Course of Machine Learning - Professor Fabio Antonio Stella - Academic year 2023/2024

Comparison of principals classification models analyzing students' alcohol consumption

Team 08: Adobati Simone, D'Amicis Andrea, Di Rocco Fabio, Mariani Chiara
Università degli Studi di Milano-Bicocca, CdLM Data Science

Abstract

What is the alcohol consumption of students in Portugal? The data analysed in this paper was collected by using school reports and questionnaire submitted in the *Alentejo* region of Portugal during the 2005-2006 school year. The goal of this research work is to predict the students' alcohol consumption based on some school and family aspects of their life using appropriately trained classification models and assessing the goodness of prediction.

Contents

1 Introduction

Alcohol consumption in the youth is a huge problem and need some active solutions to try to limit it. In this machine learning research we tried to figure out how much different choices in everyday life, with a focus on scholastic aspects, could impact on students.

This study will consider data collected during the academic school year 2005-2006 in two Portugal school located in *Alentejo* region. More in detail, the two considered sources are built from school reports and questionnaires taken in two different schools and courses.

The present work intends to predict students alcohol consumption using different classification techniques and models in machine learning, analysing different evaluation measure to understand which classification method is the best to comprehend this problem.

We are going to exhibit and explain, in the following analysis, the valuable information that could help to identify which factors could influence students to have an higher alcohol consumption.

To correctly achieve our goal, we divided the work in four phases each one with a different aim. Here are the followed step:

- **Data preparation**: understand and prepare the dataset
- **Preprocessing**: find, change and select the features
- **Classification**: use machine learning to make some prediction
- **Evaluation**: compare the models and draw conclusions

To made our analysis we worked with Knime Analytics Platform (Knime, 2023), a software that allows the user to implement machine learning strategies with an easy and intuitive layout.

1.1 Data Features

The dataset chosen for answering this research question is the Student Alcohol Consumption, available on the Kaggle Platform (Kaggle, 2016). It is made of 1044 records each with the following 34 features:

- *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** – 5th to 9th grade, **3** – secondary education or **4** – higher education)
- *Fedu* – father's education (numeric: **0** - none, **1** - primary education (4th grade), **2** – 5th to 9th grade, **3** – secondary education or **4** – 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**)
- *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**, output target)

2 Data Preparation

The data we have obtained from *Kaggle* were provided in two different files, one containing records collected from Math course at *Gabriel Pereira* school and the second one containing students data from Portuguese course at *Mousinho da Silveira* school. To reach our goal, we firstly needed to merge the data into one single dataset, making sure to shuffle it to prevent problems related to the order of the records. In the concatenating process we have added a new

boolean column named "*math*" which highlights the source file for each record. At the end, we can do our analysis with a dataset containing 1044 records (395 from Math, 649 from Portuguese) and 34 attributes.

3 Preprocessing

The dataset is composed by quantitative and qualitative features, both of them do not present missing or duplicated values.

3.1 Choice and optimization of the dependent variable

First of all we chose the dependent variable of our analysis to be "*AlcoholConsumpt*", a variable that we calculated aggregating the attributes "*Walc*" and "*Dalc*". To understand how much weight associate to each attribute, we have summed all the occurrences in order to calculate the proportion of the variables. The obtained results were 60% weight for "*Walc*" and 40% for "*Dalc*".

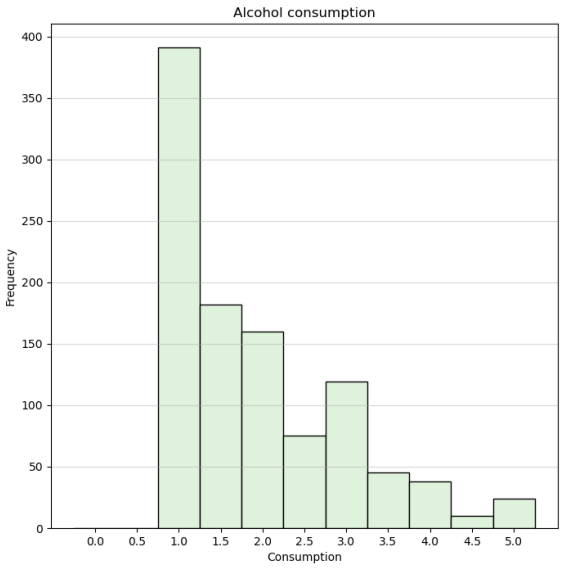

Figure 1: Alcohol consumpt distribution

In this situation we have a numeric dependent variable, "*AlcoholConsumpt*", that can range from 0 to 5. The figure 1 shows how alcohol consumption is distributed among all the values. To correctly predict the goal variable we had to turn it into a binary value. We chose a threshold of 2.0 to split values in two categories:

- **true**, assigned to values greater than or equal to 2.0
- **false**, assigned to values less than 2.0

This threshold was the optimal value to split the records into similar size classes, although a bit of class imbalance remains in the process which is not statistically relevant for the analysis. The frequency of these values is respectively 574 for false class (54.98%) and 470 for true class (45.01%). The aggregation of the variables "*Walc*" and "*Dalc*" created some type of redundancy into the dataset, so we had to remove them.

3.2 Discretization

After finding out and optimizing the prediction goal, we noticed that some attributes could be represented using a more appropriate type. Here we show the list of the updated version of some attributes:

- *sex-male*: boolean "M" = true
- *address-urban*: boolean "U" = true
- *famsize-LE3*: boolean "LE3" = true
- *Pstatus-together*: boolean "together" = true
- *guardian-family*: boolean "mother" or "father" = true
- *schoolsup*: boolean "yes" = true
- *famsup*: boolean "yes" = true
- *paid*: boolean "yes" = true
- *activities*: boolean "yes" = true
- *nursery*: boolean "yes" = true
- *higher*: boolean "yes" = true
- *internet*: boolean "yes" = true
- *romantic*: boolean "yes" = true

3.3 Feature Selection

Afterwards we faced the problem of choosing the right and most impacting attributes to analyse and predict our dependent variable. The first phase was to discard categorical attributes that limit us on selecting some types of models. In addition, for their inherent nature, these attributes should have no effect on the alcohol consumption prediction. We identified them as "*school*", "*Mjob*", "*Fjob*" and "*reason*". With this assumption, we analysed the dataset creating descriptive statistics. In particular, we computed the rank correlation matrix that allowed us to pick the most correlated features with the dependent variable.

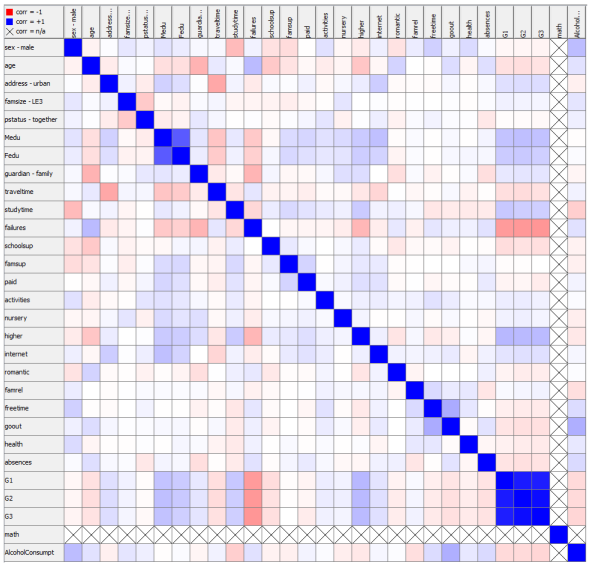

Figure 2: Features correlation matrix

Following the results shown in figure 2, we have seen that there are not significant correlations with the dependent variable "*AlcoholConsumpt*". In order to select enough features for our analysis, we included those which have a correlation higher than 0.06 and lower than -0.06. Because "*G3*" is highly related to "*G1*" and "*G2*", we only took "*G3*" to avoid distortion caused by the others. At the end of this phase, the remaining features in the dataset are "*sex-male*", "*age*", "*famsize-LE3*", "*studytime*", "*failures*", "*nursery*", "*higher*", "*famrel*", "*freetime*", "*goout*", "*absences*", "*G3*", "*AlcoholConsumpt*".

4 Classification

Classification is a type of supervised machine learning task where the goal is to categorize input data into classes. The input data, also known as instances or examples, is represented by a set of features. The process of classification involves learning a mapping or decision boundary from the input features to the class labels based on the training data. Once the model is trained, it can be used to predict the class labels of new, unseen instances.

4.1 Classifiers

Predict the dependent variable "*AlcoholConsumpt*" is not the only goal that we set for our analysis. We were also interested in making a comparison between various classifier categories. We took one classifier for each group and the list of our choices are the following:

- **Probabilistic** – Naïve Bayes (on Knime "*NaiveBayes 3.7*")
- **Separator** – Support Vector Machine (on Knime "*SPegasos 3.7*", with 500 epochs)
- **Heuristic** – Random Forest (on Knime "*Random Forest Learner*", with 500 decision trees)

In all this report, we will refer to Naïve Bayes as **NB**, to Support Vector Machine as **SVM** and to Random Forest as **RF**.

4.2 Models

For each of these classifiers, we firstly needed to settle on some models to work with. In our analysis we decided to have two major categories: holdout and cross validation. For each one we have two models, one faster and easier to execute and one more accurate that can bring more precise results.

4.2.1 Holdout

The first approach we used was "*Holdout*" and it was executed as follows: the entire dataset was split into two parts, using stratified sampling, that makes partition based on sampling strategies to keep correct proportion in the records. We also specified the random seed to ensure consistency across different runs. The training set, used to train the model, was composed by 67% of the records. The remaining records were not utilized for learning the relationship between the "*AlcoholConsumpt*" variable and the independent variables, but used in the test set to evaluate the performance of the model.

4.2.2 Iterated Holdout

This method is an evolution of the "*Holdout*", it is based on cycle and to the possibility to test more than one partitioning. At the end of the process, it performs an average to have a more stable and precise evaluation results. For each one partitioning the size of training and test set is, as before, respectively 67% and 33%. Whereas the split of the sets is inside a cycle, we had to remove the fixed random seed to prevent that all the iterations had in output the same results. The number of iterations we chose to make was 50 in order to have the most precise response possible and to cover more different scenarios.

4.2.3 Cross Validation

This model, also called K-Folds, use a different strategy to train and evaluate the analysis

made. The number of k fold you can use can vary, although the value we considered optimal for our case was 5. The amount of selected folds is used by the model to partition the dataset into k equal splits. In our case the dataset was divided into five subsets and the models were trained five times. During each iteration, four subsets served as the training set, while one subset functioned as the test set. The test set in every iteration is always a different one to prevent the same scenario to be recreated. This process facilitates a more robust comparison among all precedent models.

4.2.4 Leave One Out Cross Validation

This is the last model we implemented to make our analysis and it is also the most accurate and time consuming. It is based on the standard cross validation but it sets the k to the number of records minus one and for this reason this approach can only be used on small dataset.

5 Evaluation

To comprehend how much precise and accurate a classification procedure is, it is necessary to evaluate using some type of measures, plots and statistical comparisons.

5.1 Performance Measures

In regards to performance there are some useful indexes which can give the user an idea about how good the models are. We started focusing on *accuracy*, evaluated for each model and each classifier, and then we added the *f-measure* for a deeper comprehension of the quality of our predictions.

5.1.1 Accuracies

Accuracy is a metric that measures the proportion of correctly predicted instances out of the total instances in the dataset.

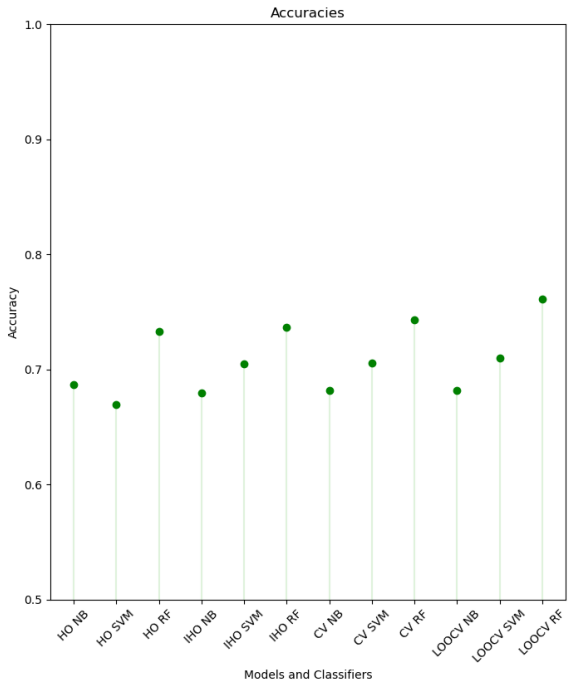

Figure 3: Model and classifiers accuracy

As we can see in the plot 3, the accuracy tends to increase with the model becoming always more advanced. This is in accordance with the theory because the partitioning become always more sophisticated. Another conclusion we can achieve is that Random Forest classifier, with this metric, over performs every other classifier with every model.

Models	NB	SVM	RF
Holdout	0.687	0.670	0.733
Iterated Holdout	0.680	0.705	0.736
Cross Validation	0.682	0.706	0.743
LOOCV	0.682	0.710	0.761

Table 1: Accuracy values

In the table 1 we can see all the accuracy values given by the models.

5.1.2 F-Measures

The F-measure, also known as F1 score, is a metric in machine learning that combines precision and recall into a single value. It is particularly useful in binary classification settings, where there are two classes, to perform a deeper level of analysis and comprehend conclusion.

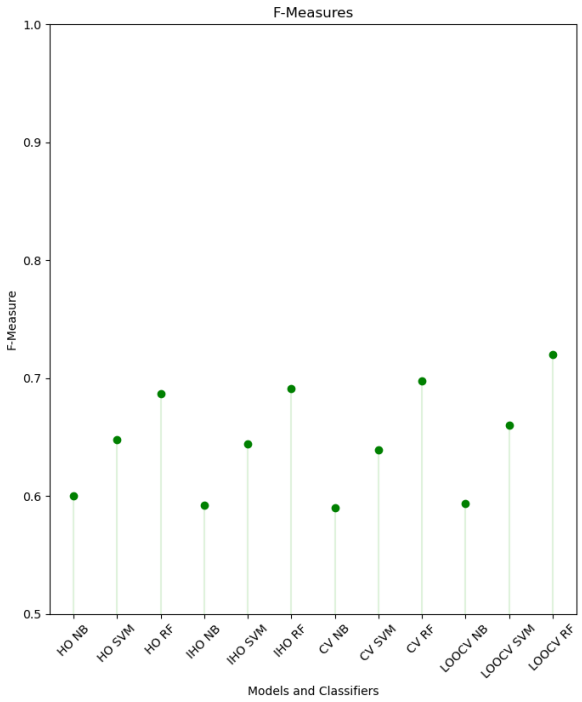

Figure 4: Model and classifiers F-Measure

As we can expect, the results shown in the figure 4, are similar to what we observed in the accuracy plot. We noticed that Random Forest is always the one that gives the best results, while the Naïve Bayes does not perform very well.

Models	NB	SVM	RF
Holdout	0.600	0.648	0.687
Iterated Holdout	0.592	0.645	0.691
Cross Validation	0.590	0.639	0.698
LOOCV	0.594	0.660	0.720

Table 2: F-Measure values

In the table 2 we can see all the F-Measure values given by the models.

5.2 ROC Curves

The ROC Curve is a graphical representation of a classification model's performance, illustrating the trade-off between true positive rate and false positive rate at different thresholds. Another useful metric related to the ROC Curve to evaluate the quality of the models is the AUC (Area Under the Curve), a single metric summarizing the overall performance of a classification model. Higher AUC values indicate better discrimination ability.

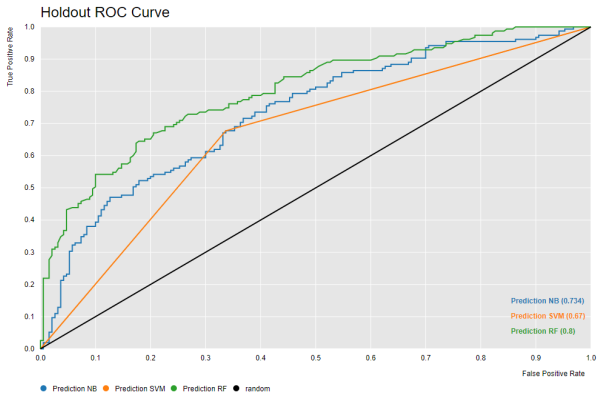

Figure 5: Holdout ROC Curve

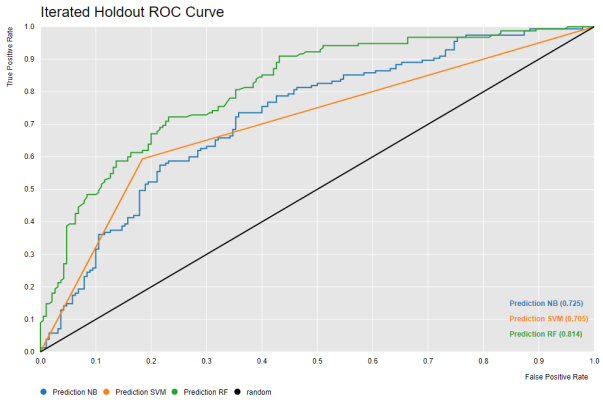

Figure 6: Iterated Holdout ROC Curve

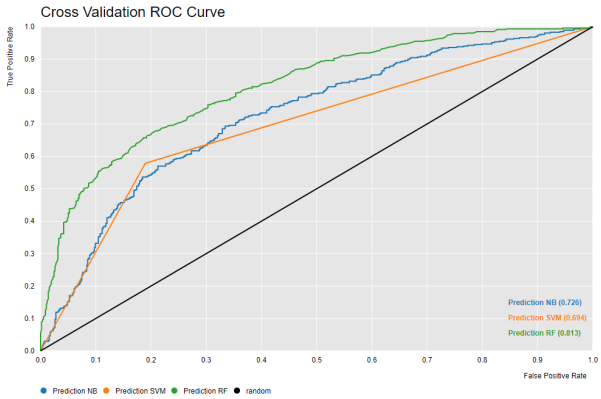

Figure 7: Cross Validation ROC Curve

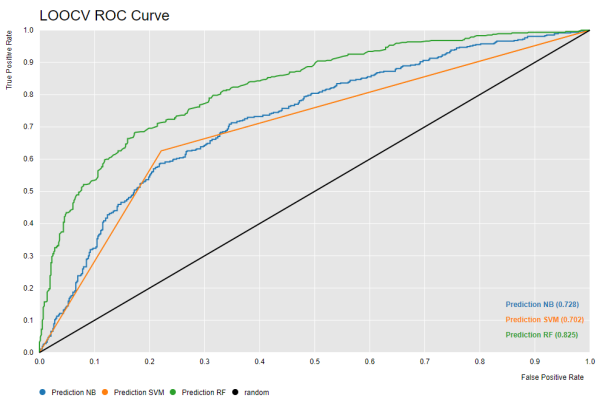

Figure 8: LOOCV ROC Curve

The first detail that stands out is that all the ROC Curves, shown in the figures 5, 6, 7, 8, for SVMs have only one decision threshold, this can be explained with the fact that the SVM's ROC tends to be linear when the problem is easily separable. Conversely to accuracy and f-measure, with this evaluation we can observe a better performance with Naïve Bayes than SVM. Although we can still see that the best classifier is the Random Forest, obtaining values of AUC around 0.8.

Models	NB	SVM	RF
Holdout	0.734	0.670	0.800
Iterated Holdout	0.725	0.705	0.814
Cross Validation	0.726	0.694	0.813
LOOCV	0.728	0.702	0.825

Table 3: ROC Curve values

In the table 3 we can see all the AUC values given by the models.

5.3 Confidence Interval

The comparison between two classifier is not easy, because it depends on the utilized models. For this reason, an additional analysis can be done with studying confidence intervals and testing statistic importance. To correctly calculate these statistical values, it is required that the two test sets are assumed to be independent and their size must be sufficiently large. For this reason, we decided to make the analysis between Iterated Holdout and Cross Validation models. In the first phase, to calculate these values, it is necessary to choose the percentage of confidence that we want to achieve. In our case this value is 95%, giving a normal distribution Z of 1.965.

Models	Lower	Upper
IHO NB - CV SVM	-0.053	0.105
IHO NB - CV RF	-0.014	0.141
IHO SVM - CV NB	-0.102	0.056
IHO SVM - CV RF	-0.038	0.115
IHO RF - CV NB	-0.133	0.024
IHO RF - CV SVM	-0.108	0.047

Table 4: Confidence Interval comparison between models

As we can see from the table 4, we calculated six confidence intervals, representing all the combinations between the classifiers of the two models selected. To take some conclusion we had to understand the sign of lower and upper limit. Based on the sign, there are three possible results:

- if 0 belongs to the interval the difference between models' error is not statistically significant
- if lower limit > 0, the second model is better than the first one
- if upper limit < 0, the first model is better than the second one

All our intervals are in the first of the three cases, which means that we cannot assume any statistical conclusion about the comparison made.

6 Conclusion

Concluding the report, in this analysis we made interesting comparison between some of the most used classification models and classifiers while doing prediction about our dependent variable, students' alcohol consumption in Portugal. Using some metrics, plot and statistical analysis we found out that the Leave One Out Cross Validation combined with the Random Forest classifier was the best pair to reach the goal. Even if the level of quality reached by our analysis is still considered good, it is not optimal.

Random Forest is usually able to manage non-linear and complex relationships between features and is very robust with respect to noisy data, outliers or irrelevant variables and has fewer sensitive hyperparameters than SVM. Furthermore, Naïve Bayes, due to its independence assumptions and limited ability to capture complex relationships between features,

probably failed to generalize as much as the Random Forest. The performance of every models is better than 0.50 that is the result given in all metrics by the random classifier. Therefore, more or less, they have learned patterns in the data.

More accurate data and more quantity of records, could potentially lead us to a more precise results and prediction. Another way to try to increase metrics values is to tweak thresholds and weights because changing them can highly affect the results of the classification.

The minimal class imbalance present between the two classes is another possible cause of noise inside the data. Removing records from the prevalent class does not lead us to a more accurate prediction.

These are all problems found in our analysis, we did some choices about them that we have explained in this report.

References

Kaggle. 2016. Student alcohol consumption. [Online; accessed 24-Jan-2024].

Knime. 2023. Knime documentation. [Online; accessed 24-Jan-2024].