other	42/23/181/36/282	-	-
famsup	Yes / No	398 / 251	-	-
paid	Yes / No	39 / 610	-	-
internet	Yes / No	498 / 151	-	-
nursery	Yes / No	521 / 128	-	-
famrel	Quality of family relationships (0-5)[3]	-	3.93	0.96
Dalc	Workday alcohol consumption (0-5)[3]	-	1.50	0.92
Walc	Weekday alcohol consumption (0-5)[3]	-	2.28	1.28

Table 2: Dataset attributes summary statistics

The Kaggle link provides a script in R, in which there is a set of attributes used to merge the two files, along with a statement that this way you get 382 common students. There is no mention of this in the original report [1], and the stated attributes used for this merging do not use all of the attributes in the dataset (13 out of 30). It was found that merging the Mathematics (n=395) and Portuguese (n=649) datasets according to these attributes did not allow for a fully intersecting dataset (e.g. whilst one entry shared attribute values across the 13 stated attributes, the values across the rest could have differed, with ones that should not have due to common sense). Thus, it was decided to disregard the advice on Kaggle on merging the datasets due to a lack of sources, information and reasoning behind this method.

The following analysis thus pertains only to the Portuguese exam dataset due to its larger size, keeping in mind that an analysis of the mathematics dataset could provide different insights into student performance, as various attributes could have differing effect with the different nature of the subject and possible study methods.

As can be seen in Table 2, that none of the numeric attributes are normally distributed because the mean values are not equal to zero and the standard deviations are not equal to one. Most of the students in the survey live in urban areas, have parents that live together, have family educational support, didn't have paid tutors, did have internet, and did attend nursery school. Additionally, most mothers had jobs related to working from home, services, or listed as other. Furthermore, fathers mostly had jobs related to services or listed as other.

[2] 0 - no previous education, 1 - primary education (4th grade), 2 - 5th to 9th grade, 3 - secondary education, 4 - higher education

[3] 1 - very low to 5 - very high

3 Data Visualisation

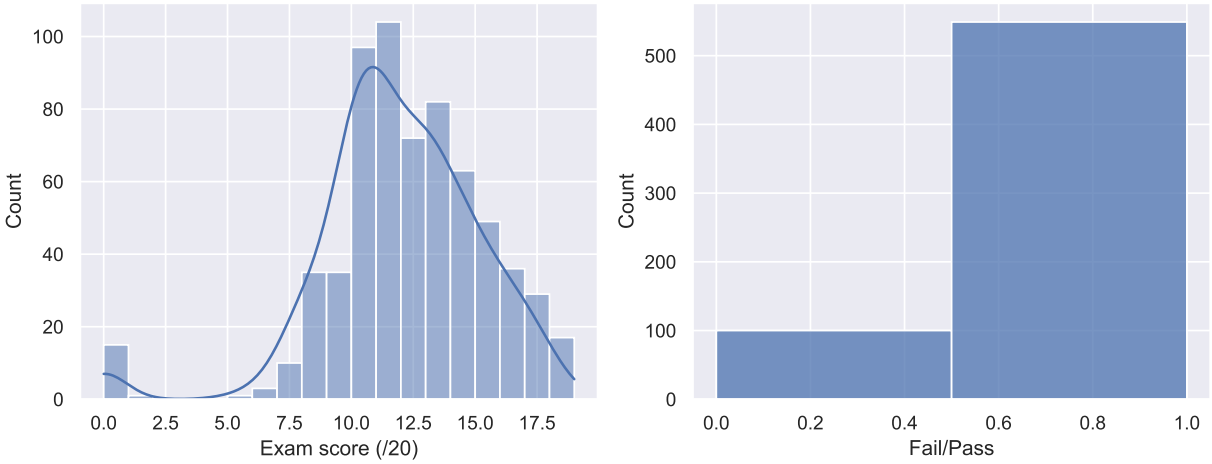

Figure 1: Grades distribution ($\mu = 11.9, \sigma = 3.23$), and transformed pass/fail data

Guessing from the peak in the left tail in Figure 1, it seems that some students who have gotten a zero score for the exam have simply not tried to write anything at all, possibly also explained by lack of exam scores between 1 and 5 - as this small space could also be caused by e.g. some exam questions being easy and doable by all. Save for this, the score distribution is not too entirely unlike a normal distribution.

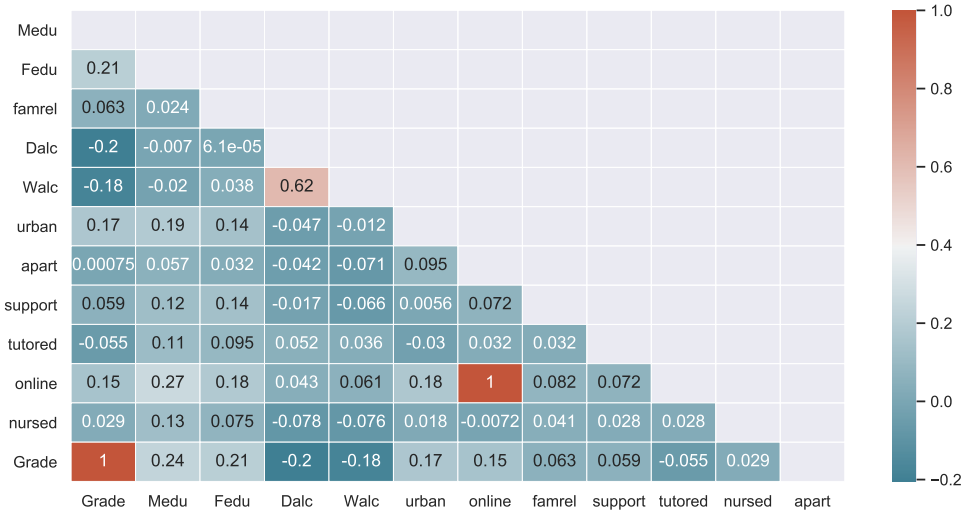

Figure 2: Data correlation heatmap

As can be seen in Figure 2, a sorting of the correlation matrix can show us (bottom row) the most correlated attributes compared with the students' grade. It can be seen that the parents' education, the student's alcohol consumption[4] and their address (whether they live in an urban environment or in a rural instead) are the more strongly correlated attributes. The nominal yes/no attributes were one-hot encoded for the purposes of this correlation analysis, and the second outcome removed (e.g. as a created attribute 'apart', signifying that the parents of the student live apart, has an equal and opposite correlation to other attributes as the one-hot-encoded attribute 'together' present in the attached Python code). Parent's job (a categorical variable with 5 different categories each) was also one-hot-encoded, but left out of this correlation analysis for the purposes of visualisation clarity.

Among the attributes themselves, there are some correlations that make a lot of sense - for example $\rho = 0.62$ of a student drinking on workdays and weekdays, $\rho = 0.18$ between living in an urban area and having access to internet (online), $\rho = 0.27, 0.18$ and $\rho = 0.19, 0.14$ of parents' education level and living with access to internet in an urban environment and more.

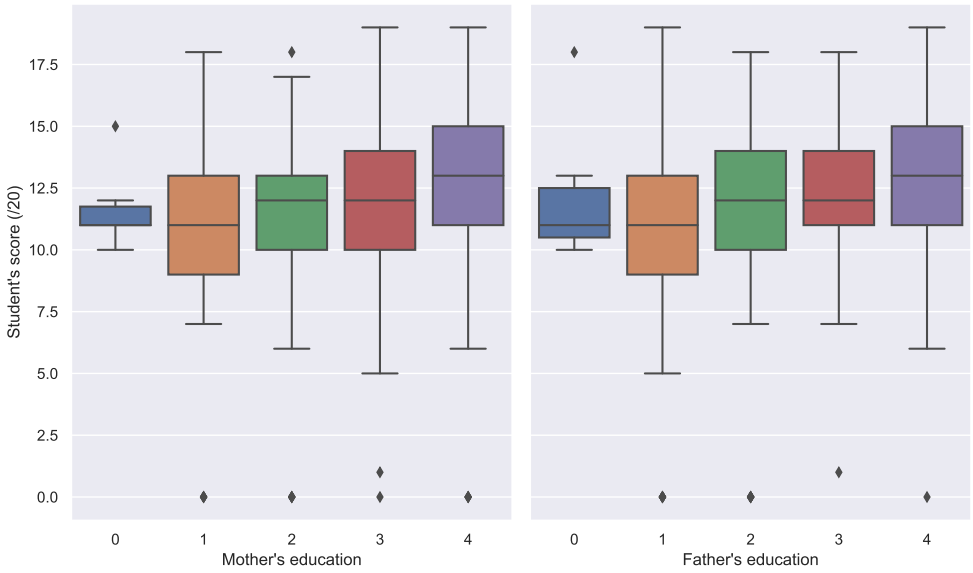

Figure 3: Parents' education and the student's grade

Through more exploratory data analysis in which box plots were built for all of the attributes and analysed, the parents' profession were found to have some impact on the student's score as well - for example as seen in Figure 4, having a father who works as a teacher led to highest mean grades. It is very interesting to see that children with parents who did not even complete primary school education seem to outperform those whose parents have at

[4]The authors were amused by the fact that alcohol consumption, whilst having a $\rho = -0.2$ for the Portuguese exams, were way less correlated with the Maths exam grade ($\rho = -0.05$), and are happy to have chosen a field of study more or less conducive to their hobbies.

least some, with this population almost completely passing the subject - this may be due to the generally known transformative effect of education on one's quality of life, hence those students might be extra motivated to perform well academically with respect to their peers.

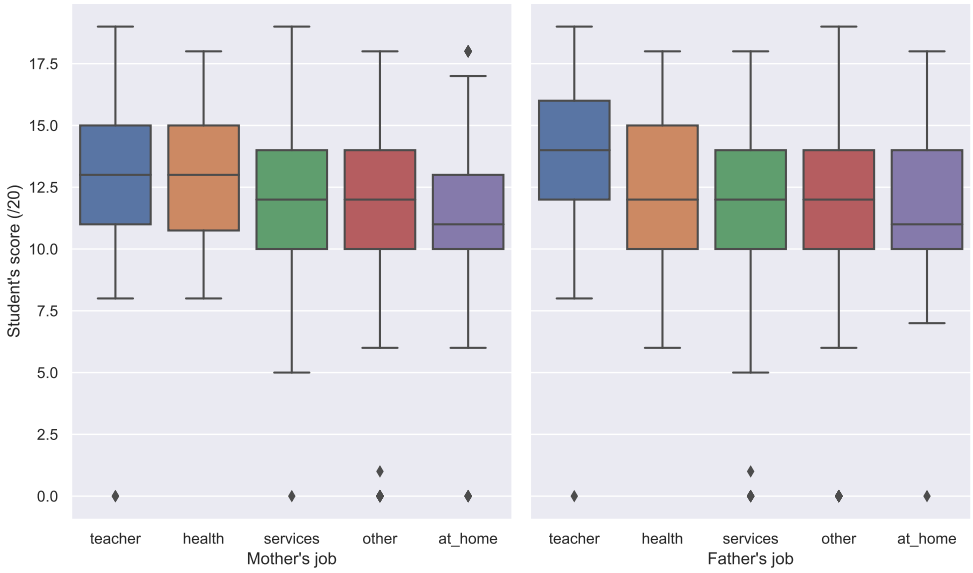

Figure 4: Parents' job and the student's grade

Through performing a principal axis analysis (PCA), using a threshold of 0.9 for variance explained, it has been found that the dimensions of the dataset could be reduced to 14 principal components, with 20 principal components accounting for almost all of the information. Due to the relatively small size of the dataset, dimensionality reduction does not seem necessary for the purposes of reducing the computing cost in later machine learning tasks.

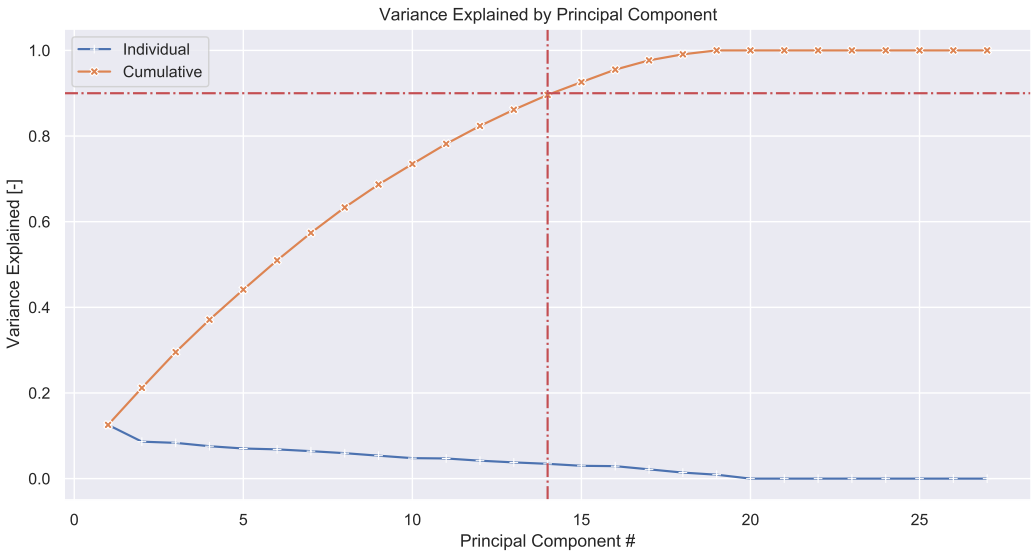

Figure 5: PCA variance explained

To find out the directions of the considered PCA components, the first 4 components of maximum absolute magnitude were extracted from the V vector for the principal axis in question. This way, some principal components relied on e.g. 1:[workday alcohol consumption, weekday alcohol consumption, if parents are together], 2:[if father works in healthcare, if mother works in healthcare, if there is no internet] and more. Intuitively, these can perhaps provide a bit of insight, yet the authors would rather to refrain from generalisations in this case.

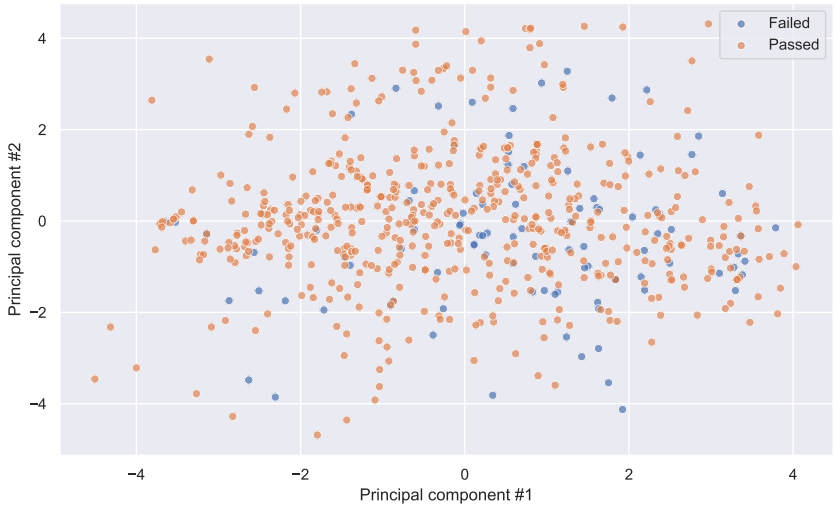

Figure 6: Data projection onto the PC axes

To be frank, it was quite difficult to find intuitive patterns of many of the data projections, no matter which principal axes were used. Were the correlations of the various attributes with the final grade more significant or the dataset larger, perhaps explainable relationships could be better extracted from the PCA.

4 Discussion

In this exploratory data analysis task, the authors have found a few interesting relationships between the attributes of socio-economic status of a student with relation to their academic performance.

Namely, the parents' highest level of achieved education seemed to be highly correlated to higher student performance. Then, factors such as the student's proclivity to consume alcohol (especially in the workdays), being situated in an urban setting (closer to the school) and having access to the internet were also identified to have more significant relationship with the student's grades.

As the relationships between the features and the grade weren't exceptionally strong (maximum correlation $\rho = 0.24$ with respect to the grade), it is the hope of the authors of the report that these attributes are sufficient to at least help with the classification task of identifying whether a student will pass or fail the final exam based on these. Other attributes from the dataset may have to be used in the next report in case the machine learning models based on this cut-out of the dataset do not perform adequately.

5 Problems

1. D) Time of day has an interval-like quality yet a value of 0 does not imply its non-existence and subsequent breaking of physics, the number of traffic lights and run over accidents are discrete with 0 having a physical meaning of no accidents, and the congestion level is a ranked categorical variable
2. A) as the infinity p-distance takes the form of $max(|x_i|) = max(7, 0, 2, 0, 0, 0, 0) = 7$
3. A) as the variance explained of the first four principal components was calculated to be equal to about 0.87 (0.29, 0.23, 0.20, 0.15, 0.13 for each PC)
4. D) as all the corresponding weights in the V vector to the attributes with a high numbers are positive (and by eliminating the other statements)
5. C) as there are 2 shared words (intersection) and 13 unique words in the two bags of words , hence the Jaccard similarity for this case is $\frac{2}{20000 - 11} = 0.0001000$
6. B) as $p(\hat{x}_2 = 0|y = 2) = 0.81 + 0.03$ from the table

Work distribution

	Task 1	Task 2	Task 3	Task 4	Questions
Dmitrij Mordasov	x	x	x	x	x
Mario Cesar Rodriguez	x	x	x		x

Table 3

References

[1] P. Cortez and A. Silva, "Using data mining to predict secondary school student performance," 2008. Proceedings of 5th FUture BUsiness TEChnology Conference (FUBUTEC 2008) pp. 5-12, Porto, Portugal.