SD201 - Mining of Large Datasets

Effects of alcohol on academic performance

Carlos Eduardo Jedwab
Daniel Victor Ferreira da Silva
Leonel Mota Sampaio Durão
Luiz Augusto Facury de Souza

1st December 2022

Contents

1 Introduction

This project has as its objective to study and analyze the effects of alcohol consumption on academic performance. For that, we have at our disposal a data set [1] originally acquired for a study to predict secondary school student performance [4]. The researchers have collected real-world data (e.g. student grades, demographic, social, school related features and alcohol consumption during workdays and the weekend), of which they were able to identify the key variables that affect educational success/failure and apply data mining models to predict a student's performance.

Our approach consists of three different methods:

- Clustering - Using KMeans++, we can study the data and observe how we could classify different group of students
- Forecasting - We aim to create a predictor for a student's performance given the available data
- Hypothesis Testing - Finally, we will try to answer the question "Students with a higher consumption of alcohol have lower average scores throughout the year?"

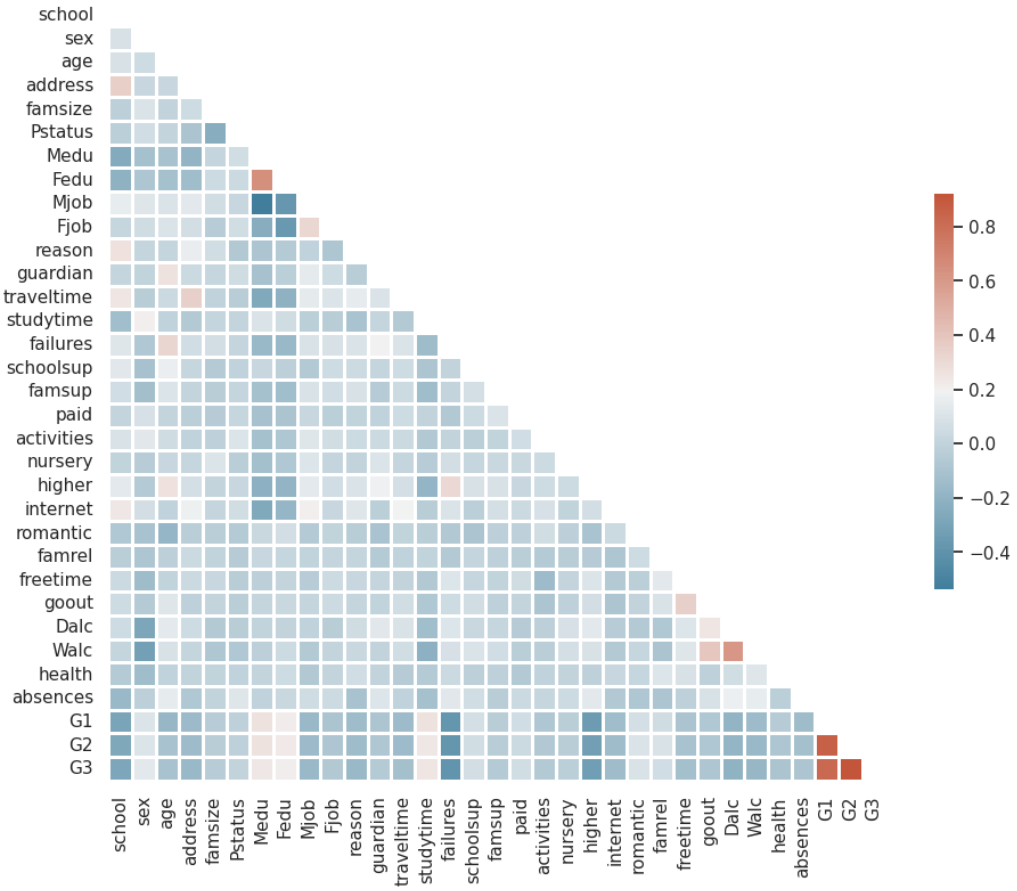

Figure 1: Correlation map between student variables.

Columns	Description
school	student's school (binary: 'GP' - Gabriel Pereira or 'MS' - Mousinho da Silveira)
sex	student's sex (binary: 'F' - female or 'M' - male)
age	student's age (numeric: from 15 to 22)
address	student's home address type (binary: 'U' - urban or 'R' - rural)
famsize	family size (binary: 'LE3' - less or equal to 3 or 'GT3' - greater than 3)
Pstatus	parent's cohabitation status (binary: 'T' - living together or 'A' - apart)
Medu	mother's education (numeric: 0 - none, 1 - primary education (4th grade), 2 â€“ 5th to 9th grade, 3 â€“ secondary education or 4 â€“ higher education)
Fedu	father's education (numeric: 0 - none, 1 - primary education (4th grade), 2 â€“ 5th to 9th grade, 3 â€“ secondary education or 4 â€“ higher education)
Mjob	mother's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at_home' or 'other')
Fjob	father's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at_home' or 'other')
reason	reason to choose this school (nominal: close to 'home', school 'reputation', 'course' preference or 'other')
guardian	student's guardian (nominal: 'mother', 'father' or 'other')
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)
studytime	weekly study time (numeric: 1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - >10 hours)
failures	number of past class failures (numeric: n if 1<=n<3, else 4)
schoolsup	extra educational support (binary: yes or no)
famsup	family educational support (binary: yes or no)
paid	extra paid classes within the course subject (Math or Portuguese) (binary: yes or no)
activities	extra-curricular activities (binary: yes or no)
nursery	attended nursery school (binary: yes or no)
higher	wants to take higher education (binary: yes or no)
internet	Internet access at home (binary: yes or no)
romantic	with a romantic relationship (binary: yes or no)
famrel	quality of family relationships (numeric: from 1 - very bad to 5 - excellent)
freetime	free time after school (numeric: from 1 - very low to 5 - very high)
goout	going out with friends (numeric: from 1 - very low to 5 - very high)
Dalc	workday alcohol consumption (numeric: from 1 - very low to 5 - very high)
Walc	weekend alcohol consumption (numeric: from 1 - very low to 5 - very high)
health	current health status (numeric: from 1 - very bad to 5 - very good)
absences	number of school absences (numeric: from 0 to 93)
G1	first period grade (numeric: from 0 to 20)
G2	second period grade (numeric: from 0 to 20)
G3	final grade (numeric: from 0 to 20)

Table 1: The preprocessed student related variables (Data set1s columns)

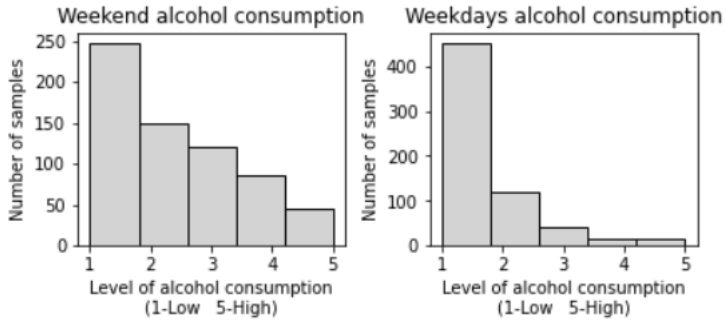

Figure 2: Alcohol consumption distribution for Portuguese students

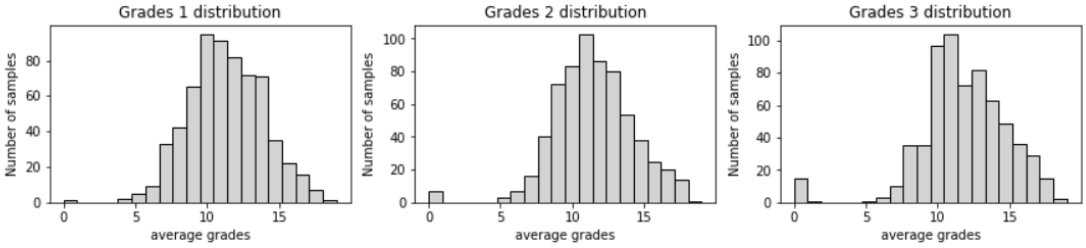

Figure 3: Grades distribution for Portuguese students

2 Data Analysis

2.1 The data set

The data was collected during the 2005-2006 school year from two public schools, from the Alentejo region of Portugal, where reports and questionnaires were answered in class by 788 students. However, in the original study, the researchers discarded 111 answers due to a lack of identification details (necessary for merging with the school reports). The data was then integrated into two data sets related to Mathematics (with 395 examples) and the Portuguese

language (649 records) classes. For both the files *Math.csv* and *Portuguese.csv* data set files, we have the student-related variables as seen in Table 1.

It is important to note that given the data from just two subjects from two Portuguese schools, we are faced with the limitation of a small data set. This prevents us from having a deeper analysis, such as analyzing the impacts on other subjects like arts and music, or generalizing these insights from data of a limited source as assertions over all the influence of alcohol consumption in academics.

The analysis is also limited by the diversity of courses in the data (just Portuguese and Mathematics). A greater range of courses could allow us to study with more depth how the alcohol consumption impacts differently between different disciplines. For example, it would be interesting to understand how creative disciplines such as arts and music have different impacts compared to sciences disciplines.

2.2 Data exploration

The data sets we adopted for our analysis, as explained by the data set construction process in the study, have no missing data and therefore eliminate any need for cleaning. Still, for each method used in the analysis, we applied specific data handling before running the code.

Moreover, with the correlations map in Figure 1, it is clear that there is a negative correlation between alcohol consumption (Dalc and Walc) and the final grades (G1, G2, and G3), as we initially expected. Another visible result is the strong positive correlation between the three student grades. That is also expected, since a student who goes well in the first exams, tends to be a student that will do well in all the exams.

Analyzing the distribution of the parameters that we are most interested in (alcohol consumption and grades), as shown in Figure 2, there is a unbalance of data, such that there are much more samples of students with low alcohol consumption than with high alcohol consumption. This is expected, since the average student will have low or medium levels of alcohol consumption, as opposed to students with high consumption habits, which represents the extreme cases. Given this, we divided the alcohol consumption for some analysis into two different groups with more balanced sizes. The first of low consumption (values 1 and 2), and the other of high consumption (values 3,4, and 5).

As illustrated in Figure 3, the three periods' grade distributions are close to a normal distribution. With only a few outlier samples of grade 0 or 1, increasing little until G3, but after some tests removing these outliers, there were no significant improvements in the results.

Finally, plotting the average grade distributions for both Portuguese language (Figure 4) and Math (Figure 5), and comparing the distributions of students with high alcohol consumption, levels 3 or higher, or low, levels 1 or 2. We can clearly see how the cases with high alcohol consumption generate a distribution with lower variance and lower mean grades.

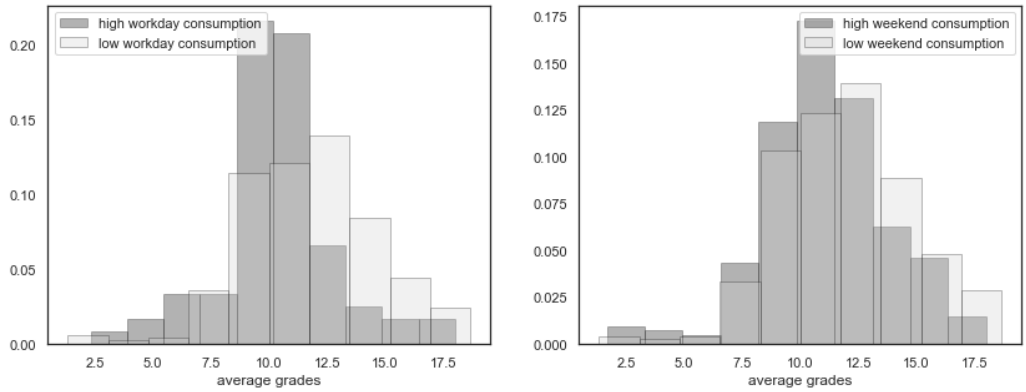

Figure 4: Average Portuguese grades distribution for low and high alcohol consumption on workdays(Left) and weekends(Right).

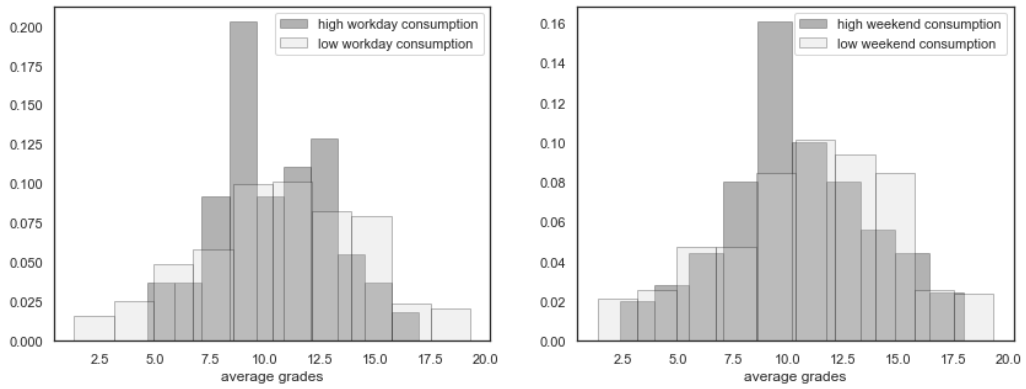

Figure 5: Average Math grades distribution for low and high alcohol consumption on workdays(Left) and weekends(Right).

3 Methods

3.1 Clustering

One of the important steps while trying to understand the data is to see if it has any sort of natural separation between different classes. For this reason, we've used the KMeans++ algorithm to cluster out data. First, it was important to normalize the data and to transform categorical data into numerical with the one-hot encoding technique.

In the KMeans algorithm, the most important hyperparameter is the number of clusters. To decide that we calculate the inertia for different numbers of clusters and by the elbow method, we keep the one after the steepest descent. With the plot it's possible to see that there's now a clear elbow and the inertia continues to lower after a very large number of clusters. We don't want a very large number of clusters, since we can't get any useful information out of this. So a good number is around 10 clusters. In the plot with each cluster represented by a different color, it's possible to see that there's no clear separation between the classes. This fact is also indicated by the large inertia with this number of clusters.

We thought that one of the causes of the lack of clear separation of the data might be because of the large number of dimensions that the data have, 33 with the one hot encoding. For this reason, we tested to reduce the dimensions of the data with Principal Component Analysis (PCA) before applying the clustering algorithms. To choose the number of components to retain, we've used the cumulative variance. With around 20 components, we have a large cumulative variance, so that's the chosen number. It's possible to see that the inertia diminishes

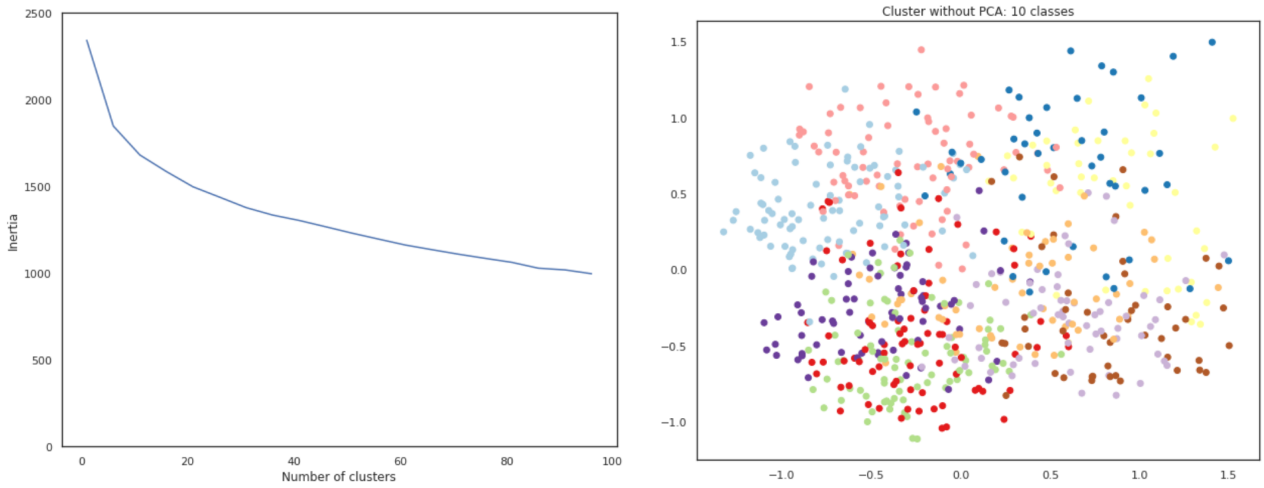

Figure 6: Inertia for PCA and initial clusters.

a little by applying KMeans after the PCA. However, the form of the curve continues the same and we choose the number of classes to be 10.

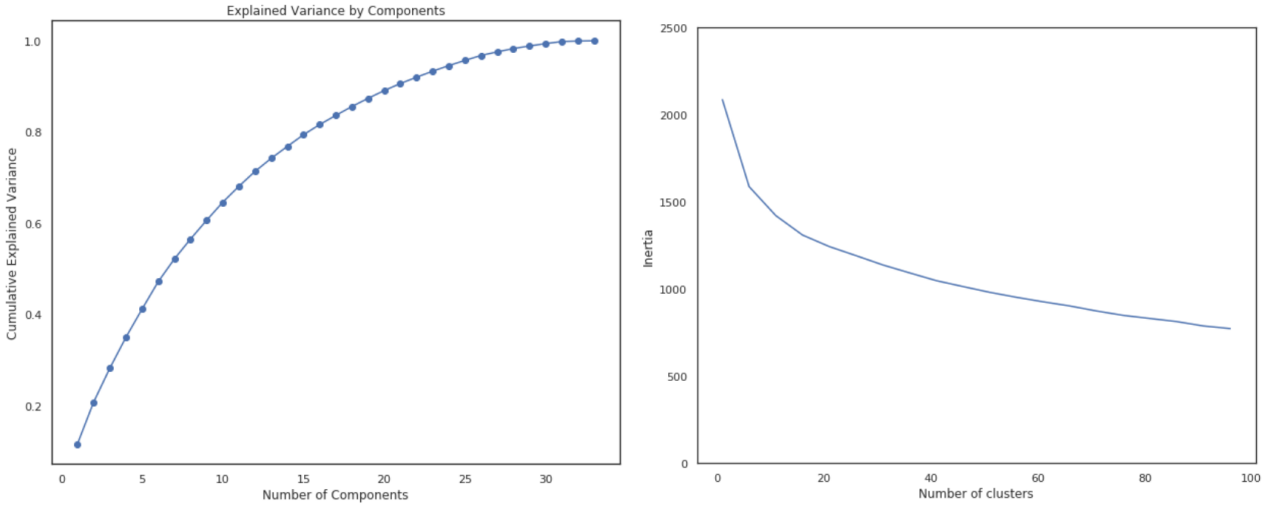

Figure 7: Inertia for PCA and initial clusters.

The results obtained remain largely the same and we can conclude that the data is not very well clustered. Using all the attributes, the students can't be separated into different distinct groups very well.

3.2 Forecasting

In order to test if the data gives enough information to predict the scores and if Dalc and Walc are relevant to the forecast, different prediction models were tested and optimized using randomized search cross-validation with 10 folds in the Portuguese data. In order to compute the experiments, the variables 'G1', 'G2', and 'G3', referring to the exam's scores, were replaced by 'G', which is the average of the 3 variables. Then, the values were normalized and, for each model, 100 different combinations of hyperparameters were randomly chosen and tested in order to obtain the combinations which results in the smallest mean squared error. After finding the values were normalized and divided into train set and test set with a proportion of 75% train and 25% test. The models with the best hyperparameters combination and their error metrics are shown in the following table, where:

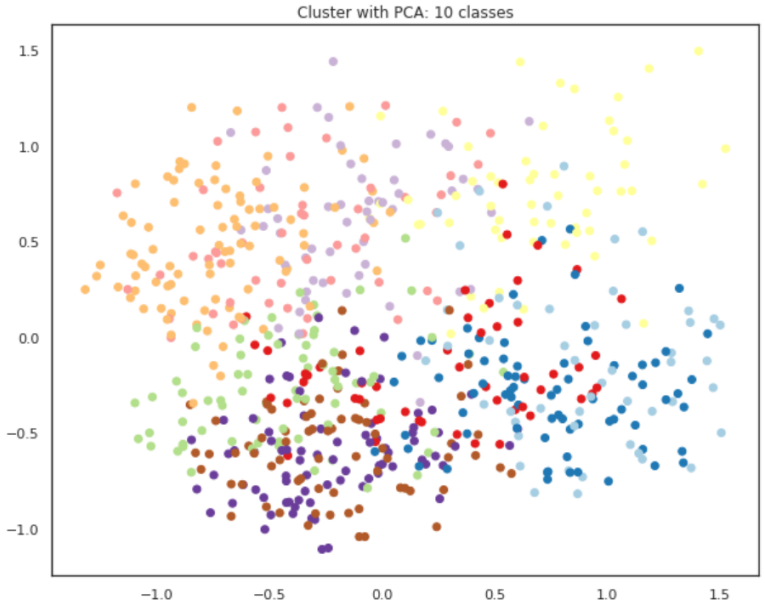

Figure 8: Inertia for PCA and initial clusters.

- LR: a simple linear regression, used as a baseline.
- MLP: a Multilayer Perceptron[6], which is a neural network with 1 or more hidden layers. Differently from the Perceptron, the MLP's hidden layers allow the model to separate non-linear data, since the hidden layers project the data into another dimension. This model can be very powerful, as it can approximate any continuous nonlinear function.
- XG: xgboost is a regularized version of the classic gradient boosting [5], introduced in 2001 as a boosting of decision and widely used today, although the idea of boosting is much older. This algorithm is considered to be very powerful for tabular data [7], so it was tested in the experiments.
- RF: the random forest algorithm [2], as we've seen in class, is another powerfull model for tabular data, a begging of decision trees.
- SVM: A Support Vector Machine[3] is an algorithm that tries to maximize the margin of the line that separates the classes. It is extremely used due to its simplicity and good enough results.
- KNN: A K-Nearest Neighbour is a simple algorithm that finds a separation based on the distance of the samples in the distance. It also yields a fairly consistent result with a low training time.
- DT: Finally, a Decision Tree was tested in order to compare its results with the XG and RF methods.

	LR	MLP	XG	RF	SVM	KNN	DT
MSE	6.97	6.89	6.73	6.82	7.13	7.52	7.31
MAE	1.97	1.91	1.94	1.95	1.99	2.13	2.05

Being the MSE and MAE formulas:

$$MeanSquaredError = \sum_{i=1}^{D} (x_i - y_i)^2$$

$$MeanAbsoluteError = \sum_{i=1}^{D} |x_i - y_i|$$

Since the xgboost method resulted in the smallest MSE, it was used in the 4 following analysis:

1. Prediction of G, the average of G1, G2 and G3, using all the variables.
2. Prediction of G using only variables that have a correlation with G superior to 0.1 in absolute value.
3. Same experiment as 2, but removing also Dalc, Walc, and the variables which have correlation, in module, inferior to 0.1 in relation with Dalc and Walc. This experiment aims to observe if, by removing Walc, Dalc, and its correlated variables, the forecast will be
4. Prediction of G3 with all the variables and the average of G1 and G2

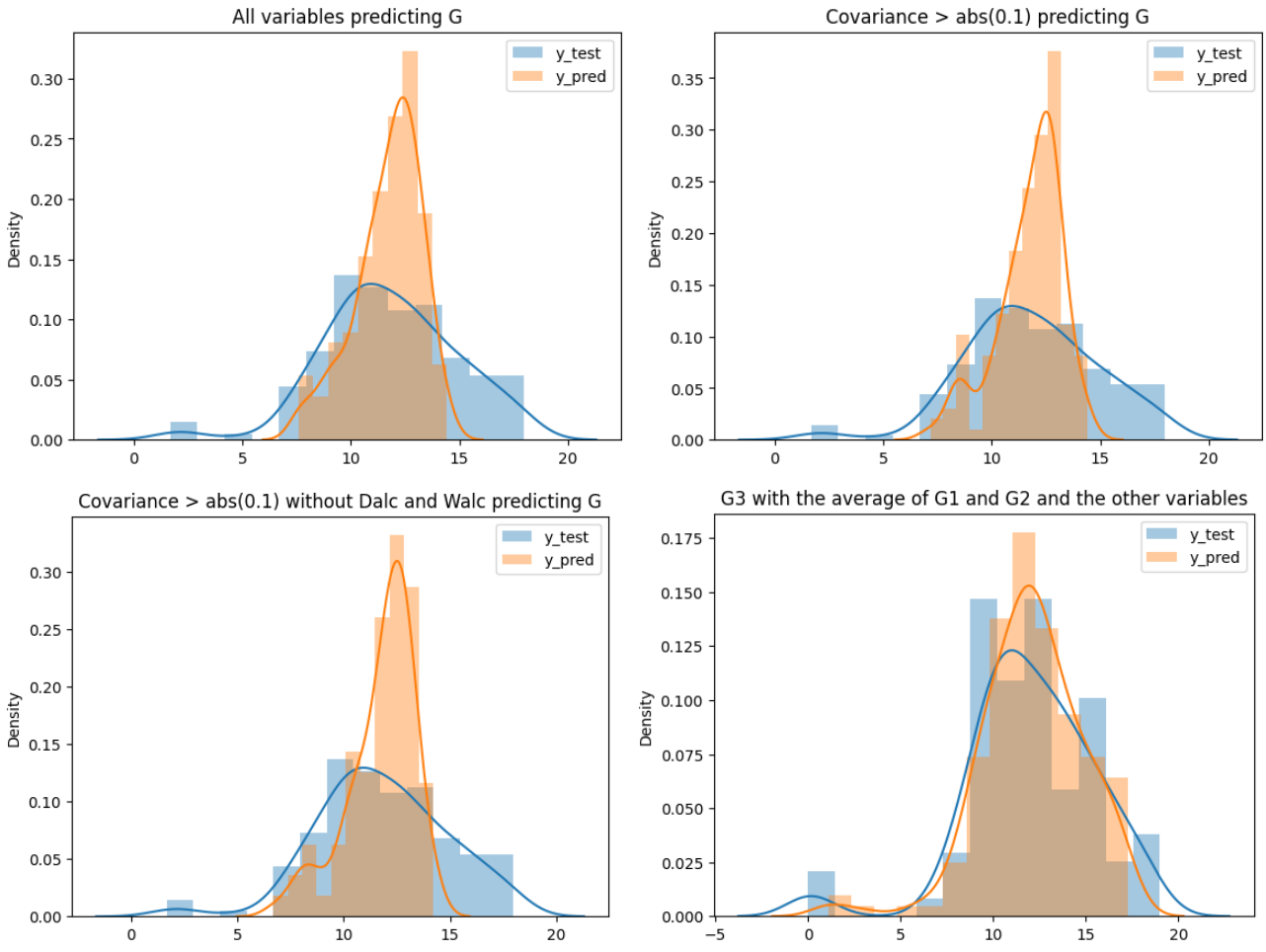

Figure 9: Estimated distribution of predicted and original target.

	1	2	3	4
MSE	6.73	7.17	7.30	2.61
MAE	1.94	1.97	1.99	1.08

In spite of the reduction in the dimensionality of the data observed in experiment 2, the error increased, so this indicates that the variables that were removed are important for the forecasting. In experiment 3, the error also increased, so the alcohol consumption variables also are important for the model. Finally, when a score variable is used along with the others in order to predict another score, the error decreases drastically, as expected, since the correlation is extremely high.

3.3 Hypothesis Testing

To show how alcohol consumption impacts academic performance, one alternative we took is a hypothesis test. More specifically, we modeled A/B and permutation tests, to answer the sub-question: "Students with a higher consumption of alcohol have lower average scores throughout the year?".

Assuming the null model that the average grades of high and low consumption are equal, and the null hypothesis that the average grade of students follows a uniform distribution, we want to reject the null hypothesis. The analysis intends to investigate these trends in grades by levels of consumption related to the different disciplines, Portuguese and Math, and different timeframes of consumption, during workdays or weekends.

Therefore, for the permutation test, we compare the difference between the averages of high and low alcohol consumption in the original data and the distribution of the difference throughout 5000 permutations of the consumption groups labels, for the null world with the average following a uniform distribution. As for the A/B test, we compute the Confidence Interval of the average grades for each consumption group through bootstrapping.

The samples were split between high and low consumption, with low consumption defined by levels 1 or 2 and high consumption by levels 3 or higher, given the scale from 1 to 5 used to create the dataset.

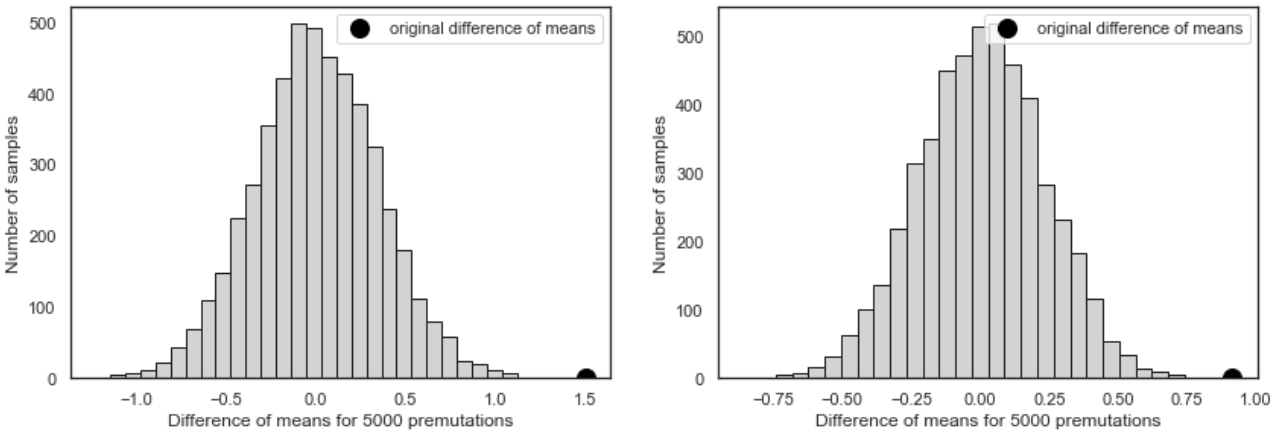

Figure 10: Distribution of the difference of means of Portuguese of permuted groups of consumption during workdays(left) and weekends(right).

Figures 10 and 11 show the results of the tests for Portuguese grades and level of consumption during workdays and weekends. The permutation test shows that, for both timeframes, the observed values for the difference in mean between the consumption groups, 1.5 and 0.91, are large enough to reject the null hypothesis, average grades of both groups from the same distribution, represented by the histogram of permutations. The A/B test also allows rejecting the null hypothesis, where the restrict 99% confidence intervals computed for the average grades of the different groups of consumption during workdays ([9.5,11] and [11.5,12.1]) and weekdays ([10.6,11.5] and [11.6,12.3]) clearly show no intersection between the intervals.

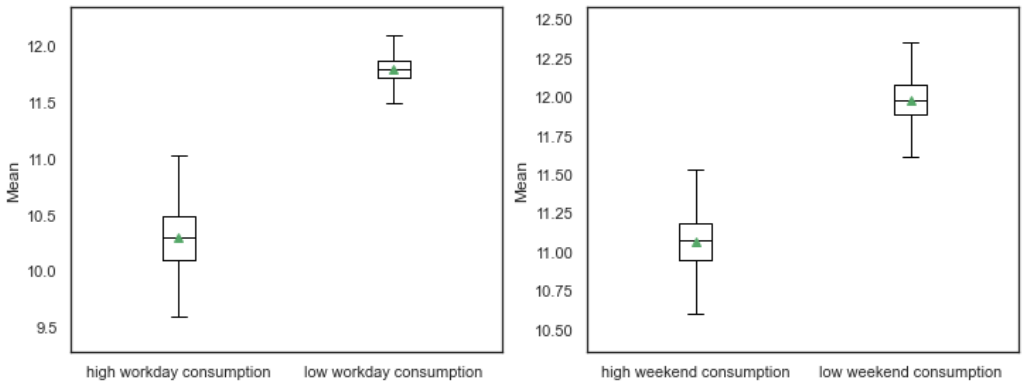

Figure 11: 99% Confidence Interval for the average grades of Portuguese for each alcohol consumption group.

Furthermore, analyzing workdays against weekends levels of consumption, the results positively support the conclusion that alcohol consumption during the week is more harmful to academic performance, with a larger difference in means for weekdays consumption groups, and increasing alcohol consumption during weekdays has a bigger impact on academic grades than weekends.

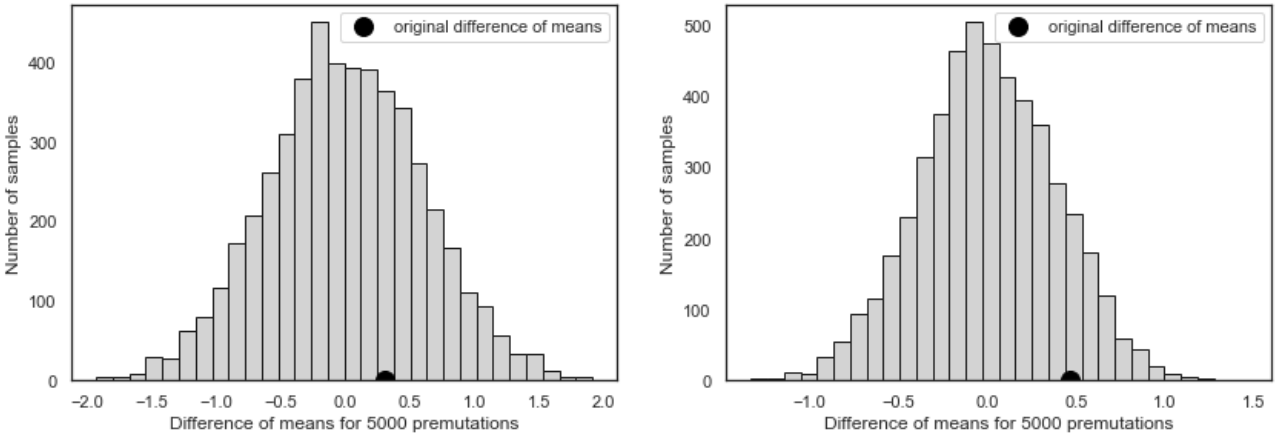

Figure 12: Distribution of the difference of means of Math of permuted groups of consumption during workdays(left) and weekends(right)

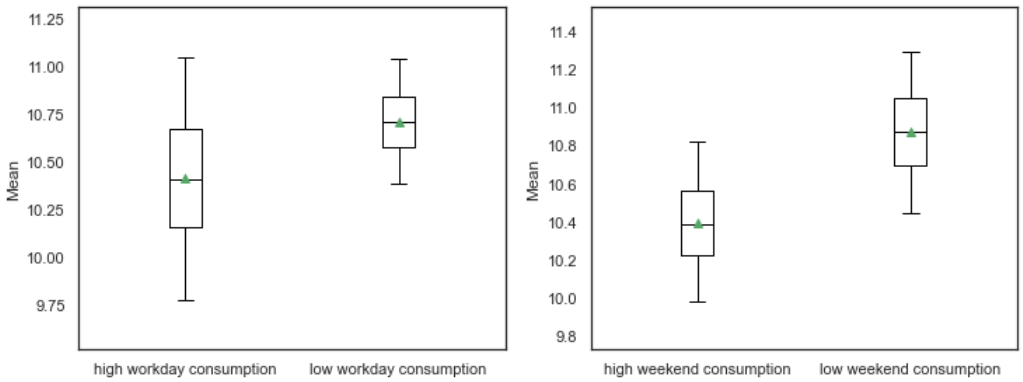

Figure 13: 90% Confidence Interval for the average grades of Math for each alcohol consumption group

On the other hand, results of both the tests shown in Figures 12 and 13 illustrate that the

different group means in the average grades of Math can not be certainly assigned to different distributions, therefore, we can not reject the hypothesis of both following the same distribution. Even with a difference in the average grades still noticeable in favor of the low consumption, it is not as expressive as for Portuguese grades, having an expressive interception of confidence intervals even with lower confidence (90%).

Analyzing levels of consumption in the different timeframes, we also can not take the same expected conclusion of greater impact on weekdays consumption, as the computed difference in means (0.3 and 0.47) was greater for weekends. Without any expressive difference from Portuguese grades data, the contrary results to the initial hypothesis could be explained by limitations on the sample size, almost half the number of Portuguese grades.

4 Conclusion

The hypothesis testing showed that it is very likely for alcohol consumption to be harmful to academic performance. Especially for consumption during workdays. A possible explanation is that a student that drinks a lot during the week, tends to not have much quality studying time during the week. On the other hand, drinking on the weekend has less impact on the study time quality. So this proposition suggests that, in fact, alcohol consumption harms students' performance.

However, another viable proposal as this one is that "good students", students with higher grades, will tend to not drink too much on the weekend, and even less during workdays. Therefore, alcohol consumption is a consequence of the dedication of the students. It could also be that there is another underlying factor that directly impacts both a student's performance and drinking habits. Another possibility, the most plausible one, is that there is a combination of all these three proposals.

Finally, the prediction model showed evidence at least for the implication that high alcohol consumption impacts performance (as this was the question we have proposed in this project). Assuming this is true, we can explain why removing the alcohol consumption variables from the model implies a larger error.

The workload was distributed as follows: Carlos, data analysis; Daniel, hypothesis testing; Leonel, clustering; and Luiz, forecasting.

References

[1] "Dataset," https://archive.ics.uci.edu/ml/datasets/Student+Performance, accessed: 2022-11-21.

[2] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[3] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.

[4] P. Cortez and A. M. G. Silva, "Using data mining to predict secondary school student performance," 2008.

[5] J. H. Friedman, "Greedy function approximation: a gradient boosting machine," *Annals of statistics*, pp. 1189–1232, 2001.

[6] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning internal representations by error propagation," California Univ San Diego La Jolla Inst for Cognitive Science, Tech. Rep., 1985.

[7] R. Shwartz-Ziv and A. Armon, "Tabular data: Deep learning is not all you need," *Information Fusion*, vol. 81, pp. 84–90, 2022.