Effects of alcohol consumption in student life

Craciun Eugen Mihai
Iancu Robert
Melinte Tudor Matei

Introduction

We are students. Every student drinks alcohol in a certain amount. This is the period in our lives when we want to grow stronger and we want to learn more to become smart, independent individuals that succeed in their lives . This is why, after searching for an interesting topic, we decided to go with the Student alcohol consumption dataset.

Our purpose is to try to understand the data, hidden relationships between features, and answer questions like, how will my alcohol consumption affect my grades, does it affect them at all? How is the importance of other factors like family status, free time, age, sex, compared to the importance of alcohol in a student's life.

For this, we are going to use Machine Learning techniques, to mine the data and reveal the hidden patterns. We are going to break the process into three phases, each one of us dealing with one. First, someone will try to plot and analyse the data to bring any interesting visual patterns we might get, but also to get a first insight of the dataset content for a better exploration. Then, we will try to classify the alcohol consumption (rated from rare to often on a scale up to 5), given different features and using different machine learning algorithms. This would let us see if the given data is enough to determine roughly how much alcohol does a student drink, given his circumstances. Last but not least, we will try the other way around. We will try to predict a student's grade, given his status, and especially, his alcohol consumption.

In this process, we are going to use different python technologies, like numpy, matplotlib, sklearn, pandas which will allow us to work with powerful machine learning algorithms, and do a relevant research of the topic we decided to go work with. First, we will decide on general aspects, like who will do which task, and the minimal common environment we will use, like a common github repo. Next, every member of the team will lead an individual task and we will commonize the results in the end in this paper.

Dataset analysis

Before performing any action specifically to modeling, we must do an analysis over our data. A better understanding, about the problem we are trying to solve and potential solutions, arises after a minimum inspection realised by a human.

In this project we try to find the best implementation for a solution which can predict the level of daily alcohol consumption for a student. The dataset we used contains data about students from two different educational institutions, students attending to math and/or portuguese language courses. This dataset offers us a lot of informations about each student, some of them even too personal.

The author of this dataset divided the alcohol consumption feature in two different columns. One is called *Dalc (Workday alcohol consumption)* and the other is *Walc (Weekend alcohol consumption)*, both have values between 1 (very low) - 5 (very high) which represent the consumption level. I guess that the main reason for building 2 different columns is due to the fact that most people are predisposed to consume more alcohol in weekend than in workday.

For most of my work, in this project, I took the decision to make an plot for each analysis of the two targets, Dalc and Walc. Another option was to combine these data, but I've thought that there is a good reason for keeping them separately. Combining them, we risk to lose paths in our search.

A quick look over the dataset

The dataset is composed of two CSV files, *student-mat.csv* and *student-por.csv,* which contains samples of data about students who are studying math and portuguese. As you can see in the below picture, the *student-mat.csv* contains about 395 rows and *student-por.csv* around 649 rows, both having 33 columns (features). The owner of dataset warns us that some students can be enrolled in both courses.

I decided to concatenate these datasets into one. For the students which are enrolled in both courses, I chose to keep just the samples from math course, due to the inequality of these datasets. After concatenation, we remained with a number of 662 students.

As a good practice, we should check the number of values in our dataset which are equal to *NaN*. To my surprise, we were lucky enough to have not even one *NaN* in our dataset. Missing data can cause a lot of complications for both data analysis and training a model.

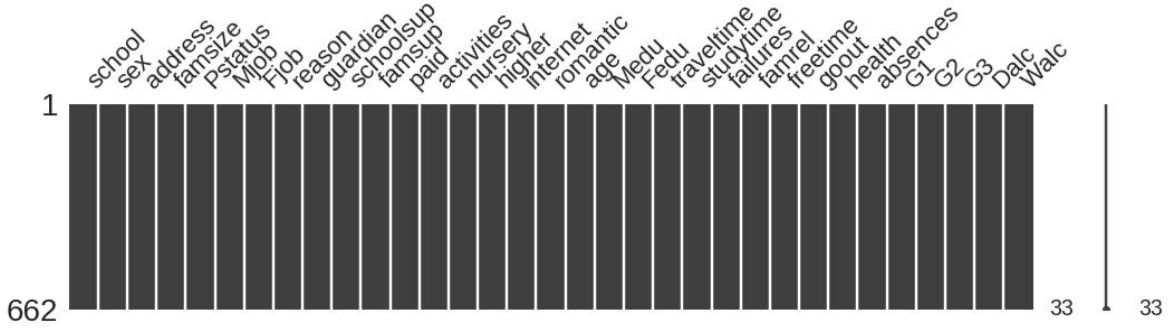


Figure 1: Number of present data for each feature

Although there were categorical features in a high proportion, I've decided to apply feature analysis techniques, specifically to continuous variables, to the data which were represented by natural numbers (between 1 - 5, the exception being the grades, which are between 1 - 20).

```
categorical_features = [
    'school', 'sex', 'address', 'famsize', 'Pstatus', 'Mjob', 'Fjob',
    'reason', 'guardian', 'schoolsup', 'famsup', 'paid', 'activities',
    'nursery', 'higher', 'internet', 'romantic'
]
numerical_features = [
    'age', 'Medu', 'Fedu', 'traveltime', 'studytime', 'failures',
    'famrel', 'freetime', 'goout', 'health', 'absences', 'G1', 'G2',
    'G3'
]
```

Regarding the fact that we have two targets, would be indicated to look over them and analyse the differences and similarities. As we were expecting there is a higher level of alcohol consumption in weekend, than in workday. The main reason would be the

responsibilities that the students have during the workday, you can't accomplish your tasks as good as being sober.

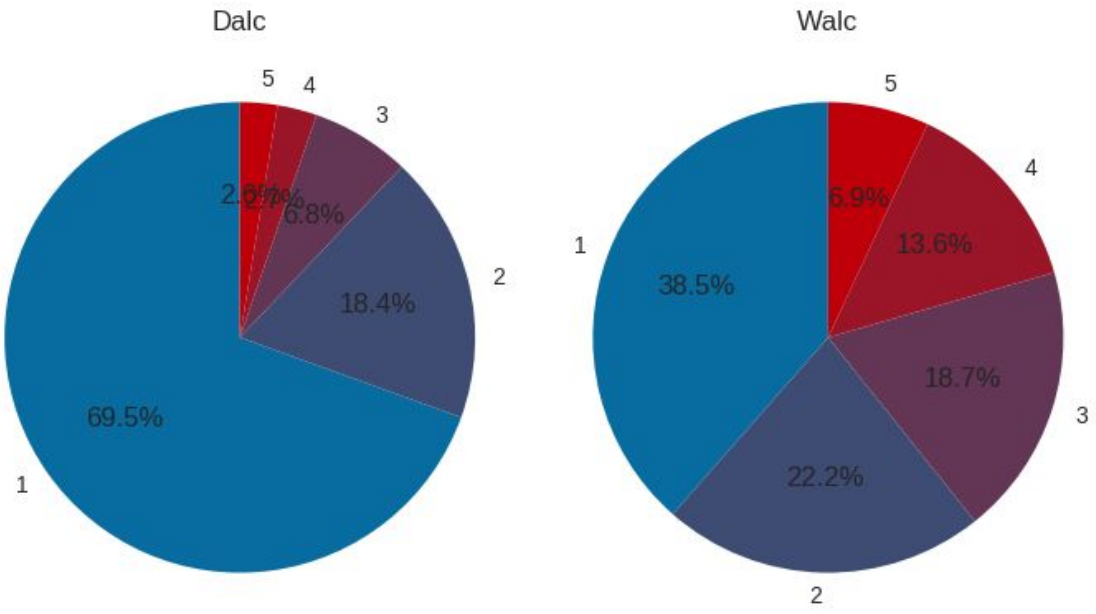


Figure 2: Percentage of alcohol consumption stages

The next thing which we are interested in is the distributions of alcohol consumption regarding the age of the student. During the whole analysis, alcohol consumption will lead in Walc. From the below plot we can notice that there is a pick for alcohol consumption, regarding the age, and that age is *"the wonderful age of seventeen years"*.

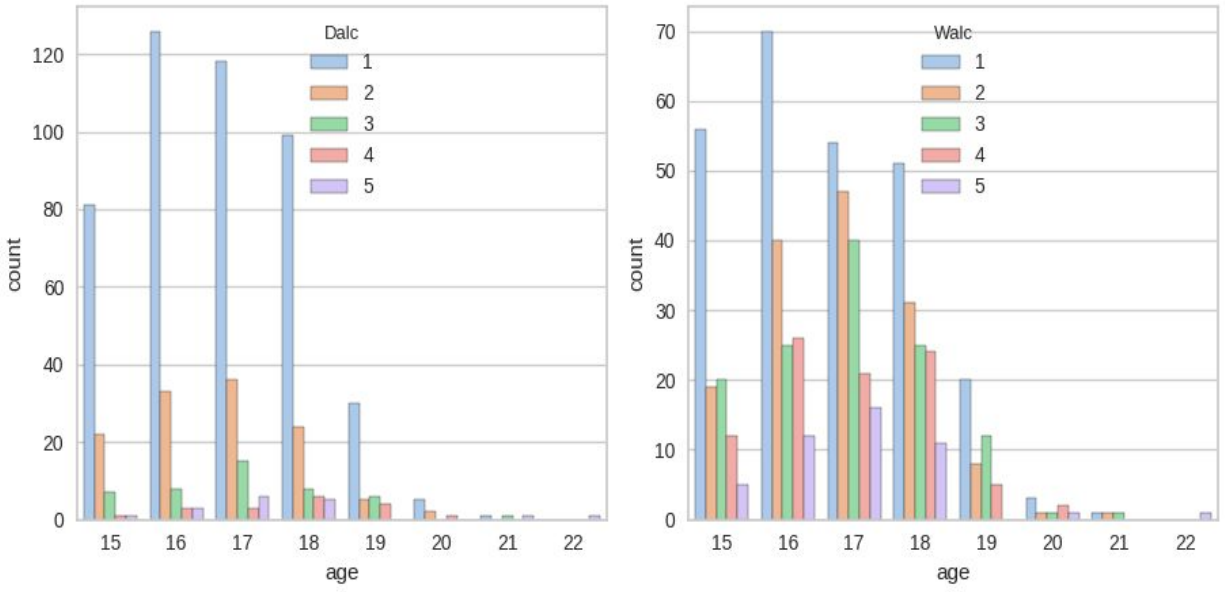


Figure 3: Density distribution on ages for each alcohol consumption value

A general analysis over the whole dataset

To ease the process of training a model, regarding the time spent on training and the complexity of it, it's required to do a general feature analysis, followed by a feature selection. After analysing all features we can conclude to a list of sufficient and necessary features. Simplification of model brings a lot of benefits, like increasing the ease of interpreting the model or avoiding the use of some dimensionalities.

The principal techniques I used are ranking in one and two dimensional for numeric features and Camér's V correlation for categorical features. First two were easy to generate due to a little help from the *yellowbrick* package, the last one, Camér's V correlation, required the implementation of the algorithm.

A one dimensional ranking of features

The main difference between *Rank1D* and *Rank2D* is the number of features used to rank the features. Rank1D use the Shapiro-Wilk algorithm to evaluate the normality of the distribution, for each feature, regarding the values specifically to the feature.

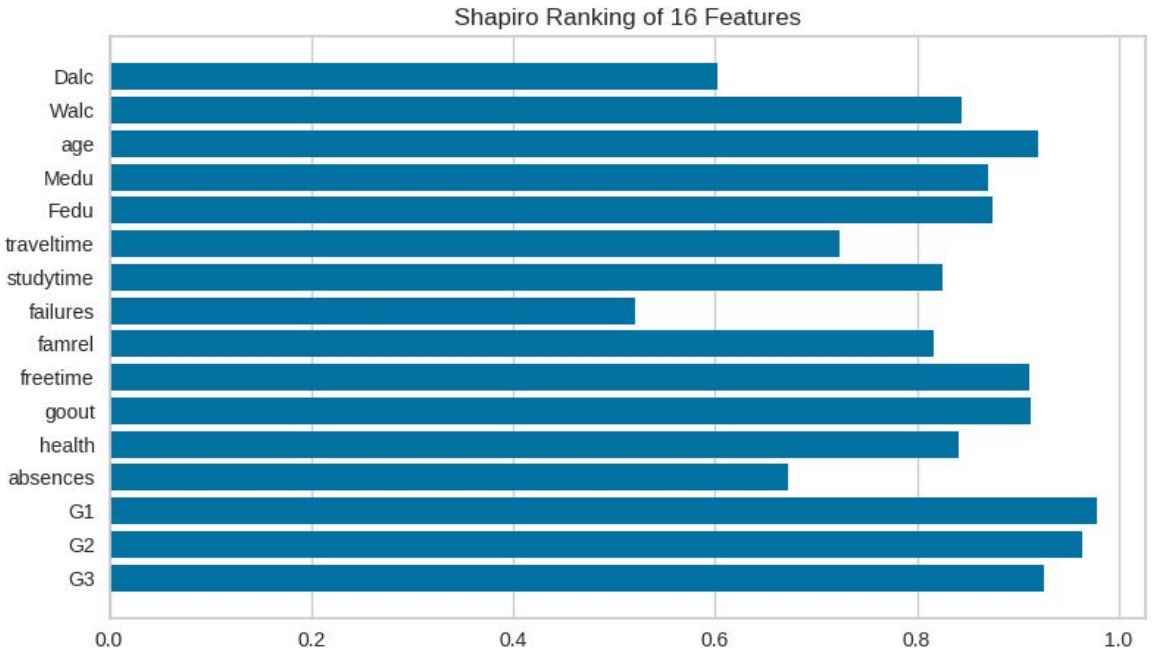


Figure 4: Rank1D evaluation for all numeric features

There are plenty of algorithms for testing the normality of a dataset, like: Chi Squared, Kolmogorov-Smirnov and Shapiro-Wilk. Normal distribution is the most widely used distribution in statistical analysis. As you can see, the features which perform the best are the grades G1, G2 and G3.

A two dimensional ranking of features

Rank2D takes pairs of features and according to what algorithm is used, it will return similarities between these. I chose the Pearson algorithm, which scores, for each pairs of features, the collinear relationships. The results obtained earlier are used to create a heatmap like the one below.

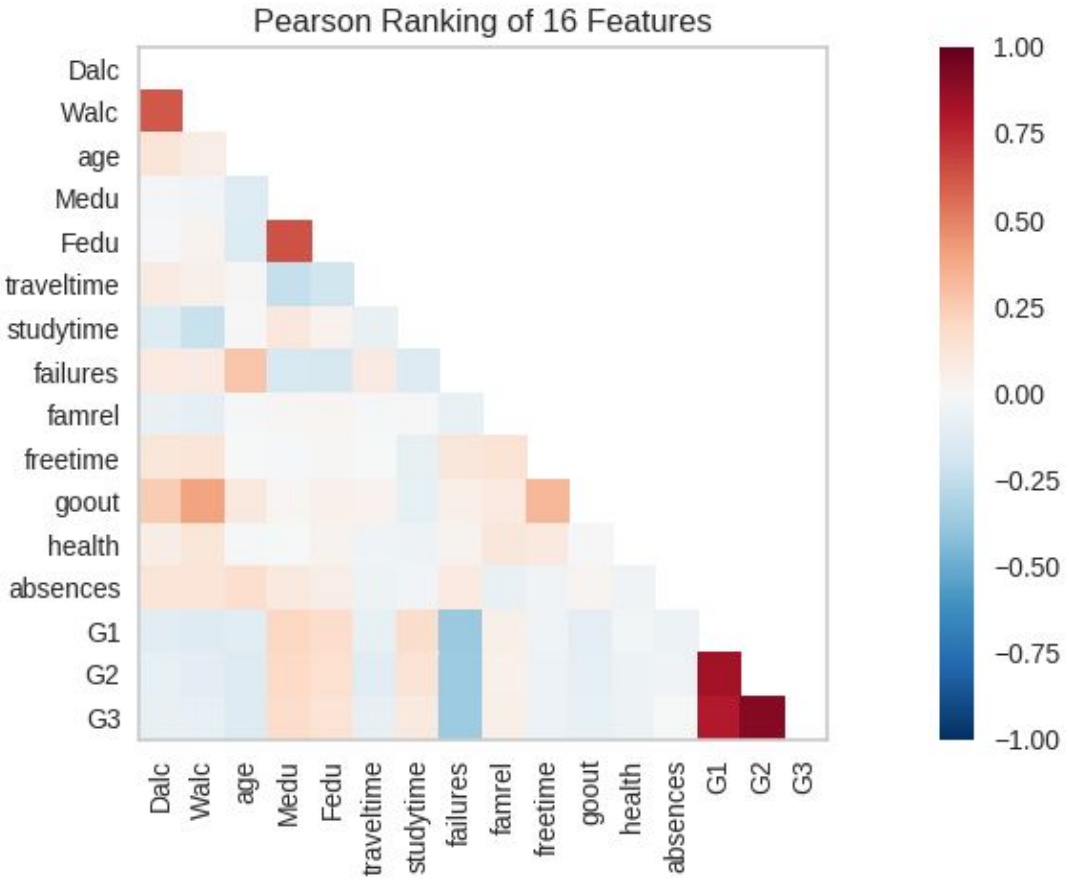


Figure 5: Rank2D evaluation for each pair of all numeric features

Obviously, the most correlated features are the grades. There exists a correlation between the alcohol consumption in a workday and a weekday. However, there are another interesting correlations: father's education with mother's education, going out with weekday alcohol consumption, going out with freetime and failures with age.

The above heatmap shows us that it's a direct link between free time, going out and weekday alcohol consumption. Most of the people who are drinking in a weekday, have freetime and hang out with their friends in a pub.

Camér's V correlation for categorical features

Regarding the fact that we can't apply Rank1D and Rank2D on categorical data, there is an alternative, named Camér's V Correlation. Although both of them can be used to obtain values for a heatmap, the first difference we can notice between Rank2D and Camér's V Correlation is the range of values. The first one returns values between [-1;+1], while Camér's V Correlation returns values between [0;1].

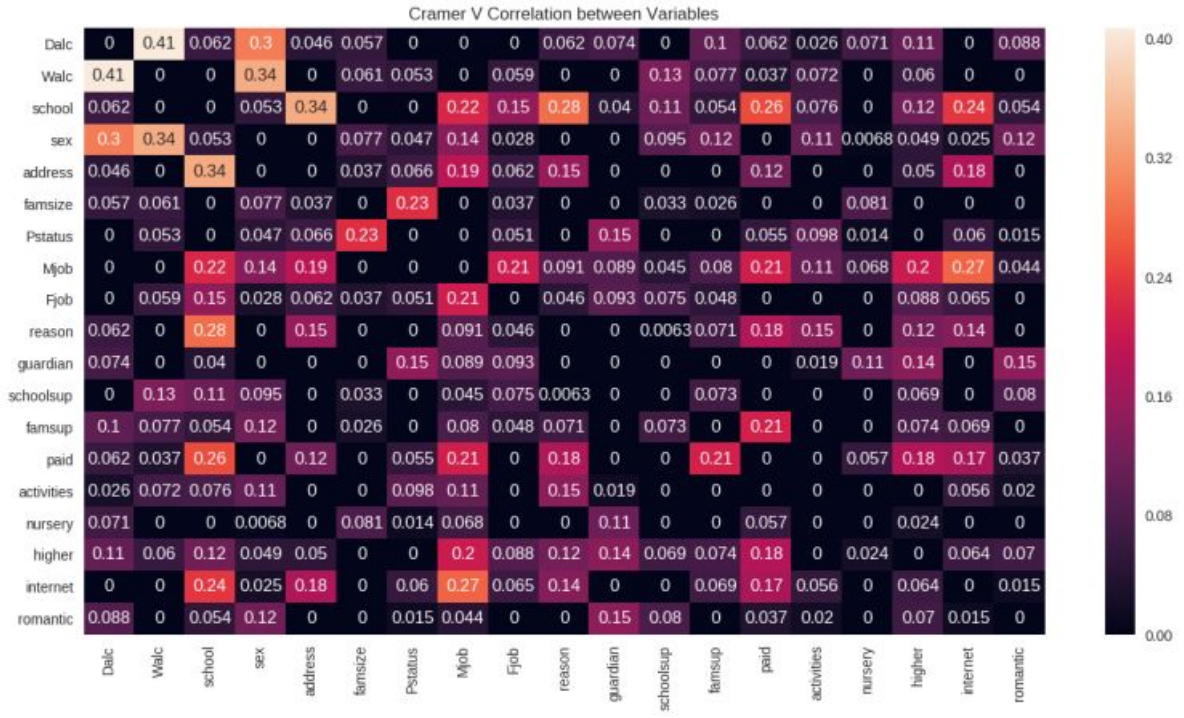


Figure 6: Camér's V Correlation Heatmap

Camér's V Correlation highlights a lot of correlations between. The most noticeable correlations are between:

- Dalc, Walc and student's sex;
- Address, reason and school;
- Family size and parent's cohabitation status;

A specific analysis on certain features

Just making a general analysis isn't enough, sometimes we need to dig deeper. There were plenty of options and combinations to plot, but I want to keep this research as general as possible, without diving too much in details.

For this section, I chose to plot the alcohol consumption in relation to with :

- The quality of family relationships, considering the family's size too;
- Some of the categorical features (romantic, internet and activities);
- Some of the numerical features (freetime, goout, health);

Analysis of features regarding family

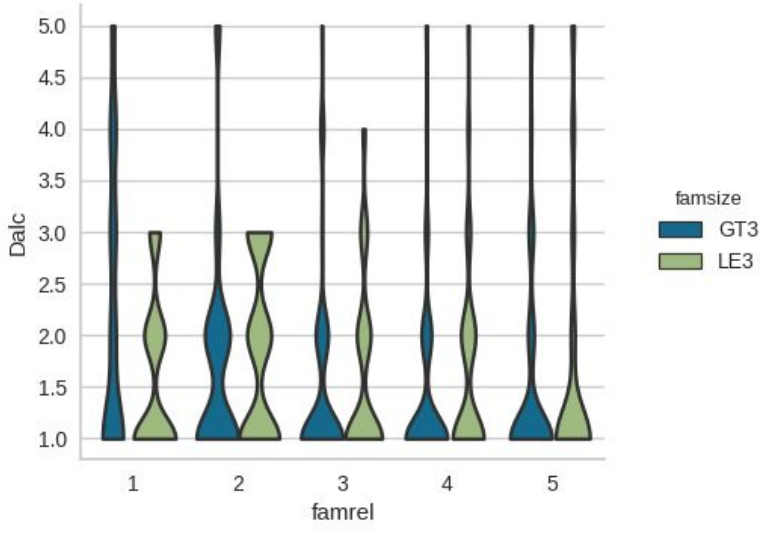


Figure 7: Workday alcohol consumption in relation to family relationship

We noticed in the earlier analysis that here are some connections between features which are linked to the family subject. I wonder what informations could offer us, the family relationship and family size, about the alcohol consumption. Indeed, the first violin plot, shows us that students with a lower quality of the family relationship tend to consume more alcohol in a workday than the ones from opposite poll.

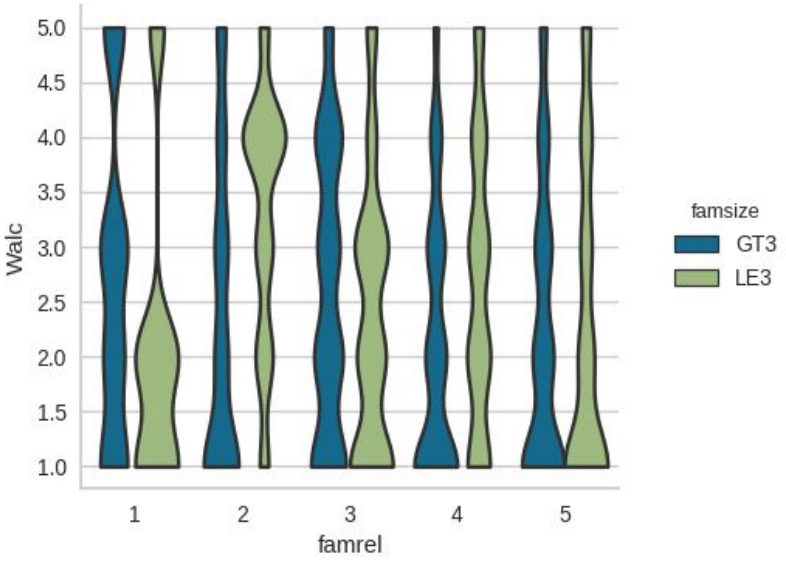


Figure 8: Weekday alcohol consumption in relation to family relationship

In the second plot, we see that the patterns for alcohol consumption in weekend, for each family size are affected, according to the quality of family relationships. Looks like a low quality of family relationships creates imbalances of the alcohol consumptions patterns in the families with different sizes.

Analysis of categorical features

Below, I've plotted some categorical data, to observe if alcohol consumption can be manipulated by "distractions", like "romantic", "internet" and "activities". There aren't too many links between these features, the only interesting thing we notice is that plots keep a kind of static ratio, between weekday and workday, with a higher level of alcohol consumption in weekday, for sure.

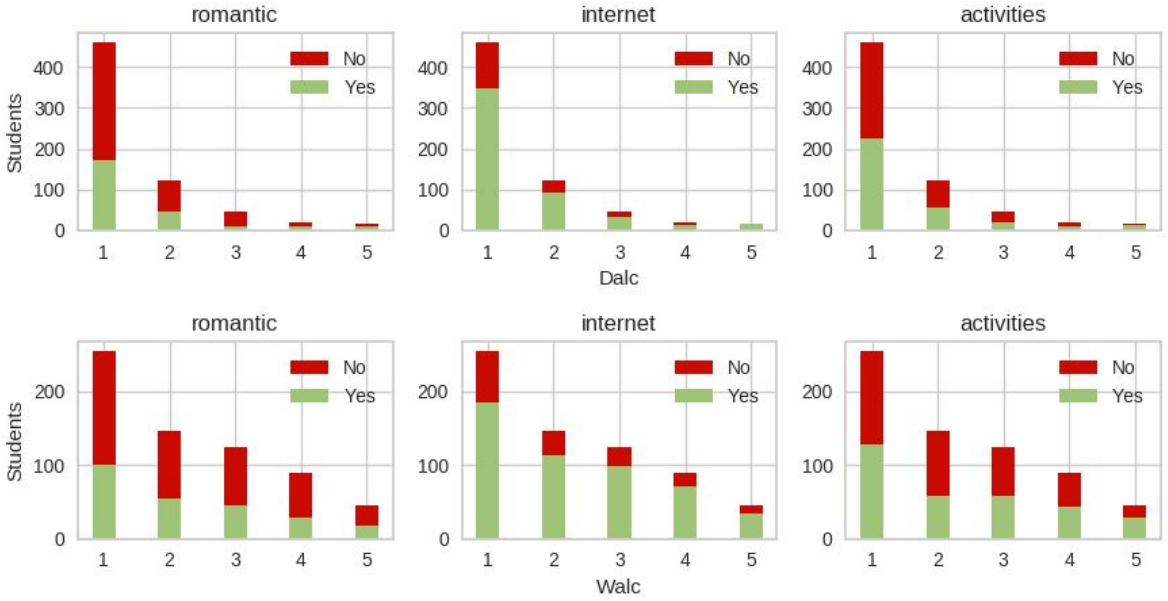


Figure 9: Stacked bar plots for 3 categorical features

Analysis of numerical features

The last figure contains density distributions of “freetime”, “goout” and “health”, regarding alcohol consumption.

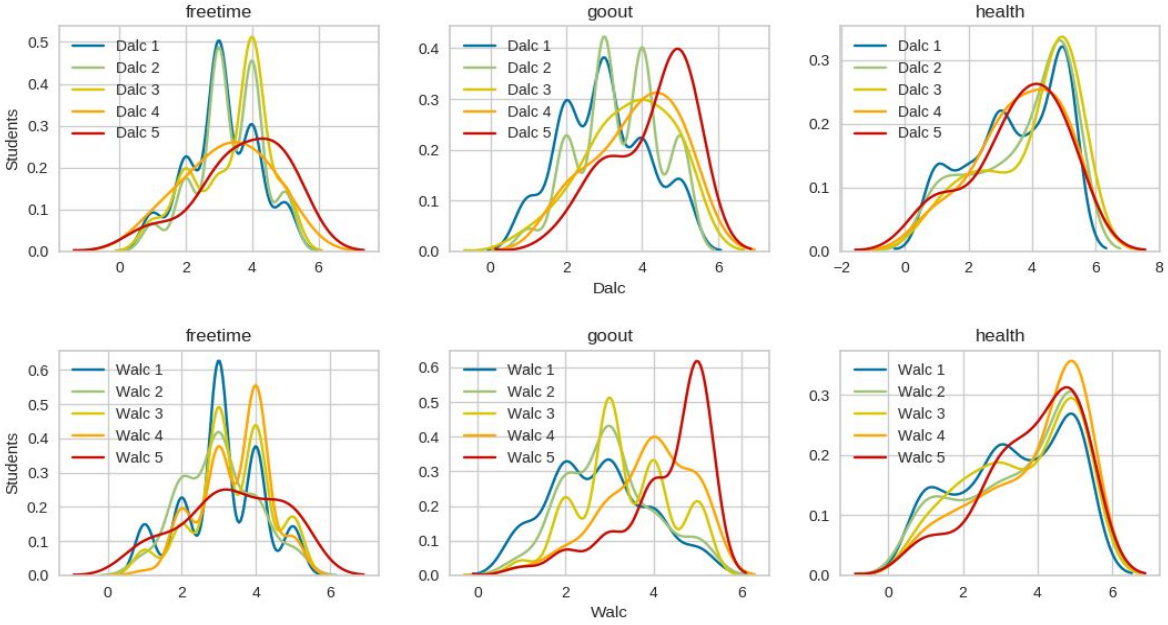


Figure 10: Density distribution plots for 3 numerical features

Alcohol consumption classification

In this part we will present how different features regarding school performance, family and social status are influencing the alcohol consumption of students. We are approaching this problem as a classification because our targets are “workday alcohol consumption” and “weekend alcohol consumption”. We will compare various classification algorithms and different approaches regarding the dataset in order to find the best parameters for this problem.

Proposed classification algorithms

Logistic Regression (Predictive Learning Model)

It is a statistical method for analysing a data set in which there are one or more independent variables that determine an outcome. The outcome is measured with a dichotomous variable (in which there are only two possible outcomes). The goal of logistic regression is to find the best fitting model to describe the relationship between the dichotomous characteristic of interest (dependent variable = response or outcome variable) and a set of independent (predictor or explanatory) variables.

Random Forest

Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks, that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. Random decision forests correct for decision trees’ habit of over fitting to their training set.

Nearest Neighbor

The k-nearest-neighbors algorithm is a classification algorithm, and it is supervised: it takes a bunch of labelled points and uses them to learn how to label other points. To label a new point, it looks at the labelled points closest to that new point (those are its nearest neighbors), and has those neighbors vote, so whichever label the most of the neighbors have is the label for the new point (the "k" is the number of neighbors it checks).

Support Vector Machine

Support vector machine is a representation of the training data as points in space separated into categories by a clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall.

Dataset preprocessing

For all trained models we used one hot encoding for categorical features and standardization by removing the mean and scaling to unit variance for numeric features. #We tried different approaches regarding the feature selection.

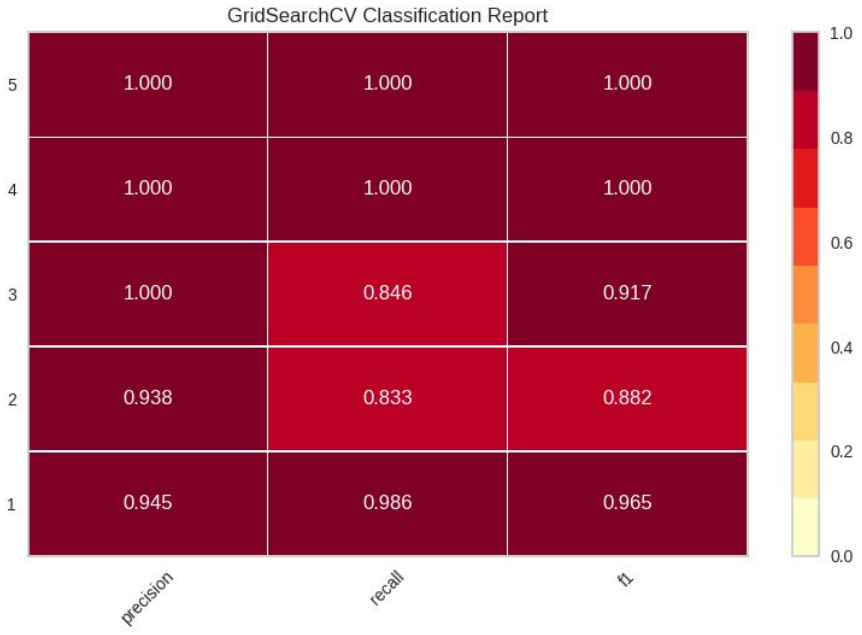
Results

For each set of features we will see how the model performs for both workdays alcohol consumption(Dalc) and weekend alcohol consumption(Walc) and compare the results. We will a table with scores for each model and custom parameters so it can be easy to reproduce the result.

Using all features

For the first model we started by using all the available features. We can see that Random Forest algorithm has the best performances on both targets, while the other algorithms have lower scores when the target is Walc.

Model	Score	Best parameters	Target
random_forest	94.95%	n_estimators': 50	Dalc
svc	76.26%	C': 0.9, 'kernel': 'linear'	Dalc
knn	71.72%	leaf_size': 5, 'n_neighbors': 12, 'p': 3	Dalc
logistic_regression	65.66%	C': 0.1, 'solver': 'lbfgs', 'tol': 0.0001	Dalc
random_forest	87.88%	n_estimators': 400	Walc
logistic_regression	54.04%	C': 0.1,'solver': 'lbfgs', 'tol': 0.0001	Walc
svc	51.01%	C': 0.6, 'kernel': 'linear'	Walc
knn	46.97%	leaf_size': 5, 'n_neighbors': 6, 'p': 3	Walc

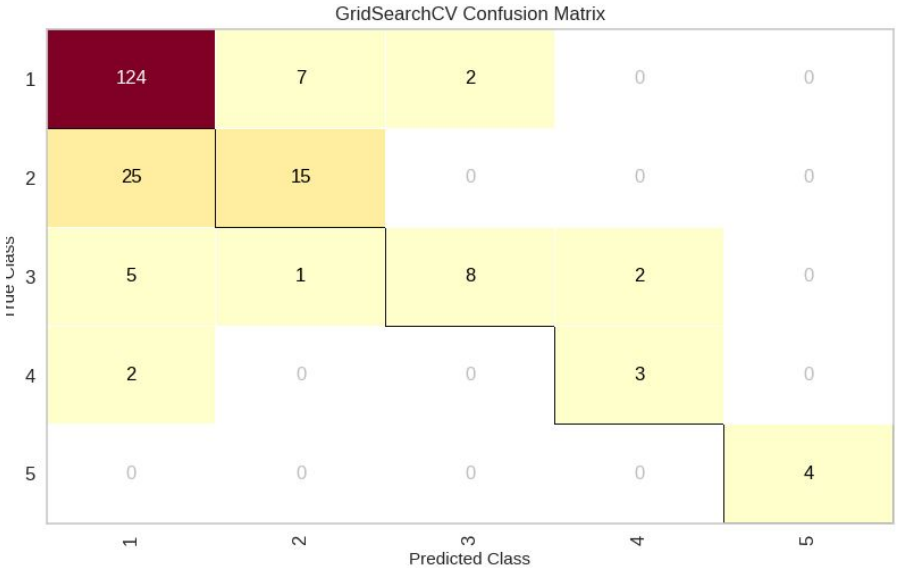


Classification report of the random forest model with 94.95%

Removing school and reason from dataset

After the first set of results we wanted to see how it would affect the models if we remove some of the features. In this set of features we removed the school because it had only two options and it didn't seem to be very relevant. Also we removed the reason column because it meant the reason for choosing that school. The results where not that different from the first set. There is a small decrease in score for the random forest model, but for the other models there is a small increase.

Model	Score	Best parameters	Target
random_forest	92.93%	n_estimators': 50	Dalc
svc	77.78%	C': 0.9, 'kernel': 'linear'	Dalc
logistic_regression	73.23%	C': 0.1, 'solver': 'liblinear', 'tol': 0.0001	Dalc
knn	68.18%	leaf_size': 5, 'n_neighbors': 4, 'p': 2	Dalc
random_forest	88.89%	n_estimators': 650	Walc
svc	54.55%	C': 0.6, 'kernel': 'linear'	Walc
logistic_regression	50.00%	C': 0.1, 'solver': 'lbfgs', 'tol': 0.0001	Walc
knn	42.42%	leaf_size': 5, 'n_neighbors': 16, 'p': 2	Walc



Confusion matrix of the SVC model with 77.78%

Removing school, reason and G1, G2

The next step in trying to remove some of the features was to remove G1 and G2 because G3 is in most of the cases the average between those two values. After we removed those features we observed a decrease for all models.

Model	Score	Best parameters	Target
random_forest	90.91%	n_estimators': 100	Dalc
logistic_regression	75.76%	C': 0.1, 'solver': 'lbfgs', 'tol': 0.0001	Dalc
svc	73.74%	C': 0.6, 'kernel': 'linear'	Dalc
knn	68.18%	leaf_size': 5. 'n_neighbors': 16, 'p': 3	Dalc
random_forest	85.86%	n_estimators': 200	Walc
svc	53.54%	C': 0.9, 'kernel': 'linear'	Walc
logistic_regression	51.52%	C': 0.1, 'solver': 'liblinear', 'tol': 0.01	Walc
knn	40.40%	leaf_size': 5, 'n_neighbors': 6, 'p': 2	Walc

Keeping only family and social features

After seeing the results obtained by removing G1 and G2, we wanted to try and see how much are the family and social status influencing the alcohol consumption. We removed all the columns that had any connection with the school activity. We removed school, reason, studytime, traveltime, failures, schoolsup, paid, activities and absences. The results show that family plays an important role this problem.

Model	Score	Best parameters	Target
random_forest	93.94%	n_estimators': 800	Dalc
svc	71.21%	C': 0.6, 'kernel': 'linear'	Dalc
logistic_regression	69.70%	C': 0.1, 'solver': 'liblinear', 'tol': 0.0001	Dalc
knn	69.19%	leaf_size': 5, 'n_neighbors': 6, 'p': 3	Dalc
random_forest	83.84%	n_estimators': 600	Walc
svc	50.00%	C': 0.9, 'kernel': 'linear'	Walc
logistic_regression	45.96%	C': 0.1, 'solver': 'lbfgs', 'tol': 0.0001	Walc
knn	39.39%	leaf_size': 10, 'n_neighbors': 6, 'p': 3	Walc

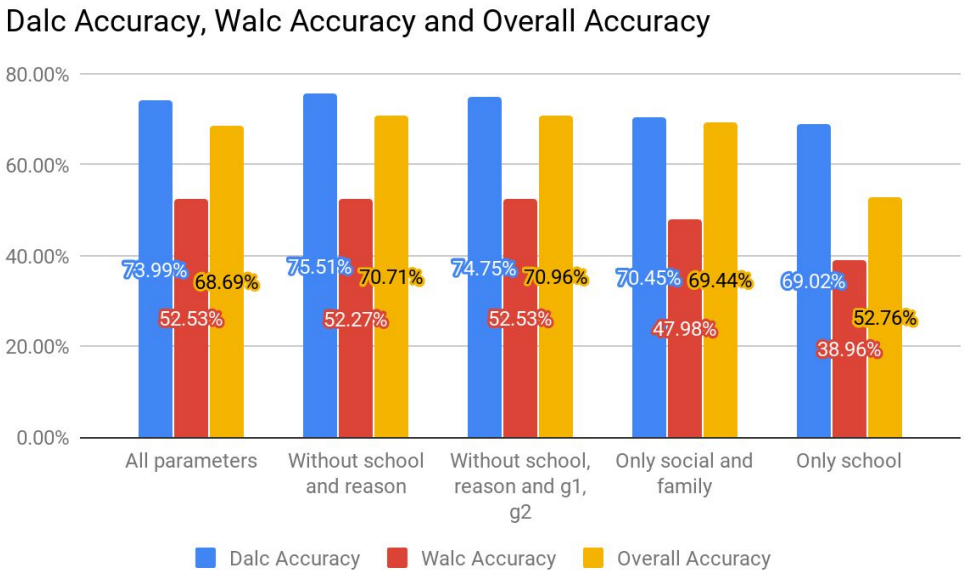
Keeping only features about school performances

After seeing how much family impacts the results on alcohol consumption, we tried the opposite and see how much impact have the school results. These results have the lowest accuracy, but the difference between models is not that big.

Model	Score	Best parameters	Target
knn	74.85%	leaf_size': 5, 'n_neighbors': 18, 'p': 2	Dalc
logistic_regression	71.17%	C': 0.1, 'solver': 'liblinear', 'tol': 0.0001	Dalc
svc	66.87%	C': 0.1, 'kernel': 'poly'	Dalc
random_forest	65.64%	n_estimators': 500	Dalc
knn	39.88%	leaf_size': 10, 'n_neighbors': 18, 'p': 2	Walc
svc	39.26%	C': 0.4, 'kernel': 'linear'	Walc
logistic_regression	38.65%	C': 0.1, 'solver': 'lbfgs', 'tol': 0.0001	Walc
random_forest	34.97%	n_estimators': 950	Walc

Overall results

The average accuracy of models based on feature sets is similar even if there was a some difference between the best models of each feature set. The Random Forest model had the best result on most of the feature sets, but the other models didn't have a big difference between their scores.



Grades regression

The direct professional bad impact which the alcohol consumption has on students are their grades. One helpful thing would be, given some background variables like parents status, age, parents education, alongside with the alcohol consumption, to be able to generate expected grades.

I chose to use Regression for this particular task to use the order relationship that is present in the grade, because it is indeed a numeric score (between 0 and 20). A classification would not work here especially because of how the data is balanced. As expected, the students' grades lay in a normal (Gaussian bell) distribution, with the majority of scores between 8 and 12 and very few 5s or 18s. The only difference here is the 0 which is pretty common, but even this is not a surprise, as 0 mostly means absence or total lack of interest. Which is not very uncommon.

Using the given dataset, I trained several Regression models to predict the final grade of a student. The used algorithms are the following:

ElasticNet

In machine learning and applied statistics, the best known model for regression is the Linear Regression model. It is a powerful tool, but has its limits and downsides.

There are a few variations of this algorithm that try to make the model better. Those would be the Lasso regressor, or the Ridge regressor. Those two use L_1 , respectively L_2 penalties. There is also a model called elastic net, which combines both penalties and gets very good results.

K Nearest Neighbours Regression

A simple implementation of KNN regression is to calculate the average of the numerical target of the K nearest neighbors. Another approach uses an inverse distance weighted average of the K nearest neighbors. KNN regression uses the same distance functions as KNN classification.

Random Tree Forest Regression

Random tree forests regression is achieved similarly to the classification, but this time there is no confidence score, just the raw number as the predicted regression value.

Regression on different features

First, I looked into the dataset, and thought that G1 and G2 are used to calculate G3 via averaging, but this seemed to not be true, as I saw grades slightly different from the average, or in some cases very different. Things like G1=10, G2=0 and G3=10 which is far from 5. So the first decision was to give it a try and try to predict G3 only from G1 and G2, without any other additional features. The results were next.

Model	best_params	MAE	R2
Random Forest Reg	**{'max_depth': 5, 'n_estimators': 400}**	**0.91**	**0.83**
SVR	**{'C': 100.0, 'gamma': 0.1}**	**0.91**	**0.83**
ElasticNet	**{'alpha': 0, 'fit_intercept': True, 'l1_ratio': 1.0}**	**0.86**	**0.82**
KNN Regression	**{'n_neighbors': 12}**	**0.99**	**0.82**

It’s impressive, how using only two features, the prediction is almost perfect, not even 1 point difference mean average error.

After these results, I decided to test every feature, except the grades before, and see how much the data reveals about the grades of the student, the results were indeed much less convenient, but it brought up pretty decent scores

Model	best_params	MAE	R2
Random Forest Reg	**{'max_depth': 15, 'n_estimators': 400}**	**2.26**	**0.45**
SVR	**{'C': 10.0, 'gamma': 0.1}**	**2.38**	**0.29**
ElasticNet	**{'alpha': 0.05263157894736842, 'fit_intercept': True, 'l1_ratio': 1.0}**	**2.54**	**0.24**
KNN Regression	**{'n_neighbors': 22}**	**2.71**	**0.12**

After trying with all features. I decided to test what impact has the alcohol consumption on the overall grade. To see this, I tried first to train the models directly on the two columns, the Dalc and Walc, the results were not very satisfying. The behaviour of the models was a simple random guess. I decided instead to try to train the models with the entire feature space (minus G1,G2 of course), but this time without the alcohol consumption. And there were some results.

Model	best_params	MAE	R2
Random Forest Reg	**{'max_depth': 25, 'n_estimators': 700}**	**2.30**	**0.38**
SVR	**{'C': 10.0, 'gamma': 0.1}**	**2.46**	**0.29**
ElasticNet	**{'alpha': 0.05263157894736842, 'fit_intercept': True, 'l1_ratio': 1.0}**	**2.61**	**0.22**
KNN Regression	**{'n_neighbors': 32}**	**2.85**	**0.09**

The results are somehow close to the others, but we can see a decrease in accuracy. The preak was obtained this time again with the Random Forest Regressor. The overall best R2 score is lower with 0.08 and with an mean absolute error higher with 0.04. Even if the SVR's R2 score has not decreased, it's MAE increased.

It seems that the best model for this task was the Random Forest Regressor. This could mean that the dataset sits in a more “categorical” space, meaning that the data is not very smooth, but has rather interesting patterns, which could be more easily revealed by a decision based model like the random forest regressor.

Conclusion

In this paper we covered the causes and the effects of alcohol consumption in students performances by comparing different features by finding what influences the alcohol consumption and school performances.

In the first part of the paper we saw that excessive alcohol consumption was present only in some cases and that students were aware of the effects of the alcohol in their lives. Also we saw the relationships between the features and how they are connected to one another.

In the classification part we found that it is a strong relationship between family and alcohol consumption and it is less affected by school performances. When using only the features that represent social status and the parents status the results were similar with the other models that also included school performances. Overall we were able to have a good prediction accuracy, the model that stands out is Random Forest having the accuracy for the most feature sets.

For the regression problem we tried to see how the alcohol consumption impacted the school performances by trying to predict the scores of the student. From the results we can see that the quantity of alcohol consumed by students helped in predicting the final grade. We were able to approximate the grade by ±2 points, by having a R2 score between 0.38 and 0.45. The random forest algorithms has the best performing results, it being helped by having most of the features categorical.

The results obtained helped see the causes and the effects of the alcohol in the lives of students, but also what stands out is that the family relationship is the most important in the young students lives.