2020

Student Alcohol usage prediction

Using ensemble classification and regression techniques

Lieuwe Meijdam S1047496 & Lennart Pikijn S1047468

Abstract

Predicting alcohol (ab)use in students has been the topic of several studies. Because multiple factors can influence a student to quit their education, data has been collected to research which variables might be causing this to happen. Student achievements are influenced by several features, amongst them is alcohol consumption. Given a set of social and educational attributes, this research aims to predict the alcohol consumption of the students based on these social and educational attributes.

Based on previous research, it has been determined that the prediction algorithms that will be used are the random forest classifier and various regression methods. Both of which were trained on a dataset containing the drinking habits along several social features of Portuguese and Math students. The data contains a feature which indicates alcohol usage on a scale from one to five. The random forest yielded an accuracy of around eighty percent at best. The regression methods yielded low accuracy (thirty-forty). It is however recommended to replicate this study with a larger body of data. Also, expertise from the social field should be taken advantage of.

Application domain

The application domain of this research at its broadest is the social domain. More specifically, one could use the trained models to predict possible alcohol abuse and its comorbid decline of educational performance. It must be noted that if the outcome of this research shall be used on a broader scale, the models should be generalized with different data. At this stage, the only feasible application of the models is for Portuguese and Math students in secondary school.

Related works

Related previous work that attempted to address the same or a similar problem:

In the context of the dataset itself it is proposed by the original poster that one could try to predict final grades for students; however, this has been done by a variety of other researchers (also published on Kaggle). So instead of conducting the study that was proposed it was decided to try and predict a different attribute in the data set: “Alcohol usage”.

When looking into related works there are some papers about predicting Alcohol usage with data mining algorithms. For example, there has been a study conducted by the University of Camerino (Pagnotta & Amran, 2016). In their study they used a workflow that is somewhat similar to this study, however their research solely relied on decision trees for classification. The conclusion was that there might be relations in the dataset which influenced alcohol consumption. However due to the lack of data and computing power their study was very limited. Another related work which tried to predict alcohol use took a different approach by mainly using descriptive statistics as the main focus point (Megan E. Patrick, 2009).

In the data mining field, there are also several studies conducted where alcohol use is used as a predictor for things like. lung cancer (Chauhan & Jaiswal, 2016) and depression chance (McEachin, Benjamin, & McInnis, 2008).

Next to related works in the data mining field there are also various in-depth studies about alcohol usage by young people/students and its relation to social factors. There are multiple studies that seek out the relation between romantic/sexual life and alcohol use, for example a study that was published by the Center of Alcohol studies (Cooper & Orcutt, 2000).

Other examples about this topic from different researchers are quite often posted on health journals (Karen L. Graves). Next to researches and institutions that conduct studies in this field there are also a lot of governmental sources that write about it for example the Dutch institution for statistics (CBS) (Susanne de Witt, 2018). However, the majority of the studies conducted in the field of health and alcohol use, are literature studies and data mining /advanced statistics are not commonly used.

Dataset and collection

The data used was obtained by a survey which was held under secondary school students. The survey (and thus the data) contains information about topics like: Social, gender and study information. The actual dataset was made public and was published on Kaggle (Kaggle, 2017). For an overview of the data please refer to the GitHub[4] page of this project.

Insight in the dataset (EDA)

For this chapter the guidelines used for the Exploratory Data Analysis are acquired from the EDA handbook by (Nist Sematech, 2012), the guidelines used for data visualization are acquired through the book "The Visual Display of Quantitative Information" (Tufte, 2001).

To get more insight in the dataset itself and to get a feeling for its properties this project started with checking for the validity of the data. The quality of the data was inspected and with simple Python code it was made clear that there were no data type conflicts and that none of the columns contained "Null" values. Since these two properties were already fulfilled, we did not have to take preprocessing steps to clean the dataset (Note that this is different from the preprocessing steps necessary for the algorithms to function).

Earlier in this paper it was stated that there was going to be a prediction on alcohol usage on students however it was discovered that there were two indicators in the dataset that related to this topic, namely: "Walc" and "Dalc" where Walc is the weekend alcohol usage and Dalc is the workday alcohol usage. In the data exploration both will be covered but in the execution of the algorithms only one of the two will be predicted ("Walc" will be the predicted label).

Since there are two different study orientations in the dataset, they were both visually explored in the same way to discover how similar or dissimilar they are. In the graph on the right (*Figure 1*) can be seen how both studies are represented in the dataset. With a count of 395 math students against 649 Portuguese students. (it is important to find the similarities in order to know if it is possible to merge the records of the students in both studies without negatively impacting the distribution of the attributes).

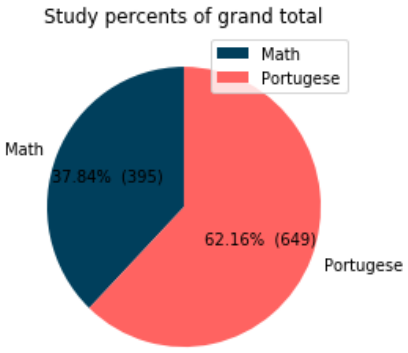

Figure 1 Study orientation distribution

In the next section there will be a comparison between math and Portuguese students where at the same time the properties of the dataset will be visually explored/explained.

When exploring the predictor label first, the distribution was made visible in *Figure 2*. When observing the bar chart and violin plot it can be seen that the predictor label has roughly the same distribution for both students of Math and Portuguese students. It is also worth noting that the majority of the students is falling into the “almost no drinking at all” category (and therefore not being a normal distribution). Since this is discovered it is important to pay attention to possible class imbalance problems in the algorithms.

To further continue inspecting the distributions of the dataset, while comparing them at the same time, similar bar plots to the ones in *Figure 2* were made for all attributes.

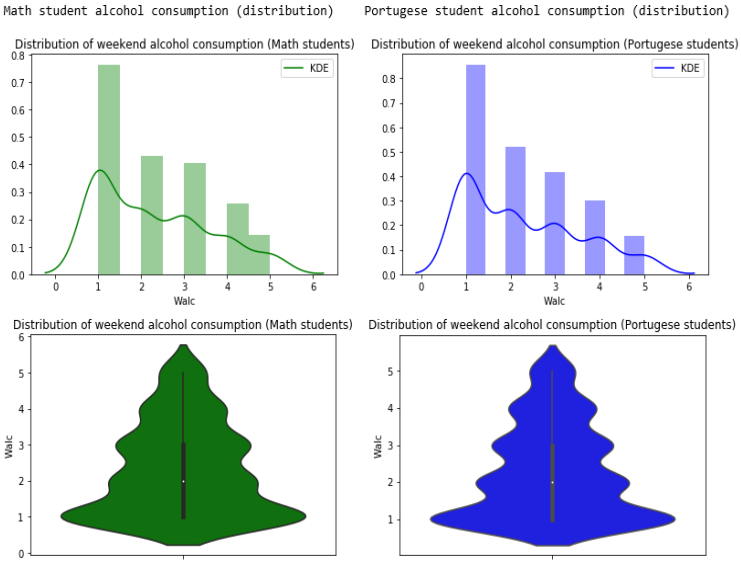

Figure 2 Predictor label distributions

Because it would be very time and space consuming to go over all of these plots, they can be inspected by referring to GitHub[4]. However, it is worth noting that in general the attribute distributions of both Portuguese and math students behave very similar and that they vary between normal distributions and others. (in the GitHub[4] these plots can also be found for a dataset with math and Portuguese students)

To see if there were attributes in the dataset that showed a strong correlation with our predictor label (“Walc”) this was tested with the help of Python code. The result showed that there was only one strong correlation (corr > 0.5 or corr < -0.5), this correlation was between the weekend alcohol consumption and the workday alcohol consumption (this was the case for both Portuguese and math students). This correlation can be observed in *Figure 3.*

Next to the correlations of attributes to the predictor labels it was also conducted between the other (numerical) attributes.

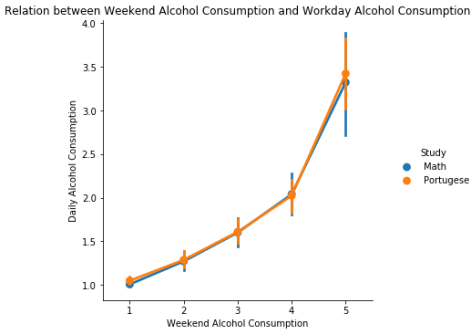

Figure 3 Correlation between weekend and workday alcohol consumption

Between the numerical attributes the following correlations were found: A correlation between Fedu and Medu meaning that students often have parents of the same educational level; Correlations between G1, G2, G3 meaning that students usually achieve similar results in different periods.

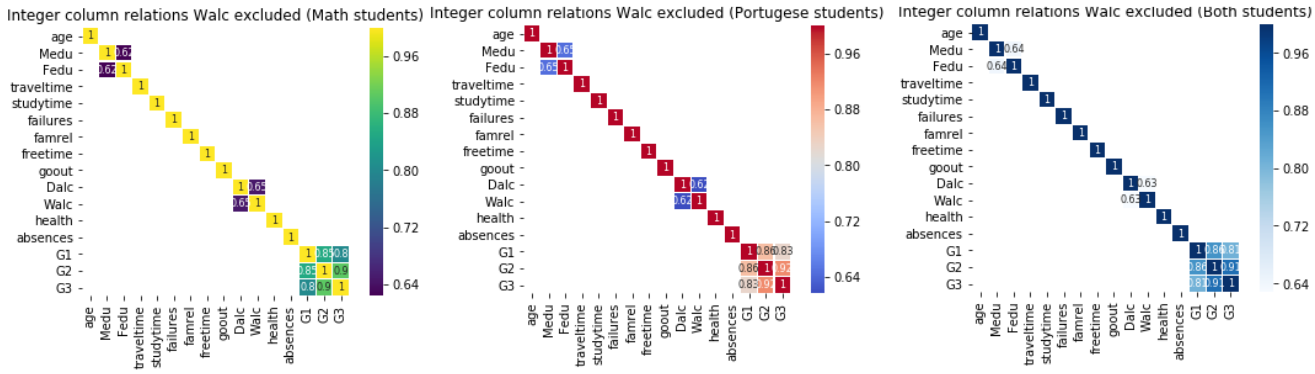

Figure 4 Correlation plots of all the numeric attributes

The impact of the correlations on the algorithms will be discussed further in this chapter.

The last step in the EDA was to inspect the feature importance of the dataset. The term “Importance” refers to the data with the most meaningful relation the label (“Walc”). The reason to conduct feature selection is that it is possible to reduce the number of features if necessary and to gain a better understanding of the relations attributes have with the label.

For this end Univariate feature selection was used, this approach will conduct statistical tests on all different attributes and will select the attributes with the strongest relationship to the label. The statistical test that was used was the chi-squared (chi^2) statistical test for non-negative attributes, these tests will display the statistical significance for categorical variables. The reason that chi^2 was used is its good use case for nominal and ordinal variables. (the attributes with the highest score were: Dalc; Absences and goout; the full list can be seen in the GitHub[4])

Methods

Random forest

The random forest classifier is an ensemble method which tries to improve the generalization performance by construction multiple decorrelated decision trees[1]. All of these trees are conducted with random sampling of training data points and random subsets of features when splitting nodes. This has as a result that the trees show a wide diversity among each other. After the algorithm is finished creating all the trees and needs to predict a label it will take a vote among all the trees, were the majority of the voted label gets chosen as a prediction. By aggregating the predictions of all the trees in the forest the algorithm is able to reduce the variance of the trees without negatively impacting their low bias (making it robust to overfitting) (Pang-Ning Tan, 2019).

Motivational arguments for the choice of Random forest are its robustness, high performance in high dimensional datasets (like the one of this study) and the ability of the algorithm to handle not only numeric but also categorical data.

Regression

The second algorithm is regression. More specifically, several variations of linear regression have been used. Linear regression fits a line with the minimal sum of squares through all data points and predicts by approximating new data to that line. Ridge regression, also known as Tikhonov regularization, is used to reduce problems with collinearity. This could be useful in models with large numbers of parameters. Lasso is a linear model which estimates sparse coefficients and prefers solutions with fewer non-zero coefficients. ElasticNet is also a linear model, but it combines L1 and L2 regularization. This combination makes learning a model possible in a case where only few of the weights are non-zero. All of these regression models are implementations of the Scikit-Learn Linear Models[2].

Motivation for the use of linear regression methods is its simplicity and ability to predict continuous labels

Optimization

To assure optimal results of the random forest classifier, RandomizedSearchCV has been used. This method randomizes the hyperparameters of the random forest and returns the configuration at which it yielded the most accurate results. Afterwards, GridSearchCV has been used.

This method cross-validates on a parameter grid, of which the parameters' values range around the values found in the randomized search for the optimal configuration.

[1] https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

[2] https://scikit-learn.org/stable/modules/linear_model.html

After this exhaustive search, the optimal configuration in a limited time will be acquired. Both of these methods are implementations of Scikit-Learn's Model Selection[3].
The parameters that were optimized/ tuned for the Random forest were (also includes the reason why they were chosen):

- the number of estimators used by the algorithm; this was important to tune since using more creates a heavier code to execute but using less means that it would theoretically be possible that some features would be missed in all subspaces used by the different trees.
- The maximum number of features, this influences the feature subset selection (default = auto; meaning no subset selection is performed). Although there are guidelines about this feature (The original creators of the algorithm stated that it should be the sqrt of the count of the attributes for classification problems), there were also several sources stating that in practice this still should be tuned. The result of the optimal parameter value was 6, since the attribute count is 41 and the sqrt of 41 is 6.4031 it seems that the original creators were right in this case. ref: "The Elements of Statistical Learning: 2nd Edition" (Hastie et. al. 2009, p. 592). This is a parameter that controls the form of randomness in the model.
- Max depth is a very important parameter which controls the maximum allowed depth of the trees. In general, a larger depth means more specific information and a smaller depth means more general information. Although max depth is described here as an important feature it was not included in the hyperparameter tuning. The reason for this is that according to the book "Introduction to data mining" (Pang-Ning Tan, 2019) a Random Forest ensemble should run with unpruned trees.
- Minimal samples required to split a node, to control the allowed size of leaves.
- The minimal samples of the leaves, to control the minimum number of samples in all of the leaves.
- Bootstrapping, this controls whether or not all of the data is used to fit the model. If this is turned off there will be no random variation between trees with respect to the selected examples at each stage. This is also a parameter that controls the randomness of the model.

For regression, since by definition there is only one optimal line, assuring optimal results was done by running its code multiple times. This way, the (randomized) splitting of the data happened several times, yielding slightly different results each time. Also, the hyperparameters were tuned in the same way as with the random forest classifier, but it was found that the default parameter settings already yielded the optimal results except for one variant of regression. To attempt to further increase accuracy, only the best features discovered in the EDA have been used to generate new regression models.

Since linear models do have some kind of assumption they make on the data there were precautions that were taken. For example, linear regression models assume that the residuals follow a normal distribution (see GitHub[4]) and this is a requirement in order to work smooth. Neighboring residuals cannot have correlations with each other (multicollinearity). The precautions to both problems can be seen in GitHub[4].

Evaluation

The models are both trained on a segment of the total data, and then tested on the remaining data. Predictions will be made on the test data and will be compared to the actual value to eventually calculate an accuracy percentage. To conclude whether a model is useful or not, a threshold of 75% accuracy has been established. Also, a confusion matrix will show how well the models have performed at predicting the actual class the students belong to.

[3] https://scikit-learn.org/stable/modules/classes.html#module-sklearn.model_selection

Results

Random forest

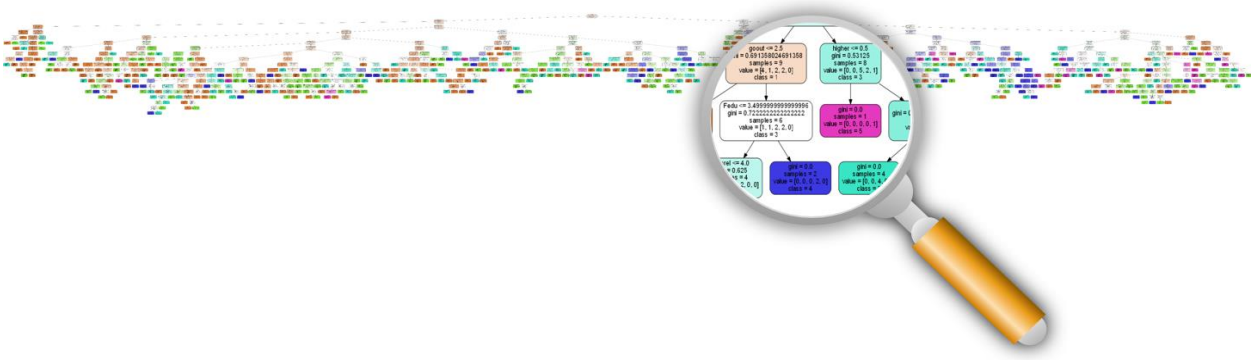

Figure 5 Random instance of a tree in the ensemble

The model for the classification algorithm “Random Forest” was developed in multiple iterations. These iterations can be observed and compared in *Figure 6*. The Accuracy metric was used to express the quality of the iteration. To be able to visually inspect the predictions of the iterations a confusion matrix was chosen as visualization technique since it is well suited for multiclass labels. (note that all of the confusion matrices below were made with the same random states)

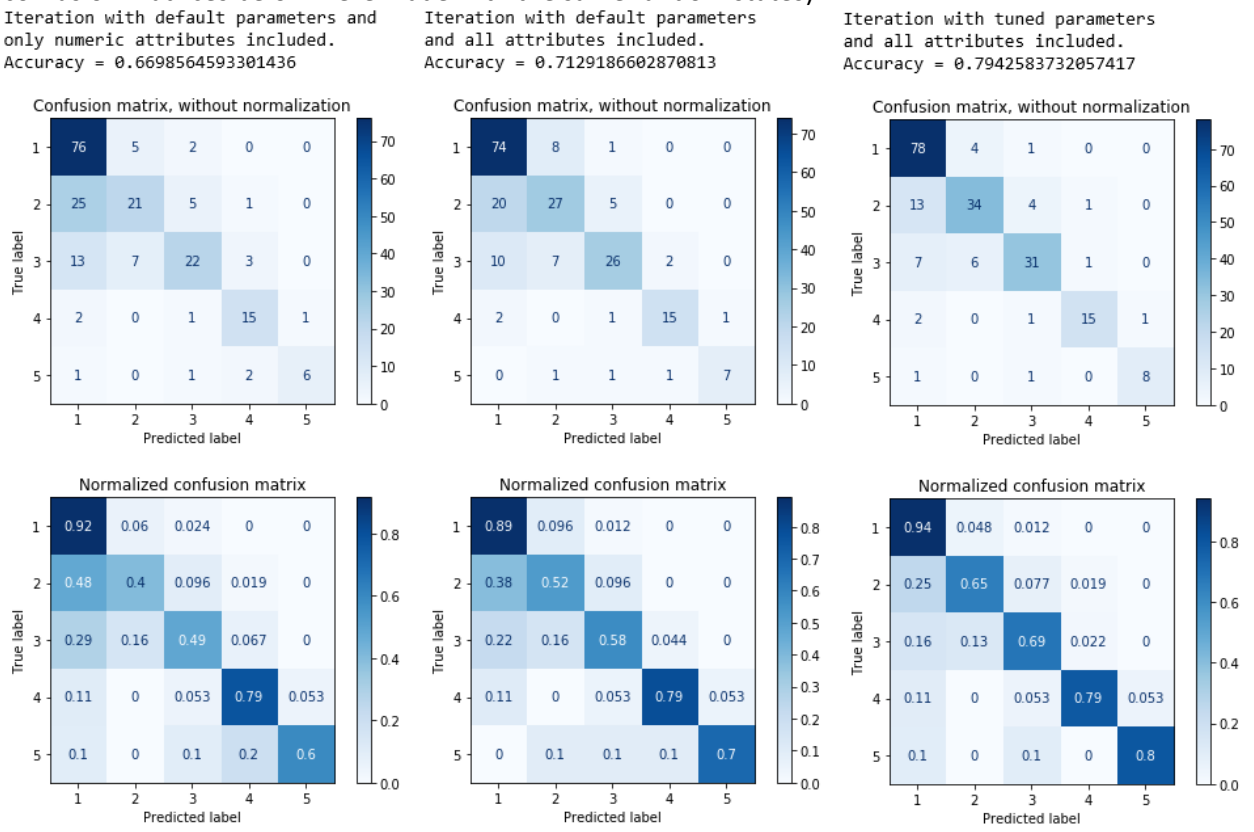

Figure 6 Confusion matrices by iteration

To further explain all of the iterations:

First iteration: In the first iteration of the trained model there were only numeric attributes inserted to the Random Forest algorithm along with the default parameters (please note that this means that the categorical attributes were ignored in this iteration). In the first iteration it can be observed that there is quite some visible spread and that some of the labels get predicted quite bad (especially label 2).

Second iteration: After observing the results of this iteration and referring to the feature importance research done in the EDA part of this project it was chosen to include all of the attributes (so also the categorical attributes) in the second iteration. After the second iteration the accuracy increased but it was still below the threshold of (0.75).

Third iteration: To further improve the accuracy of the model, parameter tuning was used. The third (and final) iteration was therefore conducted with all the attributes and with the tuned parameters. More about the tuned parameters and their roles can be found in the method section. In the last iteration it can be observed that the predictions overall are more likely to be correct and that the wrong predictions lie closer to the actual label. (more on this topic in the recommendation section)

Regression methods

Iteration 1 - using all features: in the dataset to predict a class for a student has been unsuccessful in the first several runs of the regression models, yielding at most 35% accuracy.

Iteration 2 - with parameter tuning: Practically the same results as the previous iteration.

Iteration 3 - After feature selection: the performance of all regression methods only slightly improved. On average, 40% accuracy was the best result. The results after the feature selection have been visualized in the confusion matrices *Figure 7*.

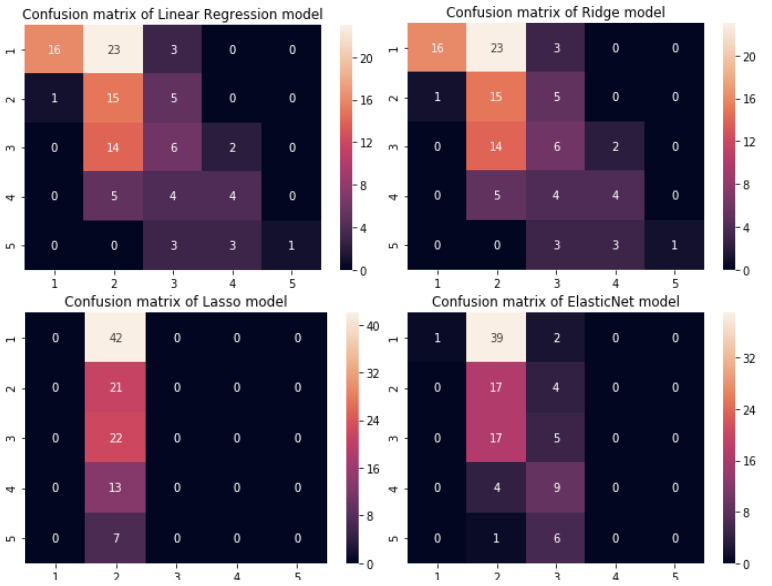

Figure 7 Confusion matrices for various regression techniques

Comparison of methods

After successfully implementing multiple models based on different techniques, the models could be compared with each other in terms of prediction accuracy. In the figure/table (*Figure 8 and Table 1)* on the right the different models can be compared in total accuracy as well as in label specific performance.

When inspecting the results, it can be said that random forest generally performed with an adequate accuracy overall (with a slight exception for label2). When inspecting the linear regression methods however it seems that they seem to mainly suggest label 2 which is likely due to underfitting.

Figure 8 Method comparison

Model	Label 1 Accuracy	Label 2 Accuracy	Label 3 Accuracy	Label 4 Accuracy	Label 5 Accuracy	Total Accuracy
ElasticNet	0.056	0.793	0.238	0.000	0.00	0.2857
Lasso	0.000	1.000	0.000	0.000	0.00	0.2762
Linear Regression	0.333	0.724	0.238	0.250	0.29	0.4095
Random Forest	0.946	0.590	0.727	0.708	0.80	0.7905
Ridge	0.333	0.724	0.238	0.250	0.29	0.4095

Table 1 Accuracy scores per level (and total)

Conclusion

The random forest classifier has performed the best, being able to use all of the attributes in the dataset. The various linear regression methods yielded an accuracy of 40% at best, making them less suited for the goal of this study. A likely reason for the lower scores of the regression methods is that this study solely used linear regression methods. Because certain requirements to use a linear regression model to its fullest (residuals should be normally distributed for example) were not met, the algorithm could not be used to its fullest potential.

The random forest method however managed to get scores that satisfied the 75% threshold, with scores going up as high as 80% accuracy (but generally around 75%). The reason that the method performed well (especially given the limitations) is that the data is able to use the stronger points of random forest. Another reason is that due to hyperparameter tuning a good configuration was discovered.

From the results and multiple executions of the algorithm on different splits it also becomes clear that although one of the methods used exceeded the accuracy threshold, it was not very stable. This stability issue has a direct relation with certain limitations of the dataset, judging from the results it can be said that the methods have a hard time (especially the regression methods) predicting labels with few data points. Also, the fact that there are as little as 100 label 5 points there might be a possibility that patterns that are found are random/ noise.

In summary of the above: The Random Forest algorithm managed to perform accurately enough to satisfy the 75% threshold, although more data samples would influence its stability positively. The various linear regression techniques did not manage to get scores that satisfy this same threshold, these methods seemed to be a bad match for this dataset due to its distributions/ shape.

Recommendations

In this study there were several interesting factors and relations found in terms of alcohol usage for students. However, the study was also limited in some factors, these factors were deep understanding of the application field (alcohol usage and healthcare for students). The other factor was a limitation on the dataset that was used, it was useable but better results would have been achievable with more datapoints.

It surely is a recommendation to further study the topic of this paper, since it has a clear application which is interesting from a social as well as a scientific perspective. However, it is recommended to attend to the two limitations stated above, this could be done by doing the next study with a multi-disciplinary team which combines field experts and data mining experts. With this combination it could for example be possible to verify the models with a deeper understanding, for example by using a weighted accuracy metric (e.g. a true label of 5 but a predict of 1 influences the model accuracy quite badly and a true label of 2 and a predict of 3 weights less heavy). In this study this was considered but was not implemented because of the lack of field specific knowledge, so it was hard to verify if this approach (or a similar) one was valid to do.

In case of further attempts of regression models, it is recommended to look into options for non-linear multi class regression types. For the reason that these regression types might be a better fit.

The last recommendation for further studies in this topic is to try to expand the dataset. In our study we did see negative influences caused by the fact that there were only 100+- label 5 records in the dataset. Since this amount is not very vast it was hard for the algorithms to accurately detect relations because it could have been random with this arbitrarily amount.

Code insight

For reproducibility and transparency, the code used in this research is available on GitHub[4] and can be seen as an appendix of this document.

The code is accompanied with comments to clearly explain how the results were achieved.

Bibliography

al., H. e. (2009). *The Elements of Statistical Learning: 2nd Edition.*

Chauhan, D., & Jaiswal, V. (2016). *An efficient data mining classification approach for detecting lung cancer disease.* Retrieved from 2016 International Conference on Communication and Electronics Systems: https://ieeexplore.ieee.org/abstract/document/7889872

Cooper, M. L., & Orcutt, H. K. (2000). *Alcohol use, condom use and partner type among heterosexual adolescents and young adults. .* Retrieved from Journal of studies on Alcohol and Drugs: https://www.jsad.com/doi/abs/10.15288/jsa.2000.61.413

Kaggle. (2017). *Student Alcohol Consumption.* Retrieved from Kaggle: https://www.kaggle.com/uciml/student-alcohol-consumption

Karen L. Graves, P. (n.d.). *Risky Sexual Behavior and Alcohol Use among Young Adults: Results from a National Survey .* Retrieved from https://journals.sagepub.com/doi/abs/10.4278/0890-1171-10.1.27

McEachin, R. C., B. J., & McInnis, M. G. (2008). *Modeling gene-by-environment interaction in comorbid depression with alcohol use disorders via an integrated bioinformatics approach.* Retrieved from BioData Mining: https://biodatamining.biomedcentral.com/articles/10.1186/1756-0381-1-2

Megan E. Patrick, L. W.-L. (2009). *The Long Arm of Expectancies: Adolescent Alcohol Expectancies Predict Adult Alcohol Use.* Retrieved from Oxford Academic: https://academic.oup.com/alcalc/article/45/1/17/121663

Nist Sematech. (2012). *Exploratory data analysis.* Retrieved from Engineering statistics handbook: https://www.itl.nist.gov/div898/handbook/eda/section1/eda11.htm

Pagnotta, F., & Amran, H. (2016). *Using data mining to predict secondary school student alcohol consumption.* Retrieved from https://bradzzz.gitbooks.io/ga-seattle-dsi/dsi/dsi_05_classification_databases/2.1-lesson/assets/datasets/STUDENT%20ALCOHOL%20CONSUMPTION.pdf

Pang-Ning Tan, M. S. (2019). *Data Mining.* Pearson.

Susanne de Witt, C. H. (2018). *Alcoholgebruik onder jongeren.* Retrieved from CBS: https://longreads.cbs.nl/jeugdmonitor-2018/alcoholgebruik-en-gezondheid/

Tufte, E. R. (2001). *The Visual Display of Quantitative Information.* Garb.

[4] https://GitHub.com/LennartPNL/DataMiningProject

Table of figures

Table of tables