Introduction to Statistical Machine Learning for Bioinformaticians and Medical Informaticians

SoSe 2022

Tutor: Dana Petracek

1	Σ

Marina Dittschar & Clarissa Auckenthaler

Assignment 8

(Submitted 04.07.2022)

Task 1

The two datasets consist of 11 continuous feature variables (fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, alcohol) and the response variable `quality`. The response was determined by at least three different wine experts, which are scores between 0 and 10. Overall we have 4898 observation of white wines and 1599 oberservations in the red wine data.

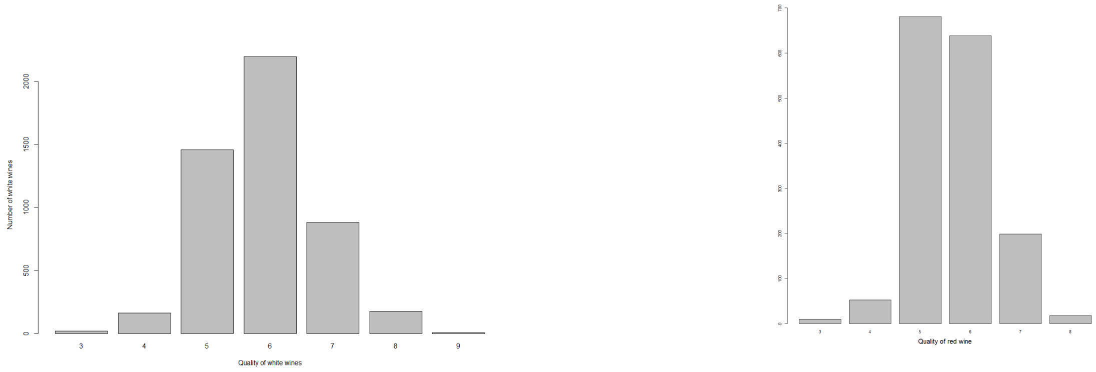
First we get an overview of our data, the table 1 displayes the minimum, maximum and mean value of each feature/attribute. Also we checked if the dataset contains any nan values. It does not so we can easily use the whole data set without any further modification.

Attribute	**White wine**			**Red wine**		
	Min	Max	Mean	Min	Max	Mean
fixed.acidity	3.8	14.2	6.85	4.6	15.9	8.32
volatile.acidity	0.08	1.10	0.278	0.12	1.58	0.528
citric.acid	0	1.66	0.334	0	1.0	0.27
residual.sugar	0.6	65.80	6.391	0.9	15.5	2.54
chlorides	0.009	0.346	0.046	0.012	0.61	0.087
free.sulfur.dioxide	2.0	289.0	35.31	1.0	72.0	15.87
total.sulfur.dioxide	9.0	440.0	138.4	6.0	289.00	46.47
density	0.98	1.039	0.994	0.99	1.0	0.997
pH	2.72	3.820	3.188	2.74	4.0	3.31
sulphates	0.22	1.08	0.489	0.33	2.0	0.658
alcohol	8.0	14.20	10.51	8.4	14.9	10.42

Table 1: Output of the summary of the two datasets (before normalization)

Wine Quality Distribution

Figure 1a shows the distribution of the quality for the white wine dataset and figure 1b the distribution on the quality for the red wine dataset. As we can see, the lowest quality for white wine is '3' and the highest is '9', the other values (1,2 and 10) do not occur in this set. In the dataset of the red wine the highest quality is a '8'. It looks like majority of the white wines have a quality between '5' to '7'. There are only 5 with the highest quality '9' and dataset is normally distributed. Also for the red wines the majority has a quality between '5' to '7' but for red wines there are less ranked as a '7' compared to the white wines. The mean of the quality for white wines ('5.8') is closer to '6' than for red wines ('5.3').



(a) Distribution of the quality for white wines

(b) Distribution of the quality for red wines

Figure 1: Plots of the distribution for the two datasets

Linear relationship

The correlation between dependent and independent variables can help us to come up with a initial list of importance. Before making prediction, we will visualize the correlations of the variables in a heatmap (5) and have an closer look on the correlations coefficients how strongly the response variable quality is correlated with the different features. We can directly see in figure 5, that the response quality is correlated with the feature alcohol for both data sets.

Correlation among variables:

For white and red wine the feature alcohol is moderately correlated with the response variable quality (2. But then the at sets differ: chlorides and density have the second and third highest correlation coefficient for white wine, where as for red wine the second and third highest correlation coefficient are volatile.acidity and sulphates. In the heatmaps 5 there are some linear correlations: alcohol is moderately correlated with the density of wine and pH is moderately correlate with fixed acidity (both datasets). In the white wine dataset density is strongly correlated with residual sugar. In the red wine data set it differs again, the total sulfur dioxide is moderately correlated with free sulfur dioxide and ph is strongly correlated with citric acid.

This table 2 and the heatmap 5 give us a first idea of the importance and from this perspective we can see, that alcohol has the highest correlation coefficient with quality.

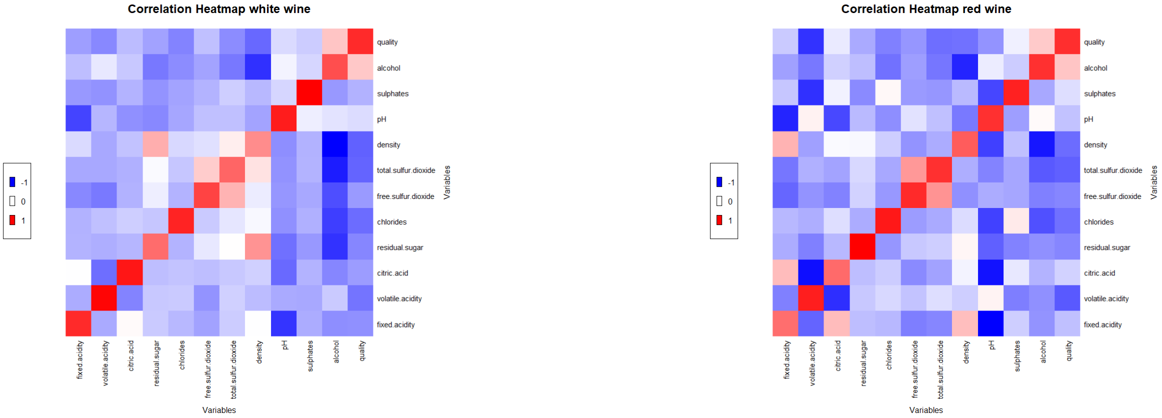
Feature	**Correlation coefficient**
alcohol	**0.435574715**
pH	0.099427246
sulphates	0.053677877
free.sulfur.dioxide	0.008158067
citric.acid	-0.009209091
residual.sugar	-0.097576829
fixed.acidity	-0.113662831
total.sulfur.dioxide	-0.174737218
volatile.acidity	-0.194722969
chlorides	**-0.209934411**
density	**-0.307123313**

(a) Correlation coefficients white wine data set with the response variable *quality*

Feature	**Correlation coefficient**
alcohol	**0.47616632**
sulphates	**0.25139708**
citric.acid	0.22637251
pH	-0.05773139
fixed.acidity	0.12405165
residual.sugar	0.01373164
free.sulfur.dioxide	-0.05065606
chlorides	-0.12890656
density	-0.17491923
total.sulfur.dioxide	-0.18510029
volatile.acidity	**-0.39055778**

(b) Correlation coefficients red wine data set with the response variable *quality*

Table 2: Correlation coefficients quality with each feature for both datasets



(a) Correlation coefficients of white wine data

(b) Correlation coefficients of red wine data

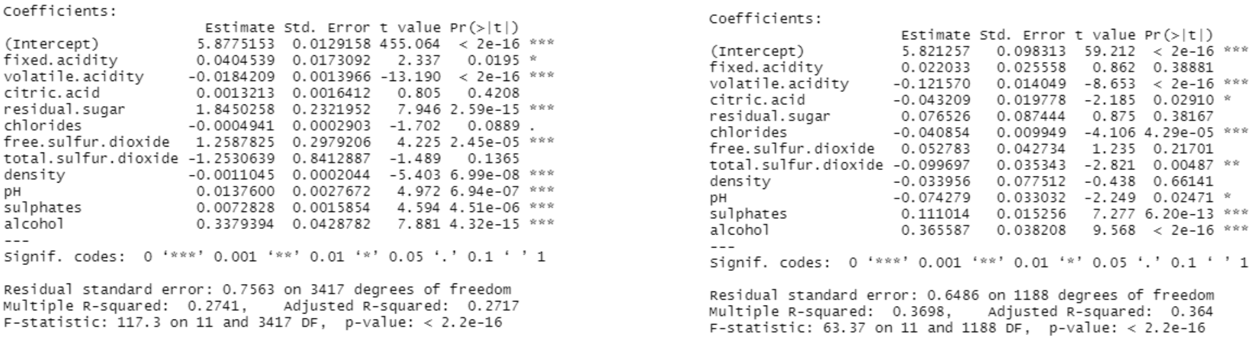
Figure 2: Correlation coefficients plotted as Heatmap for both datasets

For the modelling we split the datasets with a ration of 70% training data and 30% test data, because its often used in practice and normalized the data on the mean and variation of the training data.

Linear Models

As a regression problem we fitted a multivariate model, a model for ridge regression and one for lasso. With the modeling of a linear regression model we wanted to find out which variables are significant.We fitted a model with the function lm() and calculated the train and test MSE. We wanted to look if we get better results for ridge regression or lasso for both models (ridge and lasso) we selected the model with the best λ value and the lowest MSE.

In figure 3 we can identify that for white wines chlorides which had before a slight correlation with the quality, are now not significant. For red wines the the correlated values stayed significant. But for both models we can clearly see in table 3 that the mean squared error is quite high for both datasets. The linear model performed of red wine performed a bit better than the white wine. For our white wine data the linear regression has the worst test MSE (0.55) and the ridge and lasso performed minimal worse (0.56) but overall we can say that the linear model is not the best model to predict the quality for white wines with this data. Also for the red wine data the MSE is quite high but at least a bit lower as for the white wines. There the ridge regression generate the lowest MSE compared to the other methods, but still not really good in predicting the quality of red wines. In general using linear models to explain the variation in quality is difficult, because actual results are integer and predicted results are not integer. This amplifies that the linear model is not a 'good' model for our problem.



```
Coefficients:
                      Estimate Std. Error t value Pr(>|t|)    
(Intercept)          5.8775153  0.0129158 455.064  < 2e-16 ***
fixed.acidity        0.0404539  0.0173092   2.337   0.0195 *  
volatile.acidity    -0.0184209  0.0013966 -13.190  < 2e-16 ***
citric.acid          0.0013213  0.0016412   0.805   0.4208    
residual.sugar       1.8450258  0.2321952   7.946 2.59e-15 ***
chlorides           -0.0004941  0.0002903  -1.702   0.0889 .  
free.sulfur.dioxide  1.2587825  0.2979206   4.225 2.45e-05 ***
total.sulfur.dioxide -1.2530639  0.8412887  -1.489   0.1365    
density             -0.0011045  0.0002044  -5.403 6.99e-08 ***
pH                   0.0137600  0.0027672   4.972 6.94e-07 ***
sulphates            0.0072828  0.0015854   4.594 4.51e-06 ***
alcohol              0.3379394  0.0428782   7.881 4.32e-15 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7563 on 3417 degrees of freedom
Multiple R-squared:  0.2741,	Adjusted R-squared:  0.2717 
F-statistic: 117.3 on 11 and 3417 DF,  p-value: < 2.2e-16
```

```
Coefficients:
                      Estimate Std. Error t value Pr(>|t|)    
(Intercept)           5.821257   0.098313  59.212  < 2e-16 ***
fixed.acidity         0.022033   0.025558   0.862  0.38881    
volatile.acidity     -0.121570   0.014049  -8.653  < 2e-16 ***
citric.acid          -0.043209   0.019778  -2.185  0.02910 *  
residual.sugar        0.076526   0.087444   0.875  0.38167    
chlorides            -0.040854   0.009949  -4.106 4.29e-05 ***
free.sulfur.dioxide   0.052783   0.042734   1.235  0.21701    
total.sulfur.dioxide -0.099697   0.035343  -2.821  0.00487 ** 
density              -0.033956   0.077512  -0.438  0.66141    
pH                   -0.074279   0.033032  -2.249  0.02471 *  
sulphates             0.111014   0.015256   7.277 6.20e-13 ***
alcohol               0.365587   0.038208   9.568  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6486 on 1188 degrees of freedom
Multiple R-squared:  0.3698,	Adjusted R-squared:  0.364 
F-statistic: 63.37 on 11 and 1188 DF,  p-value: < 2.2e-16
```

(a) Summary lm model white wine

(b) Summary lm model red wine

Figure 3: Summary output of lm-model for both datasets

	Model	Train MSE	Test MSE
white	linear regression	0.5700157	0.5504366
	ridge regression	0.5720781	0.5575995
	lasso	0.5743404	0.5599133
red	linear regression	0.4164206	0.4221144
	ridge regression	0.4170434	0.4183778
	lasso	0.4185819	0.4193416

Table 3: MSE values for both data sets and different models (linear regression, ridge regression and lasso)

Classifying the Quality of wines

Because our linear models did not predict well, we decided to classify the quality into to categories. "good wines" and "bad wines".

In order to perform classification, we first had to transform the response variable. This was the case because 9 distinct response classes produced patchy and incomplete predictions on classification approaches. We therefore split the response data into two classes, in accordance with the distribution of the response variable shown in Figure 1 a). We divided the data into "good" wines with a score of 7 or higher and "not good" wines with a score of at most 6. This was encoded in our models with "1" for "good" wines and "0" for "not good" wines.

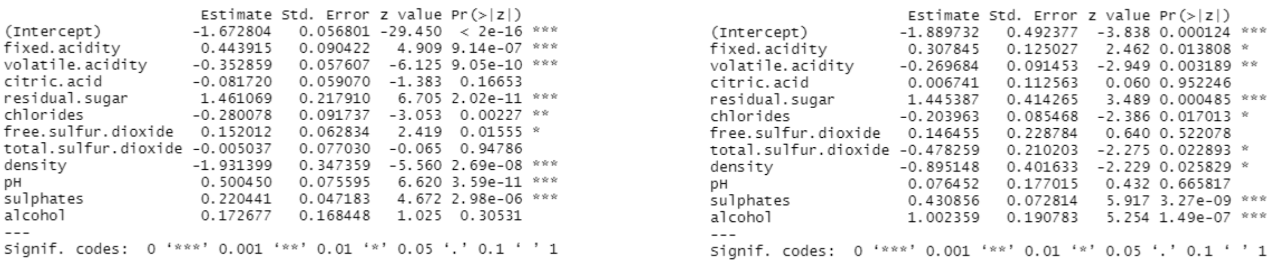
Logistic Regression

We then performed logistic regression of the data set and then classified the results by classifying all samples above 50% as "good" and all samples below that threshold as "bad". The resulting test accuracy was 81.42%, the train accuracy was 79.91%. Highly significant variables were "fixed.acidity", "volatile.acidity", "residual. sugar", "chlorides", "density", "pH" and "sulfates" (see Figure 4a). Significant was also "free.sulfur.dioxide". Interestingly, "alcohol" was not significant. When looking at the confusion matrix, it becomes clear that while overall test accuracy was quite good, the sensitivity for "good" wines was quite low at 26.5%.

For the red wine data set, test accuracy was 86.22% and train accuracy was 88.33%. Highly significant variables were "volatile.acidity", "residual.sugar", "sulphates" and "alcohol" (see Figure 4b). Significant variables were "fixed.acidity", "chlorides", "total.sulfur.dioxide" and "density". Again, the sensitivity for "good" wines was very low at 24.07% (see Figures 4).

	Predicted class	true "bad"	true "good"
white	"bad"	1102	219
	"good"	54	94
red	"bad"	311	41
	"good"	14	13

Table 4: Confusion matrices for Logistic Regression Classification



```
                        Estimate Std. Error z value Pr(>|z|)    
(Intercept)            -1.672804   0.056801 -29.450  < 2e-16 ***
fixed.acidity           0.443915   0.090422   4.909 9.14e-07 ***
volatile.acidity       -0.352859   0.057607  -6.125 9.05e-10 ***
citric.acid            -0.081720   0.059070  -1.383  0.16653    
residual.sugar          1.461069   0.217910   6.705 2.02e-11 ***
chlorides              -0.280078   0.091737  -3.053  0.00227 ** 
free.sulfur.dioxide     0.152012   0.062834   2.419  0.01555 *  
total.sulfur.dioxide -0.005037   0.077030  -0.065  0.94786    
density                -1.931399   0.347359  -5.560 2.69e-08 ***
pH                      0.500450   0.075595   6.620 3.59e-11 ***
sulphates               0.220441   0.047183   4.672 2.98e-06 ***
alcohol                 0.172677   0.168448   1.025  0.30531    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

(a) Logistic regression coefficients fitted on white wine data set

```
                        Estimate Std. Error z value Pr(>|z|)    
(Intercept)            -1.889732   0.492377  -3.838 0.000124 ***
fixed.acidity           0.307845   0.125027   2.462 0.013808 *  
volatile.acidity       -0.269684   0.091453  -2.949 0.003189 ** 
citric.acid             0.006741   0.112563   0.060 0.952246    
residual.sugar          1.445387   0.414265   3.489 0.000485 ***
chlorides              -0.203963   0.085468  -2.386 0.017013 *  
free.sulfur.dioxide     0.146455   0.228784   0.640 0.522078    
total.sulfur.dioxide -0.478259   0.210203  -2.275 0.022893 *  
density                -0.895148   0.401633  -2.229 0.025829 *  
pH                      0.076452   0.177015   0.432 0.665817    
sulphates               0.430856   0.072814   5.917 3.27e-09 ***
alcohol                 1.002359   0.190783   5.254 1.49e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

(b) Logistic regression coefficients fitted on red wine data set

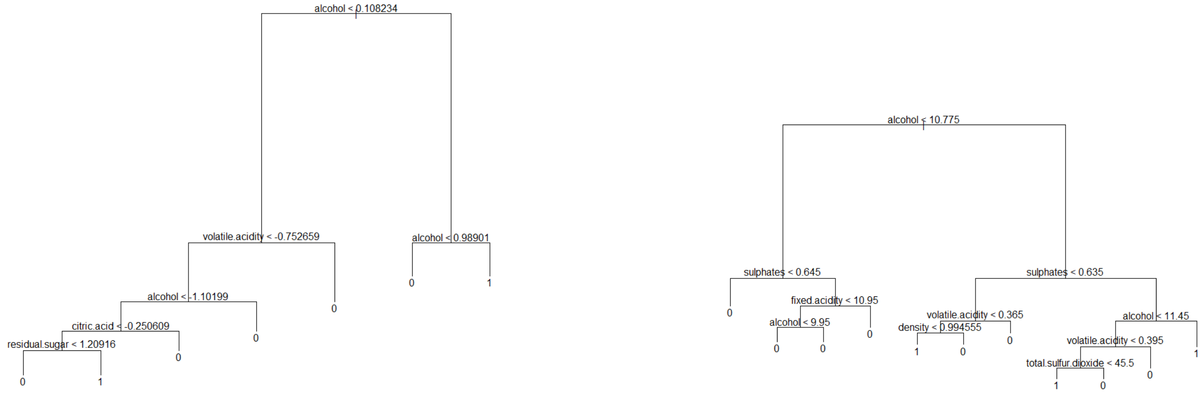
Figure 4: Logistic Regression Coefficients

Trees and random Forest

We first fit a tree with the default hyper-parameters on the data (see Figure 5a). The decision variables in this approach were alcohol, volatile.acidity, citric.acid and residual.sugar. The reported classification test accuracy for the white wine data set was 79.78%. In the fitted tree for the red wine data set (see Figure 5b), the decision variables were alcohol, sulphates, fixed.acidity, density volatile.acidity and total.sulfur.dioxide. The reported classification test accuracy for the red wine data set was 85.46%. In both cases, while overall accuracy was reasonably good, the classification on the "good" wines was not very precise, with a sensitivity of below 50%. Note that some branches on the tree fitted on the red wine data set result in the same values, this could be alleviated by pruning the tree.

	Predicted class	true "bad"	true "good"
white	"bad"	316	29
	"good"	29	25
red	"bad"	1023	164
	"good"	133	149

Table 5: Confusion matrices for the fitted trees

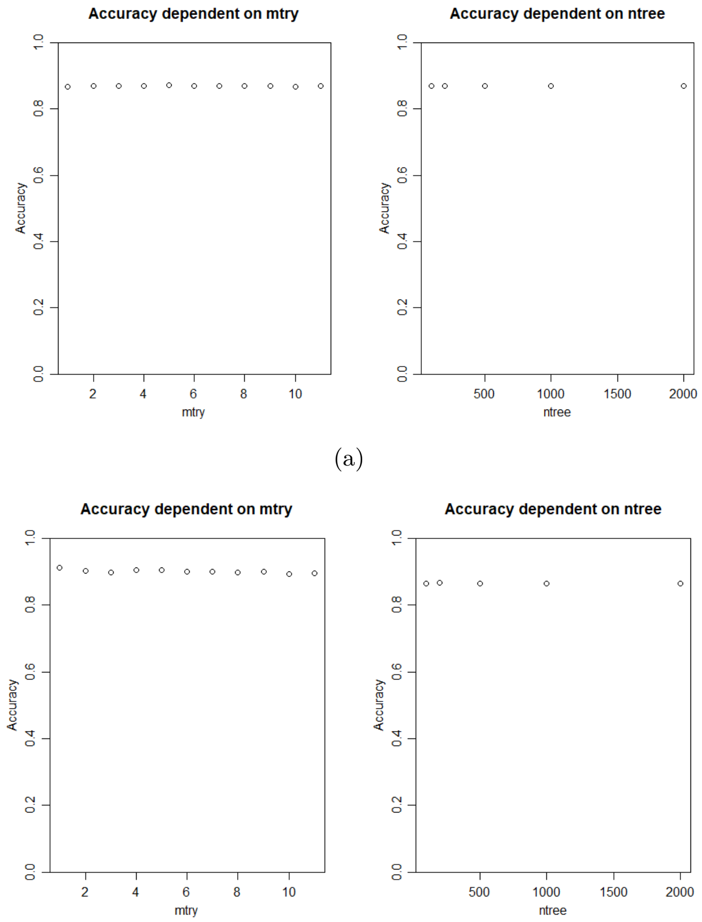


(a) Tree fitted on white wine data set

(b) Tree fitted on red wine data set

Figure 5: Trees fitted on red and white wine data sets with default hyperparameters

To estimate whether the important variables that were determined in the tree fitting were also important for other classification approaches, we also fitted a random forest model on the data sets. We tuned both the hyperparameters "ntree" and "mtry" and found that accuracy did not change by a large margin. This applied on both the white and red wine data set.



(a)

(b)

Figure 6: Accuracy dependent on different values for *mtry* and *ntree*, for the white (a) and red (b) wine data set. Evidently, the accuracy does not change significantly by tuning the hyperparameters.

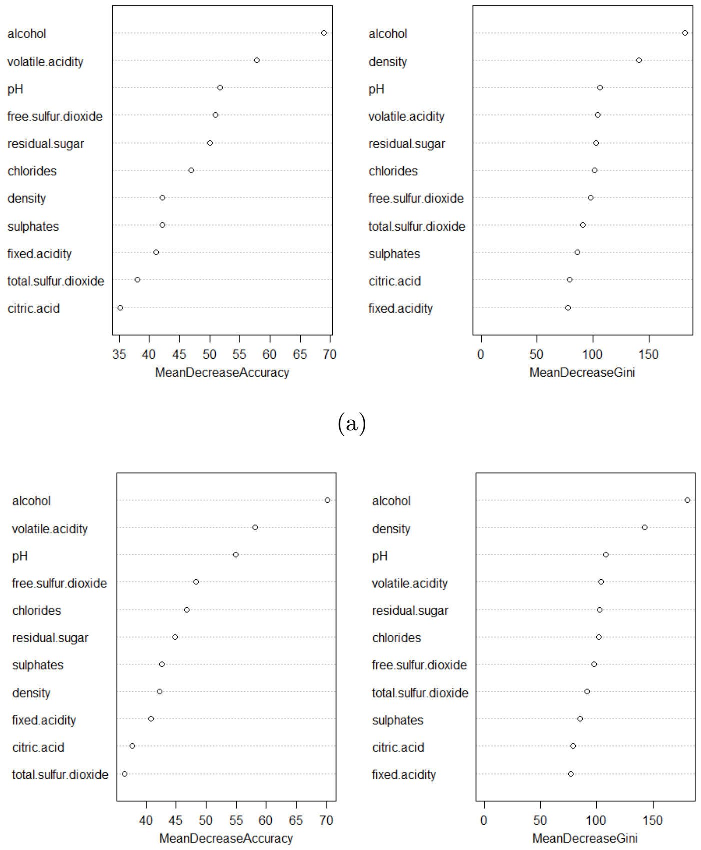
In Table 6, there are displayed the confusion matrices for the classificatino on the test sets of the white and red wine data set. Looking at the confusion matrix for the random forest model, one can clearly see that while the classification accuracies for the red and white wine data set were fairly similar, the random forest model for the red wine data set overclassifies nearly all samples as "bad" wines. This could be due to different make up of the data sets (smaller data set or less "good" wines, see Figure 1).

	Predicted class	true "bad"	true "good"
white	"bad"	1108	143
	"good"	48	170
red	"bad"	344	53
	"good"	1	1

Table 6: Confusion matrices for the random forest models

In terms of variable importance, the four most important variables with the highest Mean Decrease Accuracy for the white wine data set were: "alcohol", "volatile.acidity", "free.sulfur.dioxide" and "residual.sugar". The four highest values for Mean Decrease Gini had the variables "alcohol", "density", "pH" and "chlorides" (see Figure

7a). The values and order of the most important variables was extremely similar for the red wine data set (see Figure 7b).



(a)

(b)

Figure 7: Mean Decrease Accuracy and Mean Decrease Gini for white (a) and red (b) wine respectively. Order and magnitude for the variables are very similar in both cases.

PCA

In order to find out whether dimension reduction could be performed on the data set, we performed PCA on the normalised white wine data set (see Figure 8). The first principal component accounted for 29.29% and the second principal component accounted for 14.32%. Therefore we can see that PCA did not manage to find principal components that explained large portions of the variance. In addition to this, while some structure can be seen in the scatter plot with each data point coloured in according to its quality score, there are no resulting clusters (see Figure 8).

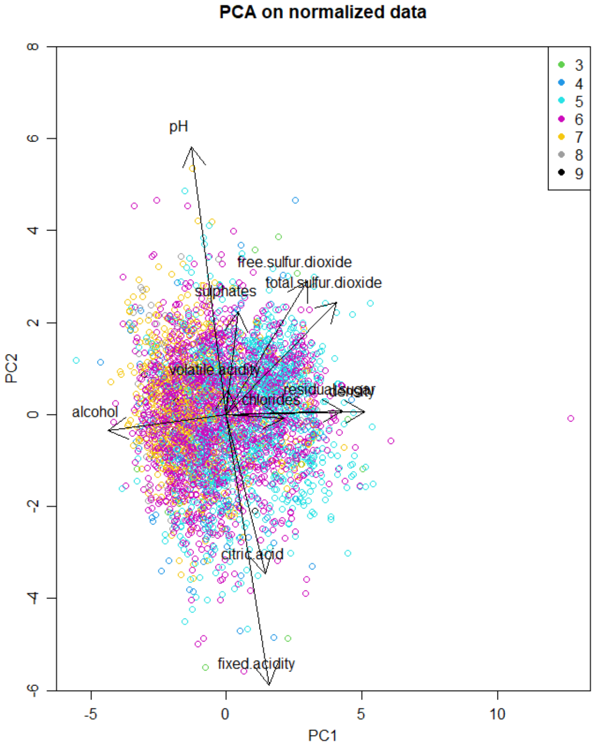


Figure 8: PCA on the normalized white wine data set

Reflection

We explored that for all our models the variable importance is almost similar: for the white wine we started with the moderately correlate variables (alcohol, chlorides and density), the linear model canceled the chlorides out (not significant) the volatile acidity is also one of the significant variables for predicting the white wine quality, only the logistic regression model of white wine displayed a not significant value for alcohol. In the Tree model alcohol, volatile acidity, citric acid and residual sugar are the decision variables for the classification and the model random forest also came up with alcohol, density and pH as main important variables. As we can see for each model the important variables are slightly different but nearly all of our models showed up with alcohol as the main important variable, except logistic regression. What is not really surprisingly based on the correlation for the beginning. The best performance of the prediction is reported by logistic regression and the tree model.

The quality of red wine started with the moderately correlated variables (alcohol, sulphates and volatile acidity the linear model supported them also. Even in the tree model alcohol, sulphates and volatile acidity are main decision variables, the model added fixed acidity and total sulfur dioxide to the list of decision variables. Only the random forest model differs a bit more and came up with alcohol, density and pH as main important variables (same as for the white wines).

Because some of our models do not have a good fit based on mean squared error value or accuracy we are not sure how accurate the values are for the variable importance. Even our PCA did not mange to separate the data well.

Overall we can say, that alcohol is for both data sets a very important factor in determining high quality wines.