Effects of Alcohol on Study

Min Cho **Yara Derungs** **Maria Doikova** **Joëlle Mattarelli**

github link: https://github.com/esgahtsustnoed/G3G-Project-14.git

1 Abstract

The consumption of alcohol among adolescents is an ongoing subject of discussion since it is the leading cause of mortality and is associated with health and cognitive issues. Contradictory results have been yielded on existing studies about the correlation between alcohol consumption and academic performance, requiring further research on this subject. By analyzing data from two Portuguese schools, this report aims to delve into this relationship. In addition to school grades in mathematics and Portuguese, the datasets included supplementary factors such as study habits, extracurricular activities, and socio-demographic variables gathered from surveys and school reports. Several pre-processing steps were carried out in order to train and evaluate the data employed by four supervised machine learning techniques (Decision Tree, Random Forest, Gradient Boosting Regression, Support Vector Regression) and two feature selection methods (Univariate Feature Selection, L1 Regularization) to then chose an optimal model for the final test. The performance results of the chosen model, assessed by evaluation metrics concluded that alcohol does not appear to have a significant impact on school performance of the students surveyed for this project .

2 Introduction

Alcohol consumption in adolescence is a very hotly debated topic. In Switzerland, deaths attributed to alcohol consumption are the leading cause of mortality in people aged between 15-34 years (Marmet et al. [2014]). Furthermore, alcohol consumption at a young age is associated with an accelerated decrease in gray matter at the expense of an attenuated increase in white matter. This is probably due to the fact that adolescence is a particularly vulnerable period for neurodevelopment. Binge drinking and heavy alcohol use, appear to be associated with poorer cognitive functioning in areas such as learning, memory and attention (Lees et al. [2020]). Several studies have been done on the topic concerning the relationship between alcohol consumption and school performance, but the correlation between academic achievement and alcohol consumption among students remains inconsistent (El Ansari et al. [2020]). In the study conducted by (Paschall and Freisthler [2003]) alcohol does not appear to have a significant impact on students' academic performance, while the study by (Sung et al. [2016]) shows a below-average performance in both boys and girls. Another study instead observes a marginal but statistically significant reduction in the school performance of males, while the reduction in females does not appear to be significant in the students who participate in alcohol consumption. (Balsa et al. [2011]).
The following project aims to further study the still open question of whether alcohol consumption affects high school performance by analysing the data of high school pupils of two Portuguese schools.
Following a series of data characterisation and pre-processing steps, the data was analysed through the use of four supervised machine learning models, due to the labelled nature of the data. Finally, the prediction outputs of the models were discussed in regard to the hypothesis.

3 Methods

3.1 Description of Dataset

Two distinct datasets were provided, containing information about the performance of students in the subjects of Mathematics and Portuguese. The data was collected from two Portuguese schools using questionnaires and school reports. To increase the number of data points for the training of the models, the decision was made to merge the two datasets, which is possible because the features in both datasets are exactly the same. An additional feature was added to indicate the subject distinction.

3.2 Data Transformation

The first step was to look at the data types and adjust them where necessary using the `Project14_data-dictionary.csv` file. The binary features values 'yes' were substituted with the number one, while 'no' were replaced with zero.

The dataset includes three grades obtained throughout the year: G1 (grade for the first period), G2 (grade for the second period), and G3 (final grade). Instead of taking into account individual grades, the study looked at the average grade to better understand how the consumption of alcohol affects overall academic performance. As a result, it was determined to compute the average of the three grades and save it as the target variable in a new column called `'mean_grade'`. Additionally, the average grades were rounded and converted to integers to align with the high school grading system in Portugal, which ranges between 0 and 20 and only contains whole numbers. Subsequently, the columns G1, G2 and G3 were removed, as they would otherwise be the only significant features for the models.

3.3 Handling Missing Data and Outliers

In the entire dataset no missing values were present. However, upon thorough inspection of the distribution of the grades, a remarkable number of zero grades were revealed. Nevertheless, it was established that all students had achieved at least one grade. To address this issue, the decision was made to calculate the mean for the target variable without considering the zero grades. Instead of removing whole rows containing zeros, an alternative approach was adopted to treat them as potential missing values. This being due to the fact, while poor grades can be achieved by pupils, it is considerably unlikely that such an extensive number of zero grades would be recorded throughout the year by numerous students from two separate schools, as presented in the dataset. Several other reasons might have led the students to achieve zeros. Assumptions of failed or unattended exams due to an illness are more likely, especially when only the second period grade, G2, is recorded as zero. The presence of two consecutive zeros following the first period grade, G1, gave indications about possible withdrawals prior to graduation. Based on these findings, it became more likely that missing data had already been addressed by replacing it with zero, and does not reflect the real performance of these students. Therefore, it seemed appropriate to exclude them from the analysis.

Furthermore, outliers from numerical features could be detected using the interquartile range (IQR) method. The upper whisker value, which is expressed as 1.5 times the IQR above the third quartile was calculated for each feature. Any data points located above the upper whisker were considered as potential outliers and therefore removed from the dataset.

3.4 Data Statistics

In order to gain a better insight into the data, further steps were carried out.

Descriptive statistics were generated to analyze the distribution of numerical features, revealing information such as the mean, standard deviation, minimum and maximum values, as well as quartiles. Additionally, histograms were plotted to visually represent the distribution of each numerical variable. In addition to analyzing histograms, the Shapiro-Wilk test was conducted to evaluate the normal distribution.

The linear relationship between the numerical features and the target variable was measured using the Pearson Correlation Coefficient. Applying the F-test, the calculated Pearson correlation coefficient and its p-value with the target variable could be evaluated.

Additionally, the imbalance ratio was computed, which was an essential step to understand the class distribution in the dataset and it is particularly important for stratification.

3.5 Data Splitting

To prepare the data for splitting, the categorical features underwent one-hot encoding as a primary step. The dataset was then split into training and test sets, following an 80-20 ratio, which assured that 80 percent of the data was used to train the models, while the remaining portion was for testing it. During the splitting process, stratification was applied to the feature `'failure'` to create subsets of the dataset that maintained a proportional representation of each class. This was applied because this feature is imbalanced and with the assumption that it is general understanding that failures highly correlate with grades. To address the varying scales, ranges, and units of the numerical features, standardizing was performed ensuring that each feature contributed equally to the regression analysis, regardless of their original conditions.

3.6 Feature Selection

3.6.1 Univariate Feature Selection

The univariate feature selection (UVFS) method evaluates the relationship between each feature and the target variable independently. It is especially suitable for high-dimensional feature spaces (Pudjihartono et al. [2022]) which was the case after one-hot-encoding.
Using the f-regression scoring function the features could be ranked based on their importance. As the hyperparameter 'k' represents the number of optimal features to be selected from the dataset, it was determined by the SelectKBest class provided by scikit-learn. The best value of 'k' could be selected through a cross validation with five folds with the most relevant features that maximizes the performance metric.

3.6.2 L1 Regualarization in a Lasso Regression Classifier

L1 feature selection was performed using L1 regularization with a lasso regression classifier. Important features are selected by shrinking the coefficients of less relevant ones to zero (D'Angelo et al. [2009]). To get the best value for the regularization parameter alpha, a grid search with a five-fold cross-validation was implemented. The best trade-off between model complexity and predicted accuracy was selected by comparing the model's performance for various alpha values. By testing a variety of ranges of alpha values, the one that performed best could be found.

3.7 Machine Learning Models

The requirements for our prediction models were as follows. Since our output target `'mean_grade'` was given and it was numerical, a supervised machine learning algorithm was looked for. Thanks to the pre-processing steps, the data were found to be neither normally distributed nor linearly related. Based on the understanding of these characteristics and all other previously mentioned requirements, it was possible to select the appropriate model types to use.
The Decision Tree Regressor, Random Forest Regressor, Gradient Boosting Regressor and Support Vector Regressor (SVR) were chosen for this project. Additionally, for each model a random state of 42 was set, in order to allow for reproducibility of the final test results. A Decision Tree is an easy-to-interpret predictive algorithm that at each node compares attributes to a set of values in order to narrow down on an outcome (in the case of a decision tree regressor the outcome is numerical) in a top-down manner (Kotsiantis [2013]). On the other hand, Random Forest combines several decision trees, which allow it to make predictions with high accuracy (Liu et al. [2012]). Gradient Boosting Regression incorporates several weaker prediction models, often tree models are used, and "boosts" prediction performance to create an optimal model (Cai et al. [2020]). Support Vector Regression is based on support vector machines, but is modified to accommodate regression problems (Lin et al. [2022]).
These models were trained and validated on the 80% split of the data using a crossvalidation method (KFold) with the aim of selecting the best performing model. In this case a five-fold crossvalidation was utilized. A crossvalidation loops over the provided data, here five times, for each model. In each loop different sections of the data are chosen for training and validation, allowing the models to avoid overfitting, due to the several iterations. (Koul et al. [2018])

In order for each model to perform as best as possible, various hyperparameters were assessed and tuned to each respective model with the aim to optimize the models' performance. For each model, two to three hyperparameters specific to that model were used, while assuming a balance between the size of the dataset and the computational abilities available. The specific values applied for each hyperparameter were the default list of settings provided by Scikit-learn documentation for each model. However, the ranges were not set by default but were instead chosen to accommodate this project. Originally, larger ranges and values were used. However, the subsequent training and testing of the models was too time consuming. Therefore, the ranges for hyperparameters with numerical values were adjusted such that in an efficient manner, effective and meaningful optimal values could be obtained. The hyperparameters were tuned using a GridSearchCV function, which in essence systematically iterates through a pre-defined list of values for each of the chosen hyperparameters (=grid). It does this in search of a combination of the best 'settings' (the best value for each hyperparameter from the grid) that lead to optimal performance for the respective model to which this function is applied. The GridSearchCV function has an integrated crossvalidation. For this project, five folds were used to loop through the data and create models that would yield the best results when applied for further training and testing (Ahmad et al. [2022]) (Ranjan et al. [2019]). The chosen hyperparameters for the corresponding models with the corresponding list of values are shown in Table 1.

Decision Tree	**Random Forest**	**Gradient Boosting Regression**	**SVR**
criterion *(absolute error, squared error, poisson, friedman mse)*	**criterion** *(absolute error, squared error, poisson, friedman mse)*	**learning rate** *(0.05, 0.1, 0.15, 0.2, 0.25)*	**kernel** *(linear, poly, rbf, sigmoid)*
splitter *(best, random)*	**n estimators** *(range 10-19)*	**n estimators** *(range 50-99, in steps of 10)*	**epsilon** *(range 0.1-0.4, in steps of 0.1)*
max depth *(range 0-9)*	**max depth** *(range 1-4)*	**max depth** *(range 1-4)*	

Table 1: Hyperparameters tuned for each model

3.8 Evaluation Metrics

Due to the nature of the project being a regression problem, the appropriate evaluation metrics to be used were R-squared (R2), root mean squared error (RMSE), mean squared error (MSE) and the mean absolute error (MAE). The definitions of each can be found in Table 2.

R^2	Determines the proportion of variance in the dependent variable that can be explained by the independent variable
RMSE	Average deviation between the predicted and the actual points
MSE	Square root of the average squared difference between the predicted values and the actual values in a dataset. Average deviation between the predicted and the actual points
MAE	Average absolute difference between the predicted values by the models and the actual values in a dataset

Table 2: Evaluation Metrics

4 Results

4.1 Data Overview

The two provided datasets included 33 variables, such as student grades, demographic data, social aspects, and school-related features. The Mathematics dataset consisted of 395 datapoints, while the Portuguese dataset had 649 datapoints. The combined dataset comprised 1044 datapoints and 34 attributes without any duplicates, as summarized in Table 3. Since the complete table does not fit on one page, or the content would otherwise be unreadable, only the first and last five columns are shown here.

school	sex	age	address	famsize	...	absences	G1	G2	G3	subject
GP	F	18	U	GT3	...	6	5	6	6	maths
GP	F	17	U	GT3	...	4	5	5	6	maths
GP	F	15	U	LE3	...	10	7	8	10	maths
GP	F	15	U	GT3	...	2	15	14	15	maths
GP	F	16	U	GT3	...	4	6	10	10	maths

Table 3: Data overview (first five rows)

After adjusting the data types, the dataset consisted of 21 categorical features, five integer features, and eight binary features. Following additional pre-processing steps to get the target variable `'mean_grade'`, the dataset has been transformed to contain a total of 32 features, with the same amount of categorical and binary features but only three features remained as numerical variables. Finally, after the one-hot encoding process, the dataset expanded to include a total of 91 features.

None of the numerical features appeared to follow a normal distribution, as can be seen from the histograms displayed below in Figure 1. Statistical evidence from the Shapiro-Wilk test further supported this observation. Based on the analysis conducted using a significance level of 0.01667 (0.05 divided by the total number of numerical features), the calculated p-values were found to be significantly lower than the chosen level. Therefore, it can be concluded that none of the features exhibited a normal distribution.

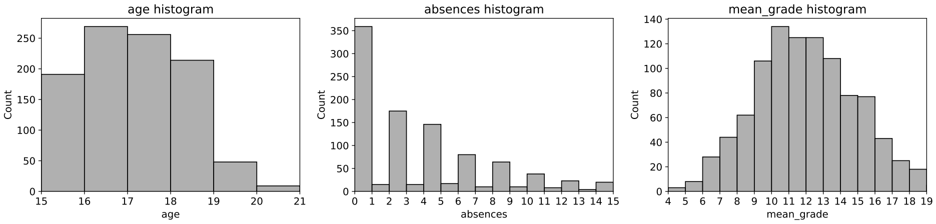

Figure 1: Distribution histograms

The F-test, the calculated Pearson correlation coefficient and its p-value led to the conclusion that the variables do not show linearity.

The outliers in this case refer to individuals who are either older than 21 or have more than 15 absences. In Figure 2, the top row shows the boxplots with the outliers and the bottom row shows the adjusted data, without outliers.

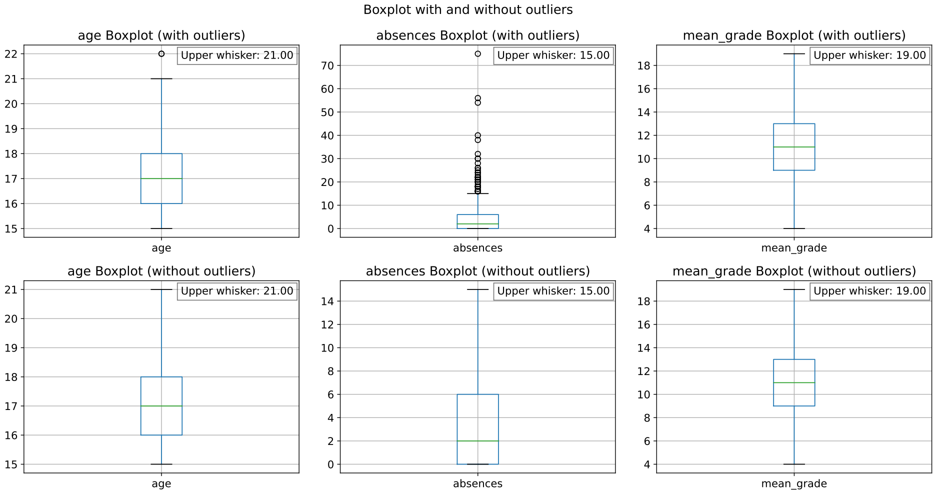

Figure 2: Boxplot outliers vs no outliers

For each of the four models, and the respective feature selection method used, the best hyperparameters were chosen. These are presented in Table 4. The models were optimized for training and testing with exactly these tuned hyperparameters.

Decision Tree (L1)	**Decision Tree (UVFS)**	**Random Forest (L1)**	**Random Forest (UVFS)**	**Gradient Boosting Regression (L1)**	**Gradient Boosting Regression (UVFS)**	**SVR (L1)**	**SVR (UVFS)**
best criterion *poisson*	**best criterion** *poisson*	**best criterion** *absolute error*	**best criterion** *absolute error*	**best learning rate** *0.15*	**best learning rate** *0.15*	**kernel** *poly*	**kernel** *poly*
best splitter *best*	**best splitter** *random*	**n estimators** *19*	**n estimators** *12*	**n estimators** *60*	**n estimators** *90*	**best epsilon** *0.4*	**best epsilon** *0.4*
best max depth *3*	**best max depth** *3*	**best max depth** *4*	**best max depth** *4*	**best max depth** *3*	**best max depth** *3*		

Table 4: Best hyperparameters tuned for each model

4.2 Feature Selection

There is some overlap between the features selected by UVFS and L1 regularization techniques. Both methods identified the following features as important: `’schoolsup’` (extra educational support), `’higher’` (wants to take higher education), `’internet’` (internet access at home), `’absences’` (number of school absences), `’school_GP` (school Gabriel Pereira), `’Medu_1’` (mother with primary education), `’Medu_4’` (mother with higher education), `’failures_0’` (no past class failure), `’freetime_2’` (low free time after school), `’Dalc_1’` (low workday alcohol consumption), `’Walc_4’` (high weekend alcohol consumption), and `’subject_maths` (the subject Mathematics). These features were deemed influential in both approaches, indicating their strong predictive power for the outcome variable. However, each technique also identified unique features. Both identified `’failures_0’` (no past class failure) as the most significant one.

4.2.1 Univariate Feature Selection

The analysis determined that the optimal number of features for prediction and simultaneously best ’k’ was 28. Their descending order of importance are shown in Figure 3. Additional selected variables are `’address’` (student’s home address type, urban or rural), `’Fedu_1’` (father with primary education), and `’Fedu_4’` (father with higher education), specific parental occupations (`’Mjob_at_home’`, `’Mjob_health’`, `’Fjob_teacher’`), choosing the school based on its reputation `(’reason_reputation’)`, and the subject Portuguese (`’subject_portuguese’`).
However, alcohol appears to have a relatively minor influence, as it is only ranked ninth with very low workday alcohol consumption (`’Dalc_1’`) and twenty-seventh with high weekend alcohol consumption (`’Walc_4’`).

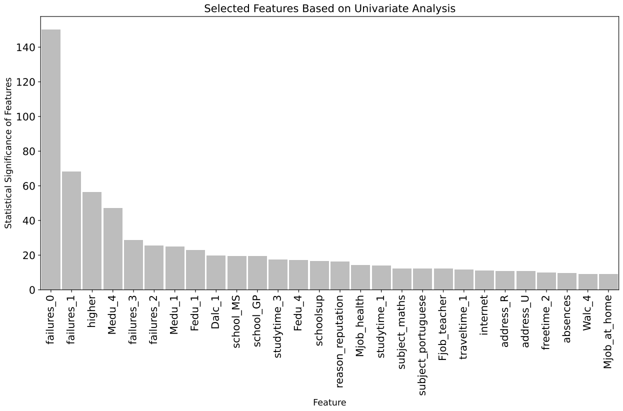

Figure 3: Selected Features UVFS

4.2.2 L1 Regualarization in a Lasso Regression Classifier

The results shown in Figure 4 display the most to least important features identified by the L1 regularization method with a best alpha found by grid search of 0.02, which gives 50 important features. Here additional features such as demographic factors (`'age'`), family educational support (`'famsup'`), extra paid classes within the course subject (`'paid'`), extra-curricular activities (`'activities'`), romantic relationship (`'romantic'`), family size less or equal to three (`'famsize_GT3'`), parent's living apart (`'Pstatus_A'`), student's guardian (`'guardian_father'`, `'guardian_mother'`), 15 to 30 minutes home to school travel time (`'traveltime_2'`), weekly study time (`'studytime_2'` (2 to 5 hours), `'studytime_4'` (>10 hours)), good quality of family relationships (`'famrel_4'`), going out with friends (very low to high, `'goout_1'`, `'goout_2'`, `'goout_3'`), current health status (`'health_1'` (very bad), `'health_3'` (moderate), `'health_5'` (very good)), and very low weekend alcohol consumption (`'Walc_1'`). Even though alcohol appears here three times, the first time is only at position 20 with `'Walc_4'` then at position 24 with `'Dalc_1'` and at position 32 with `'Walc_1'`.

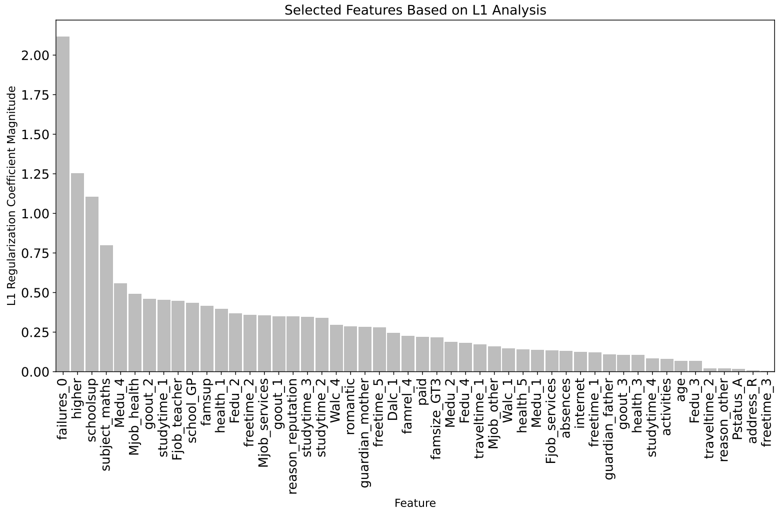

Figure 4: Selected Features L1

4.3 Machine Learning Models

The following table (Table 5) shows the results obtained from the cross-validation carried out with five folds on 80% of the dataset, namely the training set. The mean values with the standard deviations for each evaluation metric were calculated across the five folds. The performance is compared between the four models, among which the two feature selection methods for each model (L1 regularization and UVFS) are also presented for comparison.

Model	Feature selection	Train/Validation	R2 mean±std	MSE mean±std	RMSE mean±std	MAE mean±std
DT	L1	Train	0.252 ± 0.015	6.233 ± 0.185	2.497 ± 0.037	1.995 ± 0.026
DT	L1	Validation	0.186 ± 0.054	6.688 ± 0.875	2.581 ± 0.167	2.073 ± 0.116
DT	UVFS	Train	0.252 ± 0.014	6.238 ± 0.186	2.498 ± 0.037	1.992 ± 0.027
DT	UVFS	Validation	0.16 ± 0.052	6.916 ± 0.974	2.624 ± 0.183	2.11 ± 0.133
RF	L1	Train	0.339 ± 0.008	5.51 ± 0.233	2.347 ± 0.05	1.838 ± 0.04
RF	L1	Validation	0.221 ± 0.033	6.406 ± 0.811	2.527 ± 0.159	2.023 ± 0.127
RF	UVFS	Train	0.339 ± 0.014	5.515 ± 0.233	2.348 ± 0.05	1.836 ± 0.034
RF	UVFS	Validation	0.211 ± 0.024	6.499 ± 0.889	2.544 ± 0.175	2.026 ± 0.145
gb	L1	Train	0.631 ± 0.005	3.074 ± 0.079	1.753 ± 0.023	1.393 ± 0.021
gb	L1	Validation	0.279 ± 0.054	5.913 ± 0.649	2.428 ± 0.135	1.937 ± 0.122
gb	UVFS	Train	0.714 ± 0.008	2.385 ± 0.039	1.544 ± 0.012	1.21 ± 0.017
gb	UVFS	Validation	0.294 ± 0.045	5.799 ± 0.657	2.404 ± 0.137	1.914 ± 0.132
svr	L1	Train	0.694 ± 0.012	2.55 ± 0.088	1.597 ± 0.028	1.109 ± 0.017
svr	L1	Validation	0.274 ± 0.056	5.948 ± 0.572	2.436 ± 0.119	1.928 ± 0.108
svr	UVFS	Train	0.709 ± 0.011	2.421 ± 0.061	1.556 ± 0.02	1.076 ± 0.013
svr	UVFS	Validation	0.294 ± 0.04	5.807 ± 0.738	2.405 ± 0.155	1.891 ± 0.128

Table 5: Crossvalidation mean performance of each model

The best model overall was chosen by comparing the various R2 scores (validation) retrieved from the cross-validation. However, it is important to mention that this model also scored the best in RMSE, MSE and MAE. It was the Support Vector Regression (svr) that had outperformed the rest, utilizing the univariate feature selection (UVFS) method to select the best features. The optimal hyperparameters tuned to this best model were a 'poly' kernel and an epsilon value of 0.42. The 'poly' kernel maps the data from a low- to a high-dimensional space using a polynomial function (Patle and Chouhan [2013]). Epsilon on the other hand accounts for a training error. In this case, there is a distance of 0.4 between the predicted and real values (Tan et al. [2015]).
This chosen model was then tested on the remaining 20% data, namely the test set. The evaluation metrics obtained from testing the final model can be seen in Table 6, with an R2 of 0.42, an MSE of 5.194, an RMSE of 2.279, and an MAE of 1.763.

Train/Test	R2	MSE	RMSE	MAE
Train	0.707	2.447	1.564	1.091
Test	0.42	5.194	2.279	1.763

Table 6: Evaluation metrics of best model *SVR with UVFS*

5 Discussion

The aim of this project was to predict the school performance of students based on alcohol consumption. However, as seen with both feature selection methods, the quantity of alcohol that a student consumes does not appear to be a relevant predictor to how well the student does in school based on the present dataset. In order to find support for these claims all of the experiments were carried out once also by excluding alcohol from the dataset. In this case, the models' performances deviated only very slightly from the results presented in this report. This meaning that the presence of the feature alcohol seemingly does not have much influence on the final outcome. The final output of this project resulted in support vector regression being the best model with an R2 score of 0.42. It was the R2 score that was utilized in search of an optimal model due to the fact that R2 evaluates the goodness of fit, essentially evaluating the performance of models. Even though the MSE, RMSE and MAE score could be considered relatively low, the R2 score obtained is not very high and it suggests that even the best model is only capturing 42% of the variability in the target variable – `'mean_grade'`. Taking into account all four scores, it can therefore be assumed that the overall performance of the model tends to a rather poor performance. The model's performance could further be argued as less relevant for the aim of the project, as the significant feature for the hypothesis - `'alcohol'` - was not regarded of high importance to the model.

Literature to date, as previously mentioned, shows very inconsistent results on this topic. Some studies show a correlation that higher alcohol consumption poses a risk for lower grades in school student (Hayatbakhsh et al. [2011]). On the other hand, other studies show no effect of alcohol, even of heavy - alcohol consumption, on school performance (Paschall and Freisthler [2003]). In comparison to previous literature research, which shows inconsistent information on the effect of alcohol on student performance, the results obtained in this project point towards the fact that alcohol consumption and school performance in pupils are likely, not correlated.
It is important to note that, as stated in an article from 2014 (Hemphill et al. [2014]), alcohol is not the only contributing factor influencing school grades. This can also be observed in the analysis carried out in this report. Moreover, it can be assumed that there are other co-factors with which it may act in conjunction, where alcohol may become more or less influential.

There are many limitations that posed difficulties to the final outcome of the project. First and foremost, time was a limiting factor that did not allow for much experimentation with other factors that could have improved the model performance. Had this been possible, it could be that the final model could have been optimized to produce better scores. Secondly, the dataset is considerably small. Therefore, not many data points were available for the models to train on, in order to return optimal test results. It is important to further notice that too little information was provided on the collection of the data. Regarding the question at hand, it is not clear, for example, in which time frame alcohol was consumed and whether this was affected by other environmental or social factors. Furthermore, the data was gathered from surveys which are subjective and interpretation bias is a possibility.

Furthermore, only two school subject datasets for analysis were provided. In contrast, in a paper from 2014 (Golino et al. [2014]) that has a very similar experimental aim, the similar datasets, expanded to nine different subjects were used. In addition to this, the dataset is limited to only two schools in one single country. It is likely that students perform very differently based on their environment and the resources available, and therefore the considered sample is not representative on a global scale.

In conclusion, the consumption of alcohol was shown to have no significant influence on the performance of high school pupils. Nevertheless, from all models used, the Support Vector Regression with univariate feature selection was most suited to making predictions on the data for this project. This finding could provide a basis, of an already identified optimal model and feature selection method, that could be further used to explore not only this, but similar datasets. As an outlook, alcohol consumption could be assessed in older students, e.g. undergraduates and postgraduates. This would be of interest since certain neurodegenerative cognitive impacts may take more time to develop.

Additionally, this dataset was evaluated in the context of a regression problem. However, it could be a valuable observation to convert the grade into a categorical feature (e.g. <10: fail, >=10: pass), and pose the same question as a classification problem.

Furthermore, since the presented sample is not generalisable, experiments using a dataset with a larger, more randomised cohort would be interesting to look at, as this could lead to possibly the most conclusive results on this topic to date.

References

Ghulab Nabi Ahmad, Hira Fatima, Shafi Ullah, Abdelaziz Salah Saidi, and Imdadullah. Efficient medical diagnosis of human heart diseases using machine learning techniques with and without gridsearchcv. *IEEE Access*, 2022. doi: 10.1109/ACCESS.2022.3165792.

Ana I Balsa, Laura M Giuliano, and Michael T French. The effects of alcohol use on academic achievement in high school. *Economics of Education Review*, 30(1):1–15, 2011. doi: 10.1016/j.econedurev.2010.06.015.

Jianchao Cai, Kai Xu, Yanhui Zhu, Fang Hu, and Liuhuan Li. Prediction and analysis of net ecosystem carbon exchange based on gradient boosting regression and random forest. *Applied Energy*, 262: 114566, 3 2020. ISSN 03062619. doi: 10.1016/j.apenergy.2020.114566.

Gina M D'Angelo, Dc Rao, and C Charles Gu. Combining least absolute shrinkage and selection operator (LASSO) and principal-components analysis for detection of gene-gene interactions in genome-wide association studies. *BMC proceedings*, 3 Suppl 7(Suppl 7):S62, 12 2009. ISSN 1753-6561. doi: 10.1186/1753-6561-3-s7-s62.

Walid El Ansari, Abdul Salam, and Sakari Suominen. Is alcohol consumption associated with poor perceived academic performance? survey of undergraduates in finland. *International Journal of Environmental Research and Public Health*, 17(4):1369, 2020. doi: 10.3390/ijerph17041369.

Hudson F. Golino, Cristiano Mauro Assis Gomes, and Diego Andrade. Predicting academic achievement of high-school students using machine learning. *Psychology*, 5(18):2051–2063, 2014. doi: 10.4236/psych.2014.518207. URL `https://www.scirp.org/journal/paperinformation.aspx?paperid=52289`. [PDF] [HTML] [XML] 5.134 Downloads 7.944 Views Citations.

Mohammad Reza Hayatbakhsh, Jake M. Najman, William Bor, Alexandra Clavarino, and Rosa Alati. School performance and alcohol use problems in early adulthood: a longitudinal study. *Alcohol*, 45(7):701–709, 11 2011. ISSN 07418329. doi: 10.1016/j.alcohol.2010.10.009.

Sheryl A. Hemphill, Jessica A. Heerde, Kirsty E. Scholes-Balog, Todd I. Herrenkohl, John W. Toumbourou, and Richard F. Catalano. Effects of Early Adolescent Alcohol Use on Mid-Adolescent School Performance and Connection: A Longitudinal Study of Students in Victoria, Australia and Washington State, United States. *Journal of School Health*, 84(11):706–715, 11 2014. ISSN 00224391. doi: 10.1111/josh.12201.

S. B. Kotsiantis. Decision trees: a recent overview. *Artificial Intelligence Review*, 39(4):261–283, 4 2013. ISSN 0269-2821. doi: 10.1007/s10462-011-9272-4.

Atesh Koul, Cristina Becchio, and Andrea Cavallo. Cross-Validation Approaches for Replicability in Psychology. *Frontiers in Psychology*, 9, 7 2018. ISSN 1664-1078. doi: 10.3389/fpsyg.2018.01117.

Briana Lees, Lindsay R Meredith, Anna E Kirkland, Brittany E Bryant, and Lindsay M Squeglia. Effect of alcohol use on the adolescent brain and behavior. *Pharmacology, Biochemistry, and Behavior*, 192:172906, 2020. doi: 10.1016/j.pbb.2020.172906.

Guancen Lin, Aijing Lin, and Danlei Gu. Using support vector regression and K-nearest neighbors for short-term traffic flow prediction based on maximal information coefficient. *Information Sciences*, 608:517–531, 8 2022. ISSN 00200255. doi: 10.1016/j.ins.2022.06.090.

Yanli Liu, Yourong Wang, and Jian Zhang. New machine learning algorithm: Random forest. In *Proceedings of the Third International Conference on Information Computing and Applications*, 2012. doi: 10.1007/978-3-642-34062-8_32.

Simon Marmet, Jürgen Rehm, Gerrit Gmel, Hannah Frick, and Gerhard Gmel. Alcohol-attributable mortality in switzerland in 2011–age-specific causes of death and impact of heavy versus non-heavy drinking. *Swiss Medical Weekly*, 144(w13947), 2014. doi: 10.4414/smw.2014.13947. Free article.

Mallie J Paschall and Bridget Freisthler. Does heavy drinking affect academic performance in college? findings from a prospective study of high achievers. *Journal of Studies on Alcohol*, 64(4):515–519, 2003. doi: 10.15288/jsa.2003.64.515.

A. Patle and D. S. Chouhan. SVM kernel functions for classification. In *2013 International Conference on Advances in Technology and Engineering (ICATE)*, pages 1–9. IEEE, 1 2013. ISBN 978-1-4673-5619-0. doi: 10.1109/ICAdTE.2013.6524743.

Nicholas Pudjihartono, Tayaza Fadason, Andreas W Kempa-Liehr, and Justin M O'Sullivan. A Review of Feature Selection Methods for Machine Learning-Based Disease Risk Prediction. *Frontiers in bioinformatics*, 2:927312, 2022. ISSN 2673-7647. doi: 10.3389/fbinf.2022.927312.

G S K Ranjan, Amar Kumar Verma, and Sudha Radhika. K-Nearest Neighbors and Grid Search CV Based Real Time Fault Monitoring System for Industries. In *2019 IEEE 5th International Conference for Convergence in Technology (I2CT)*, pages 1–5, 2019. doi: 10.1109/I2CT45611.2019.9033691.

Dong Jun Sung, Wi-Young So, and Taikyeong Ted Jeong. Association between alcohol consumption and academic achievement: a cross-sectional study. *Chinese Journal of Public Health*, 32(4): 473–478, 2016. doi: 10.21101/cejph.a4292.

Maojin Tan, Xiaodong Song, Xuan Yang, and Qingzhao Wu. Support-vector-regression machine technology for total organic carbon content prediction from wireline logs in organic shale: A comparative study. *Journal of Natural Gas Science and Engineering*, 26:792–802, 9 2015. ISSN 18755100. doi: 10.1016/j.jngse.2015.07.008.