M1 - Stakeholder

Malika Kuhlman Hansen
Michael Dybdahl

Aalborg Universitet
September 2019

1 Definition of the problem

We have found a data set containing data about students alcohol consumption, their grades and several social conditions.

Problem statement

Are there any patterns of different groups of the students based on the features. Can we compute a model that predict if a student is heavy drinker and can a model predict the students overall grade?

2 Description of data acquisition

The data set was found on Kaggle.

1. Link (dataset): click here
2. Link (original dataset): click here
3. Link (Our upload): click here
4. Link (Colab): click here

In the second link there is a description of all the variables in the data set. We do not look into all the variables, but we usually use the following variables

- **School**, *GP or MP*
- **Sex**, *F = 0 or M = 1*
- **Age**
- **Dalc**, *Alcohol consumption in workdays, 1-5*
- **Walc**, *Alcohol consumption in weekends, 1-5*
- **Absences**, *number of school absences, 0-93*
- **G1, G2, G3**, *first, second, final periods grades*
- **Binge_drinker**, *High alcohol consumption in Weekends, low in workdays*
- **Heavy_drinker**, *High alcohol consumption in both weekends and workdays*
- **Overall_grade**, *Mean of G1, G2, G3*

3 Data Preparation

3.1 Recoding

We renamed *X1* to *ID* and we changed some of the variables from *double* to *factor*.

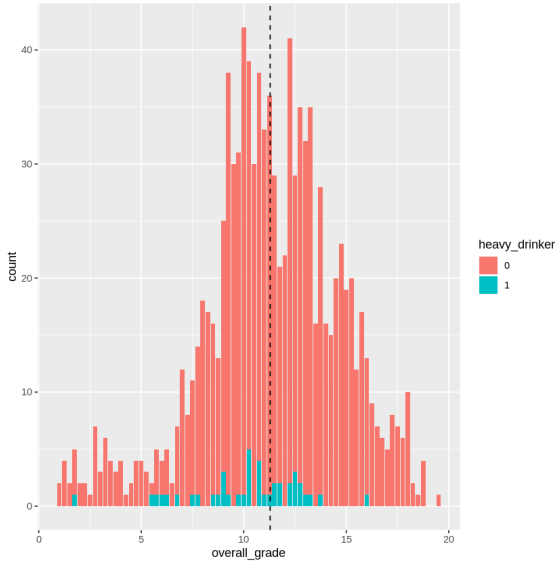
4 Exploratory data analysis

4.1 Relevant! summary statistics

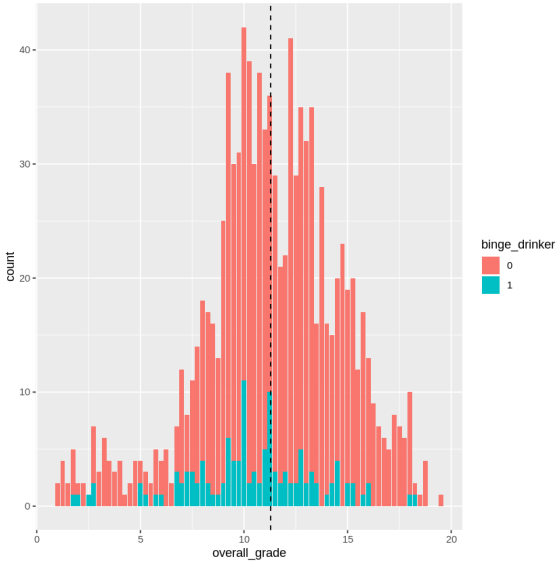
We split the data set, one with Females, one with Males, one with the school GP and one with the school MP. We have collected some of the results in the following table.

	#Observations	#Binge	#Heavy	Mean Grade	Max Grade
All	1044	120	42	11.29	19.5
F	591	42	5	11.37	18.75
M	453	78	37	11.17	19.5
GP	772	95	32	11.58	19.5
MS	272	25	10	10.46	18.75

We see that 17% of the Males are *Binge* but only 7% of the Females and 8% of the Males who are *heavy* drinkers and only 0.8% of the Females. There are way more Females who are *binge* drinkers, than *heavy* drinkers, it might be because it is normal to party in the weekends. If we look into the schools we see that there are 12% of the students binge drinker and 4% heavy at GP and 9% of the students at MS are binge drinkers and 3.6% are heavy drinkers. There are almost the same percentages of students at the two schools who are heavy drinkers.



Figur 1: Bar plot of the overall_grade colored by heavy_drinkers

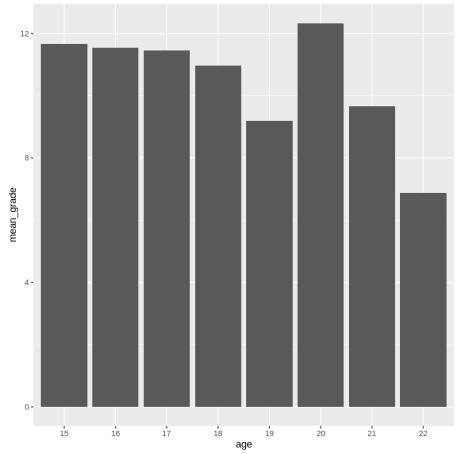


Figur 2: Bar plot of the overall_grade colored by binge_drinkers

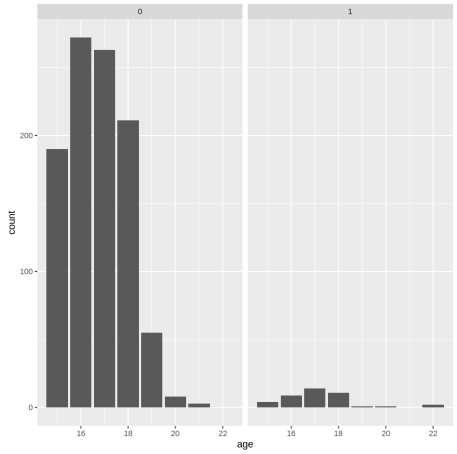
At figure 1 and 2 we look into how the overall_grade was distributed, where we colored the heavy and binge drinkers grades. Surprisingly we see that the heavy_drinkers grades is mainly around the mean grade. We see that one or two heavy drinkers get really low grades, but most of them have grades around the mean grade. If we look into the binge drinkers we see that there are more students getting lower grades, but there are also more students getting high grades.

At figure 3 we look into grades of the students grouped by their age, we can see that the 22 yo students have a low mean grade and the younger students almost have the same mean

grade. Which could mean that when you are young you want to learn and be the best, When you are older, you might have failed a lot and just don't care about your education anymore. Maybe the 22 yo students tend to drink more than the younger students, lets have a look at that.



Figur 3: Col plot showing the mean grade across the students age



Figur 4: Bar plot of the distribution of the heavy drinkers across the age.

When we look at figure 4, then we see that all the 22 yo and only some of the youngest are heavy drinkers. There are two 22 yo students that are heavy drinkers, but at the same time there are fourteen 17 yo students who tend to be heavy drinkers. Therefore the alcohol consumption may be the reason that the 22 yo students have lower grades than the younger.

6 Unsupervised ML

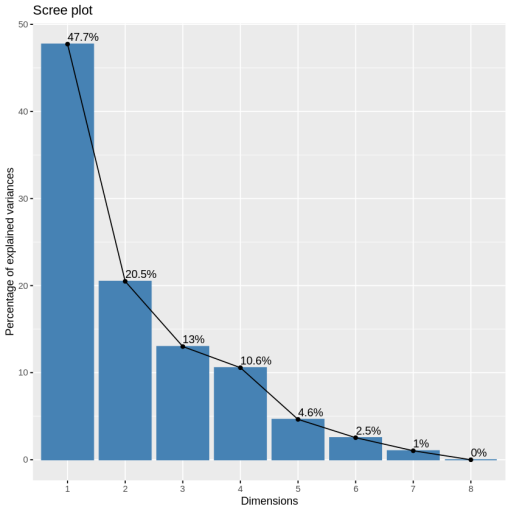
After scaling the data we computed the PCA on the data and made a scree plot of the explained variance for each dimension.

At figure 5 we see that dimension 1 explains 47.7% of the data, We are supposed to look for the bend in the graph, but we could also look into the eigenvalues. WE should choose dimensions with eigenvalues > 1, therefore we chose to use three dimensions.

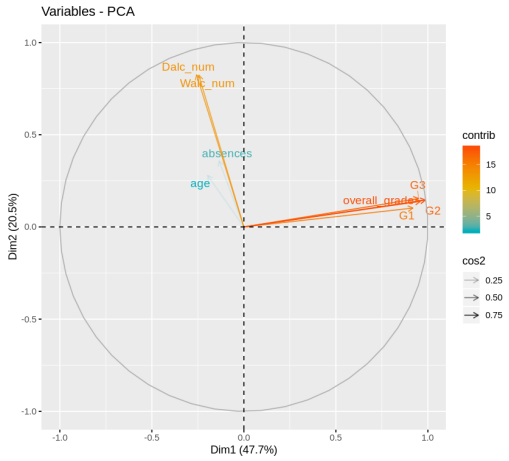
At figure 6 we see that the grades have a high contribution on the first dimension and the Dalc and walc have a high contribution on the second dimension.
We wanted do clustering at our data. By looking into the total within sum of squares, we chose four clusters. After dividing the data into these four clusters, we could then describe each clusters observations by:

- **Cluster 1**: This cluster contain students with the lowest grade, low alcohol consumption and low absence.

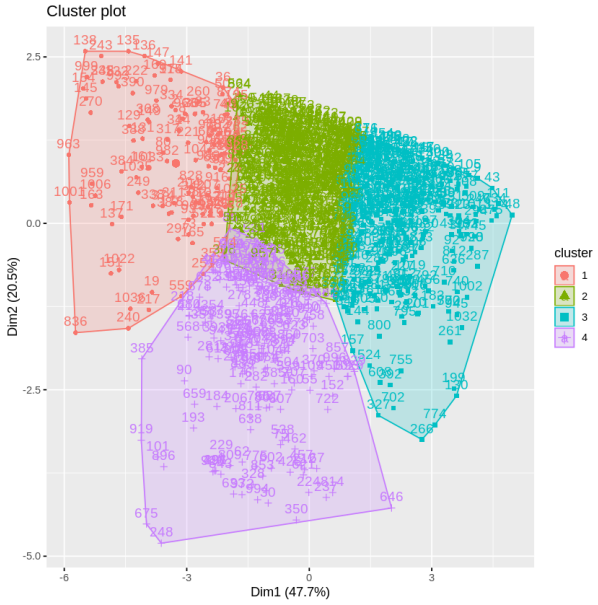
- **Cluster 2**: This cluster contain students with the average grade, low alcohol consumption and average absence.
- **Clutser 3**: This cluster contain students with the high grades, low alcohol consumption and low absence.
- **Cluster 4**: This cluster contain students with the around average grades, high alcohol consumption and high absence.



Figur 5: Plot showing the explained variance for each dimension.



Figur 6: 2-dimensional space showing the contribution of the variables



Figur 7: plot of four the clusters

We see at figure 7 that cluster 3 is placed to the right in the plot, which tells us that this cluster contains of students with high grades. Cluster 1 is placed to the left at the plot which tells us that it contains students with low grades. Cluster 2 is in the middle, so it may contain the average students. Cluster 4 is placed in middle of dimension 1, but very low at dimension 2, so it contains the student that drinks.

7 Supervised ML

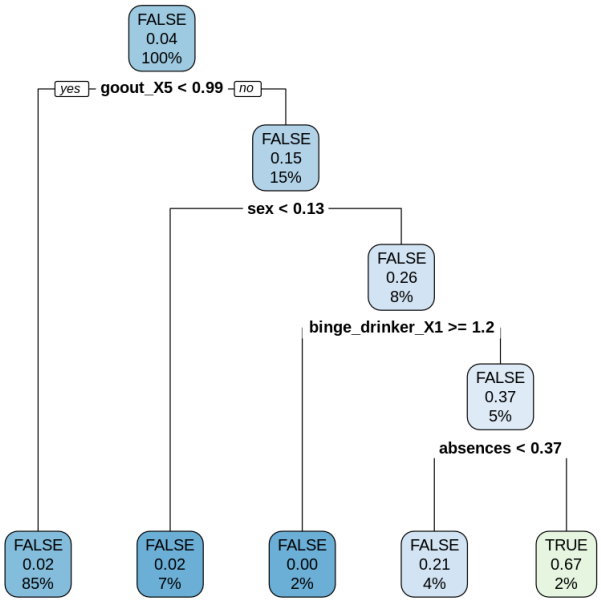
We are going make two different types of machine learning, *Classification* and *Regression*.

7.1 Classification of heavy drinkers

Here we make two classifications using *logistic regression* and *decision tree*. We want to predict the heavy_drinkers. We use the following variables:

- **School**
- **Sex**
- **Age**
- **Absences**
- **Health**
- **Romantic**
- **PStatus**
- **Medu, Fedu**
- **Failures**
- **Goout**
- **Overall_grade**
- **Binge_drinker**
- **Heavy_drinkers**
- **G1, G2, G3**

At figure 8 we see the decision tree our model uses to predict if a student is heavy drinker and it shows that if you don't go allot out with friends, are male, is not a binge drinker and have high absence. Then there are 67% chance that you are a heavy drinker.



Figur 8: Decision tree

At the same time we can see that there are 2% chance that a student is a heavy drinker, if the condition "going out with friends"is below very high, which applies to 85% of the students.
Now we compare the original values of heavy drinkers with the results from the two models:

	Test set	Logistic regression	Decision tree
TRUE	250	247	254
FALSE	10	13	6

If we should conclude if our model are good only using this table, we would say that the logistic model is good, since it predicts 13 TRUE values out of 10 and the Decision tree only predict 6 TRUE Values. The problem with this table is that we cannot see the TRUE/TRUE predictions, here we have to use a Confusion Matrix.

Logistic Regression

	FALSE	TRUE
FALSE	240	7
TRUE	10	3

Decision tree

	FALSE	TRUE
FALSE	246	8
TRUE	4	2

From these two tables we see that it is only 3 out of the 13 predictions that were TRUE/TRUE. And the Decision tree only predicted 2 out of the 6 as TRUE/TRUE. If we look into the Sensitivity for both models, we see that the logistic model has a sensitivity at 30% and the decision tree has a sensitivity at 20% which is low. So our model is not the best models to predict if a student is heavy drinkers.

8 Regression

Now we make a regression, where we predict the students overall_grade. We use the following variables:

- **School**
- **Sex**
- **Age**
- **Absences**
- **Health**
- **Romantic**
- **PStatus**
- **Medu, Fedu**
- **Failures**
- **Goout**
- **Overall_grade**
- **Binge_drinker**
- **Heavy_drinkers**
- **Walc, Dalc**

Before we ran the regression model, we knew that many of the variables would be insignificant. Therefor we reduced the model with the variables with the highest p-value, step by step. This gave us a reduced model only including the variables:

- **School**
- **Failures**
- **Overall_grade**
- **Binge_drinker**
- **Romantic**

By choosing these variables our model had an R^2 value at 19.62% which tells that this model only explains 19.62% of the data, and we considered that as low.
The $RMSE$ was 2.98 which tells that the model predict 2.98 wrong. This was considered

very high, and it is connected to the low explained variance. We needed a much higher R^2 value to reduce the $RMSE$ and thereby make better predictions of the grade.
A summary of the model is given below:

$$\begin{aligned}\text{Overall\_grade} = 12,1795 - 0,4799 \cdot \text{romantic1} - 3,1863 \cdot \text{failures1} - 3,5118 \cdot \text{failures2} \\ - 4,7338 \cdot \text{failures3} - 0,8829 \cdot \text{binge\_drinker1}\end{aligned}$$

We can conclude that this model does not predict the overall_grade very well.

8 Conclusion

We found this data set with Portuguese students and their alcohol consumption as well as further social conditions, which included 1044 students and 53 variables.

Exploratory data analysis

We could see that the average overall grade was 11,28 and that females average grade was a bit higher then males. Also the GP school had markedly higher grades than MS.
We looked at some visualisations, where we could see that the binge and heavy drinkers had almost the same distribution as those who wasn't binge or heavy drinkers. We also looked at how the grades where distributed by the age, where the older students had lower grades than the younger students, which may because this students have had trouble by passing their classes.

Unsupervised ML:

As we computed unsupervised machine learning on the data, we could see by the PCA, that the numeric variables could be reduced to three dimensions. These dimensions explained the grades, the alcohol consumption, the absence and age.
Then we found four clusters using Kmeans, where these clusters contained the bad, the average, the good and the students with high alcohol consumption.

Supervised ML:

After the unsupervised ML, we would try to predict if a student was a heavy drinker or not, using two models: logistic regression and decision tree. These models did not perform well, when predicting if a student was a heavy drinker, which probably is because of the low number of heavy drinkers in the data set.
We also computed a regression model, which should predict the overall grade. This model included five response variables, which gave a R^2 on 19,62% and $RMSE$ around 3, which we considered as not that good. If the model should perform better, it need more explanatory variables, which could increase the R^2 (explained variance) and reduce the $RMSE$, so the predictions would be better.