How to pick good wine

Quality classification

Pedro Rio
MSc Data Science and Engineering
Instituto Superior Técnico
97241

Contents

1 Introduction

Several classification methods are used in this report with the objective of selecting high quality wines from a dataset[1] that incorporates information regarding the quality and chemical composition of several green wine samples of the red variety from the north of Portugal.

In the dataset description and exploration, some differences between high and low quality wines are discussed, as well as moderate correlations between some of the variales. For each class, outliers are identified and winsorised, producing a clean dataset. That dataset, its standardisation and the top standardised principal components extracted from it are all used in the classification task.

Several models are used in the classification, ranging from logistic regressions, support vector machines with a polynomial and with a radial kernel, linear discriminant analysis and flexible discriminant analysis. In each method, the training data is used to fit and optimise the model parameters. Afterwards, each model's performance is computed in the testing data and the best model is selected.

Finally, the results obtained are discussed, future work is proposed and final considerations are presented.

2 Exploratory data analysis

2.1 Data description

This dataset contains 1599 observations and no missing values in eleven chemical features and a quality assessments for green wine samples of the red variety from the north of Portugal. These variables are fixed acidity ($g_{tartaric\ acid}/dm^3$), volatile acidity ($g_{acetic\ acid}/dm^3$), citric acid (g/dm^3), residual sugar (g/dm^3), chlorides ($g_{sodium\ chloride}/dm^3$), free sulfur dioxide ($mg/dm^3$), total sulfur dioxide (mg/dm^3), density (g/dm^3), pH, sulphates ($g_{potassium\ sulphate}/dm^3$) and alcohol ($vol.\%$) and the quality assessment is the median of at least 3 quality assessments from different enologists in a score from 0 to 10.

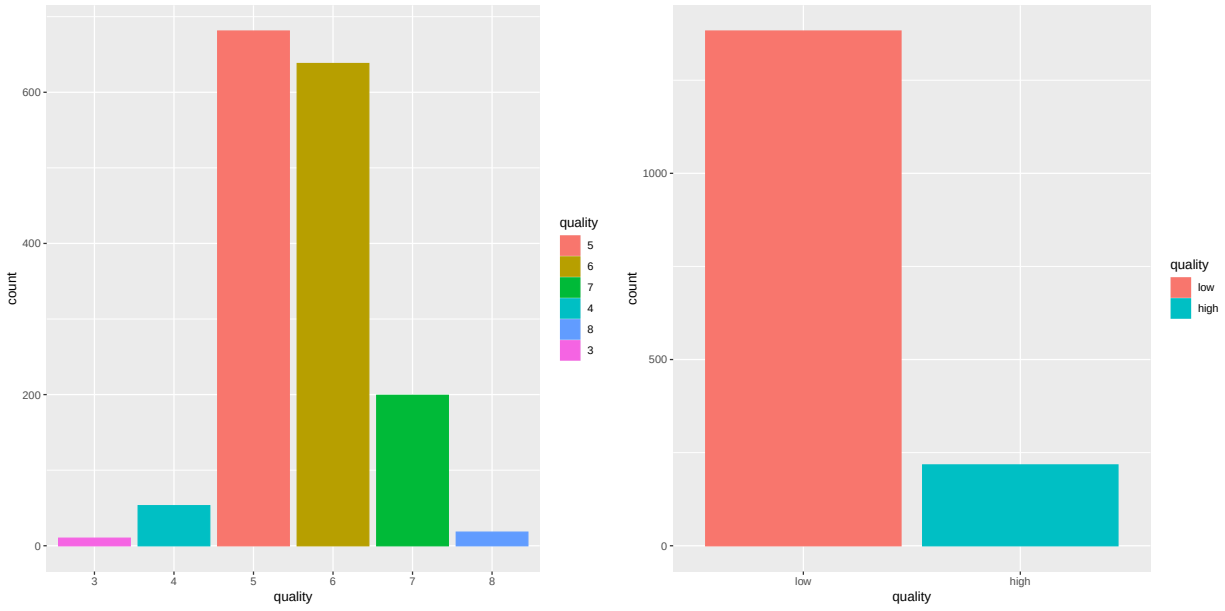

Figure 2.1: Observation counts per quality class

As the goal of the project is the classification of wines in good and bad qualities, the quality factor will be consolidated into a simpler factor with a low quality class from classes 3, 4, 5 and 6 and a high quality class from classes 7 and 8. Nevertheless, the quality classes remain quite unbalanced, with 1382 belonging to the low quality class and only 217 to the high quality class, which will give more weight to the results of the class with the most number of observations if this is not taken into consideration.

Noticeable from figures 2.2 to 2.6, the distributions of alcohol, sulphates, density, volatile acidity and fixed acidity per quality class are different and that on average wines with high quality have more alcohol, sulphates and fixed acidity and less density and volatile acidity.

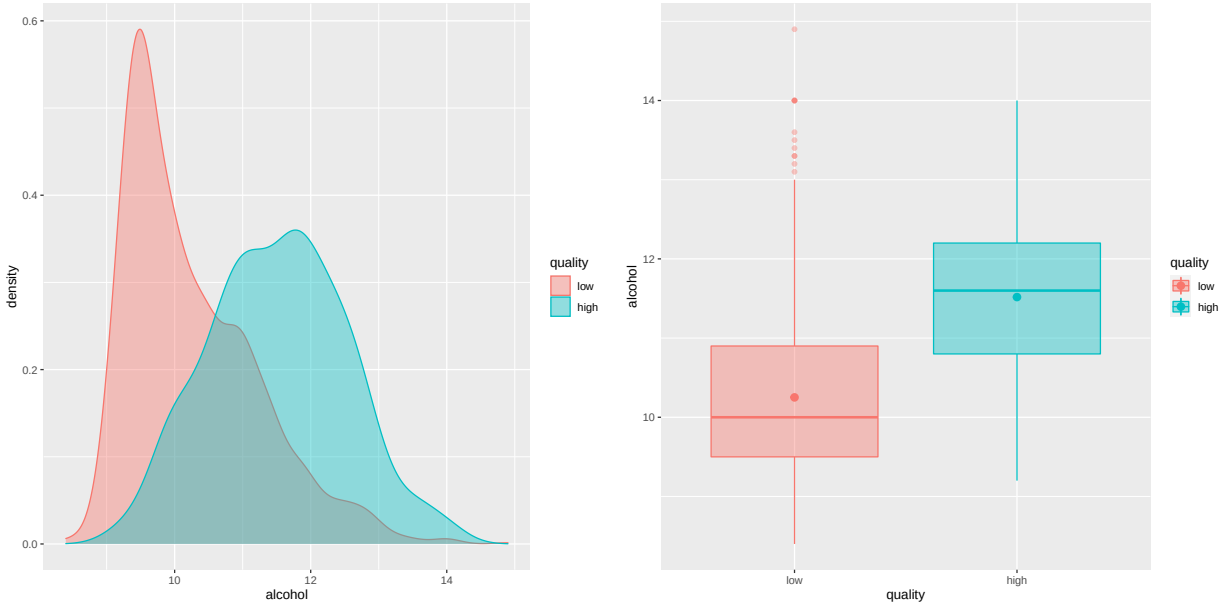

Figure 2.2: Alcohol distribution per quality class

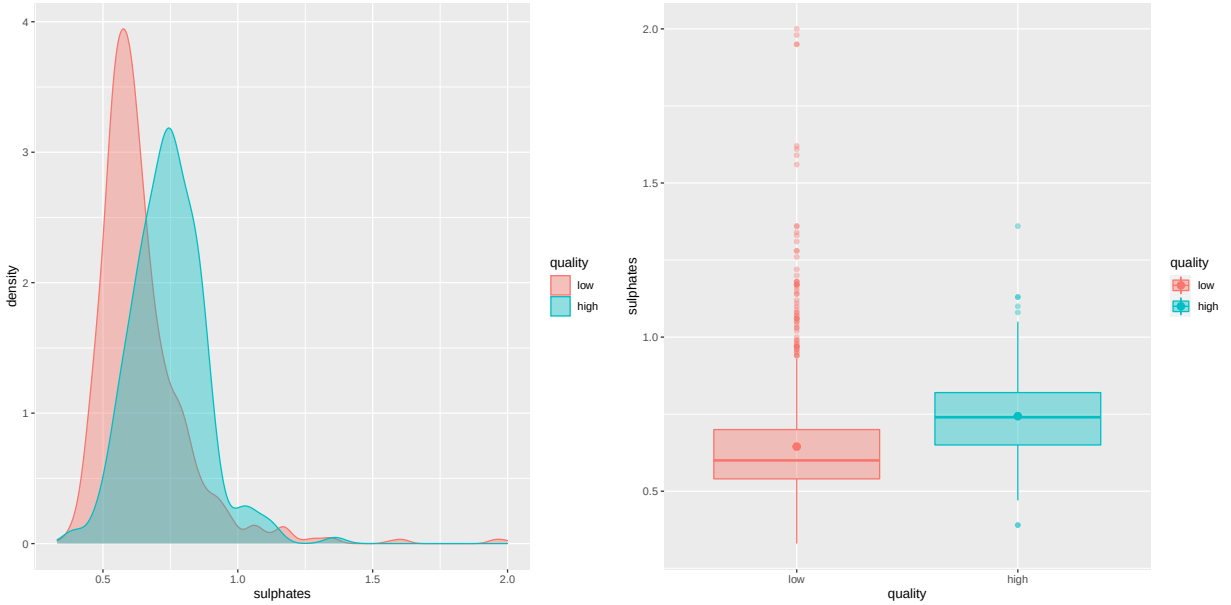

Figure 2.3: Sulphates distribution per quality class

Figure 2.7 suggests that the correlation between variables is moderately high with a maximum absolute

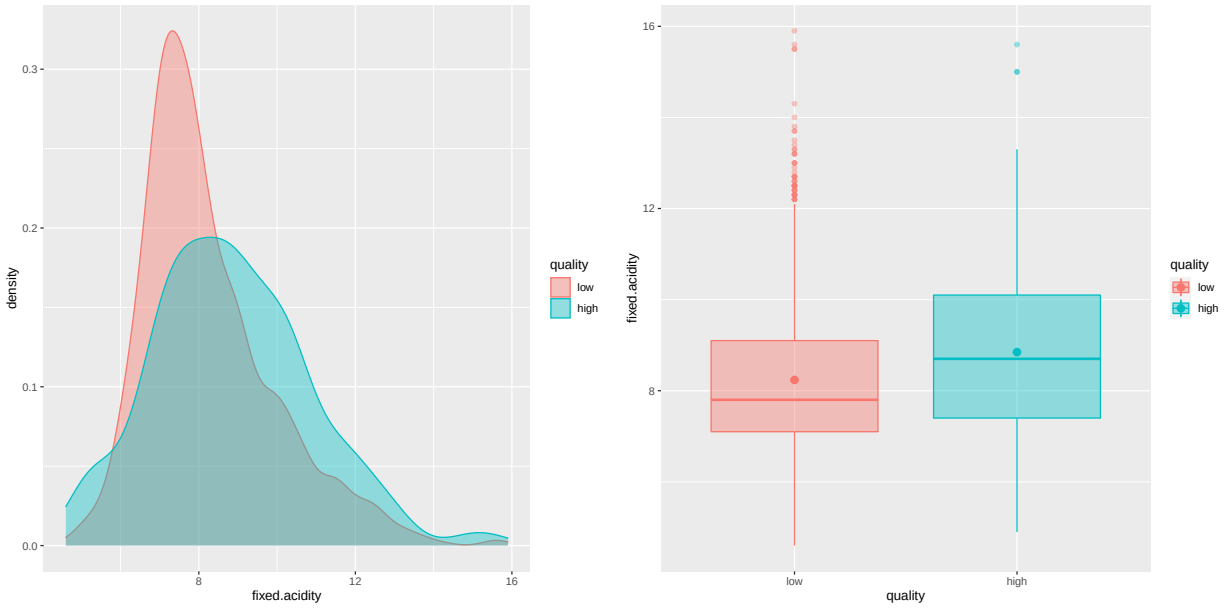

Figure 2.4: Fixed acidity distribution per quality class

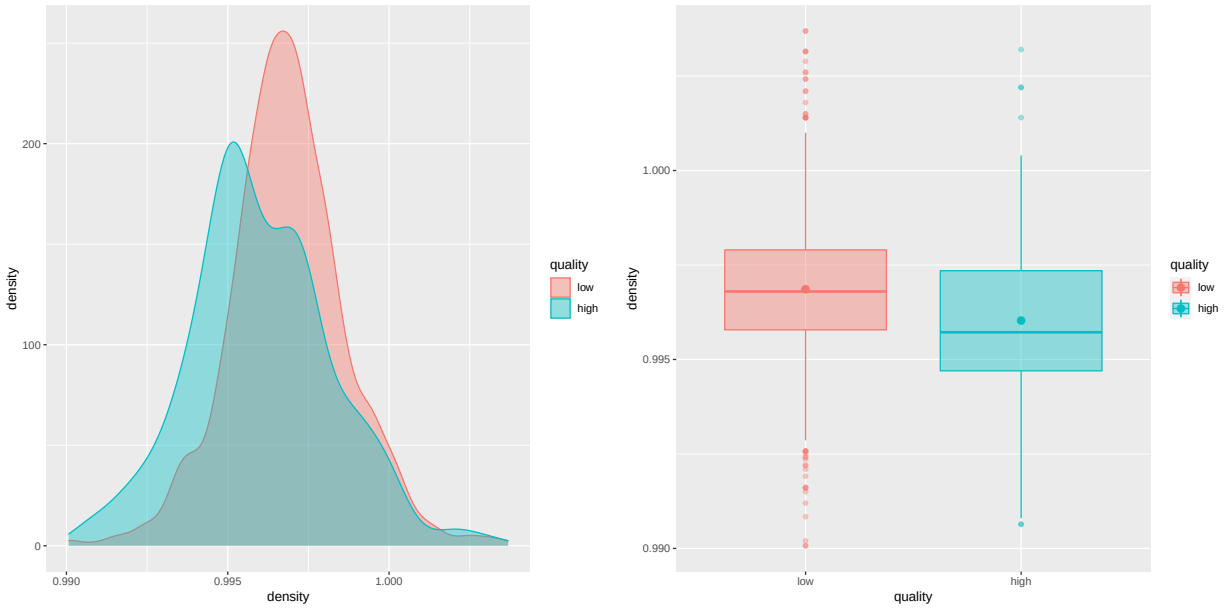

Figure 2.5: Density distribution per quality class

value of the correlation coefficient of 0.7, positive between fixed acidity and citric acid, fixed acidity and density, free sulfur dioxide and total sulfur dioxide and negative between fixed acidity and pH. Given the absolute value of the correlation coefficient, some care must be taken to preserve the assumption of no multicolinearity that some classification models make and remove variables that have a high correlation or that can be explained with linear combinations of other variables.

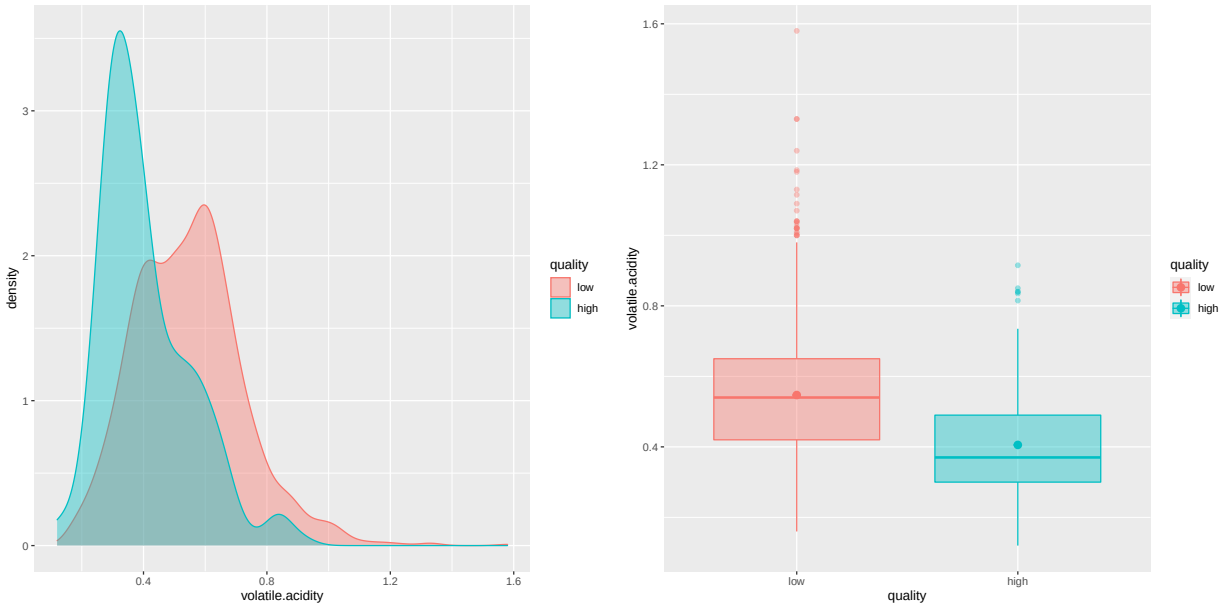

Figure 2.6: Volatile acidity distribution per quality class

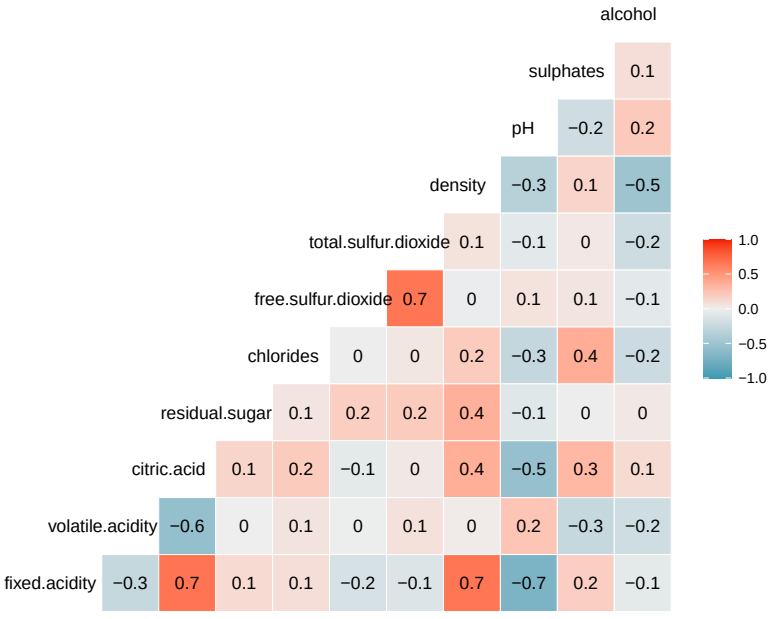

Figure 2.7: Correlation among variables

2.2 Outlier Detection

As presented in figure 2.8, both classes of wine quality display the existence of outliers that are identified through the score and orthogonal distance to the origin in the robust principal components space [2].

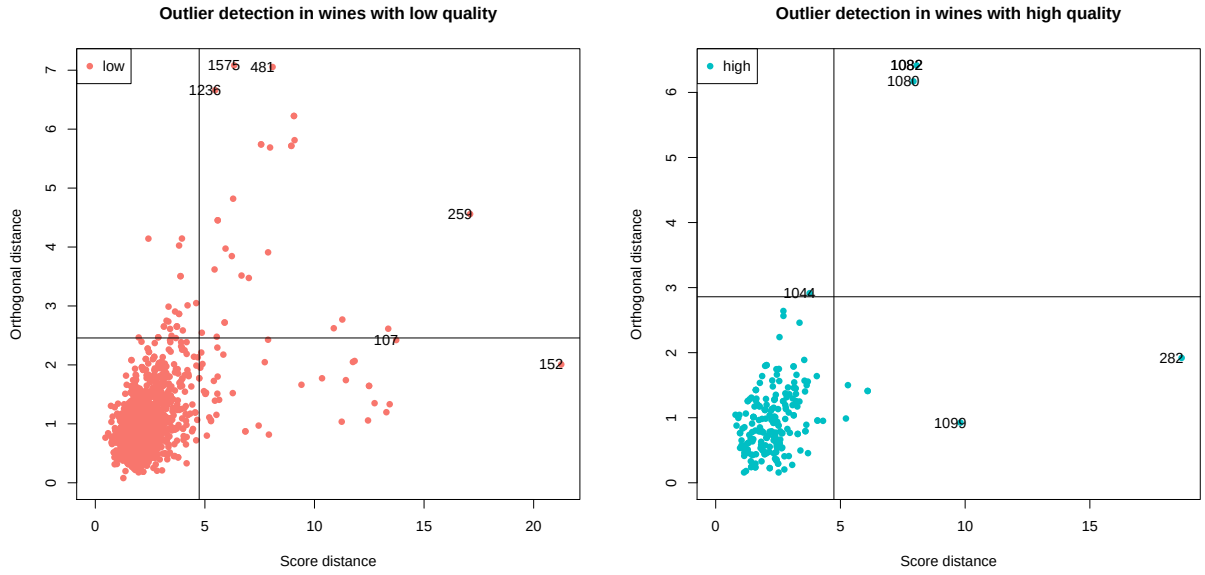

Figure 2.8: Outlier detection per quality class

Each class was subsequentially winsorised [3] in order to minimize the influence of extreme values without removing them, leading to the results in figure 2.9.

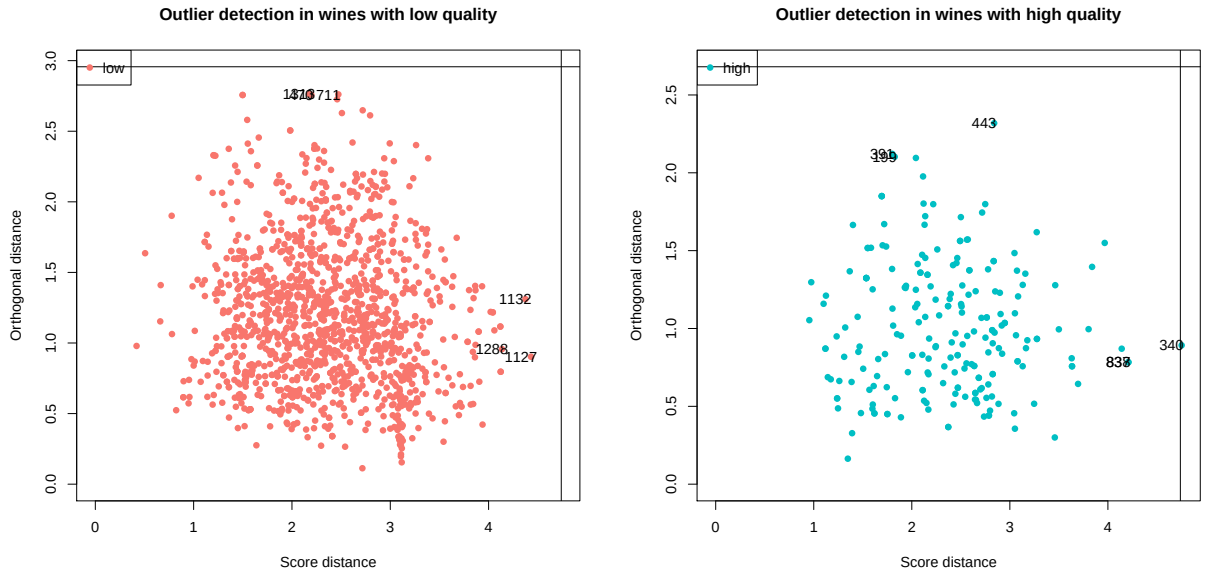

Figure 2.9: Outlier detection per quality class after winsorisation

2.3 Data Reduction

As there are significant differences in the range of values of each variable with the total sulfur dioxide having a range of 283 mg/dm^3 while density has a range of 0.01 g/dm^3, it becomes necessary to standardise the dataset so all the variables can contribute proportionally in the prediction. For this reason, the classification tasks will also use a scaled and centered dataset.

Another dataset transformation used in the classification models is the principal component extraction, with the objective of reducing the dimensionality of the data. In this case, as outliers were previously identified and treated, the classical standardised principal components were computed instead of the standardised robust principal components. The analysis presented in table 2.1 suggests that the first 6 principal components explain 85.95% of the variance in the sample and the variables that most contribute to each principal component are presented in figure 2.10.

Table 2.1: Standardised Principal Components Explained Variance

	PC1	PC2	PC3	PC4	PC5	PC6
standard.deviation	1.75	1.46	1.26	1.04	0.94	0.86
proportion.of.variance	0.28	0.19	0.14	0.10	0.08	0.07
cumulative.proportion	0.28	0.47	0.61	0.71	0.79	0.86

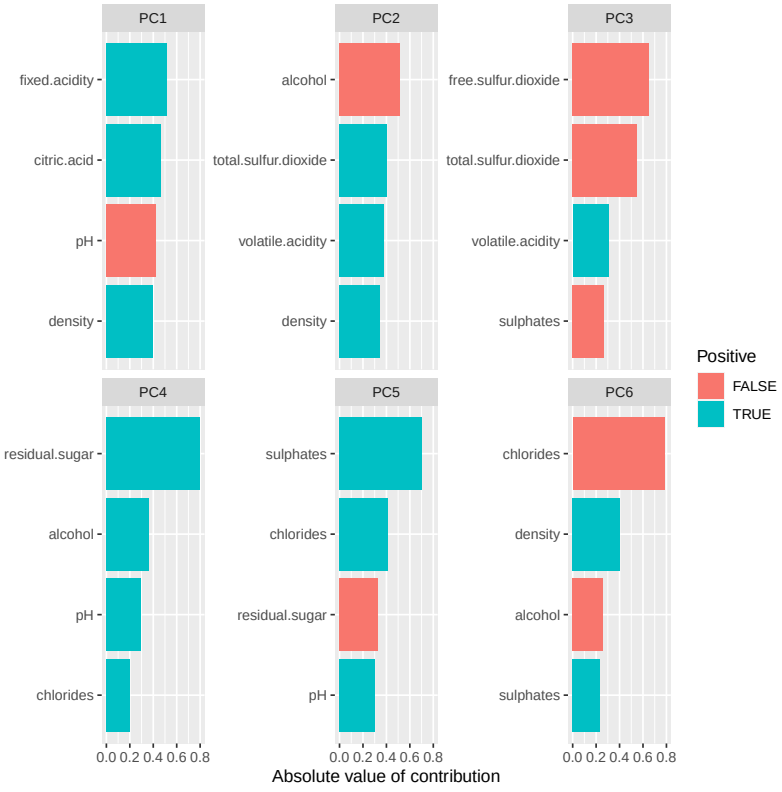

Figure 2.10: Top principal component loadings contribution

3 Classification and discrimination

The classification uses several methods, from generalised logistic regression, support vector machines to linear and flexible discrimination analysis and the models are trained on 80% of the data, with hyper parameterization in a 10 fold cross-validation repeated over 10 times.

3.0.1 Generalized Logistic Regression

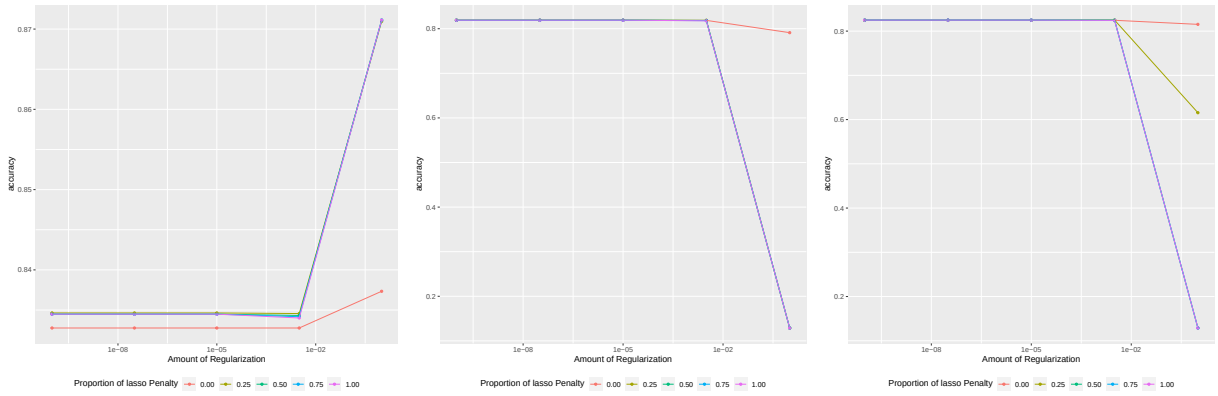

Figure 3.1: Logistic regression hyper parameter grid

An elastic-net logistic regression classifies the dataset, with an accuracy of 83.7% , using a penalty of 1 and a mixture of 50% lasso and 50% ridge regularisation. In the standardised dataset the accuracy is 81.19% using a penalty of 1e-10 and a mixture of 25% lasso and 75% ridge regularisation. Regarding the standardised principal components, the accuracy is 81.5% with a penalty of 1e-10 and a mixture of 50% lasso and 50% ridge regularisation.

3.0.2 Support Vector Machines with Polynomial Kernel

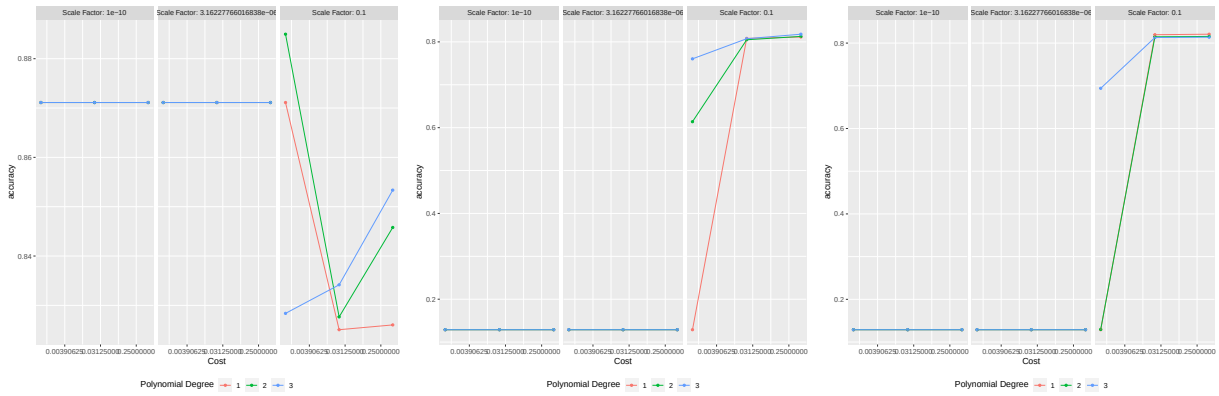

Figure 3.2: Support vector machine with polinomial kernel hyper parameter grid

A support vector machine with polynomial kernel generates an accuracy of 80.56% in the datase with a cost of 9.77e-4, a degree of 2 and a scale factor of 0.1. In the case of the standardized dataset, the accuracy is 82.13% with a cost of 0.5, a degree of 3 and with scale factor of also 0.1. Using the standardised principal

components, this method attains an accuracy of 80.56% , again with a cost of 0.5, a degree of 1 and once more a scale factor of 0.1

3.0.3 Support Vector Machines with Radial Kernel

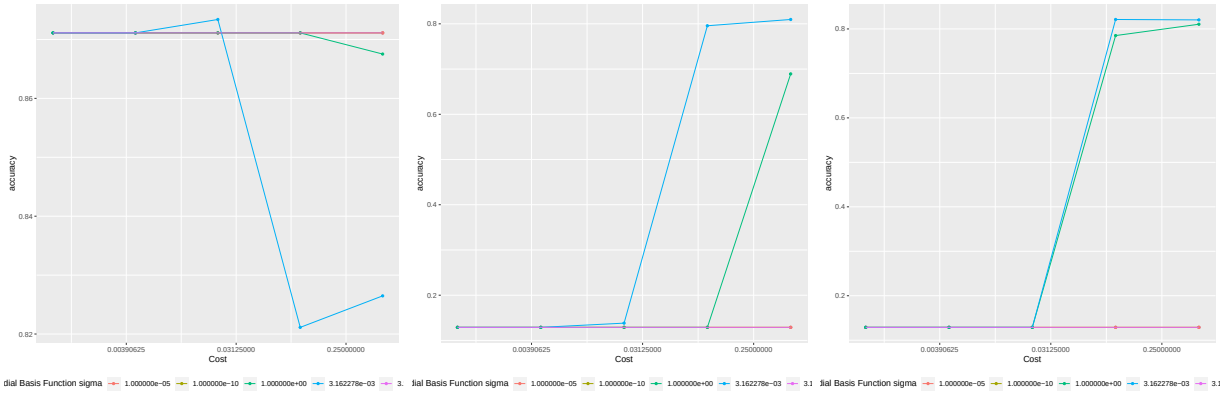

Figure 3.3: Support vector machine with radial kernel hyper parameter grid

In the dataset, the application of a support vector machine with a radial kernel generates an accuracy of 84.95% , using a cost of 0.0221 and a radial basis function sigma of 0.0031. Using the standardized dataset, the accuracy is 80.25% , with a cost of 0.5 and a radial basis function sigma of 0.00316. Applying this method to the standardised principal components allows an accuracy of 79.94% , with a cost of 0.105 and a radial basis function sigma of also 0.00316.

3.0.4 Linear Discriminant Analysis

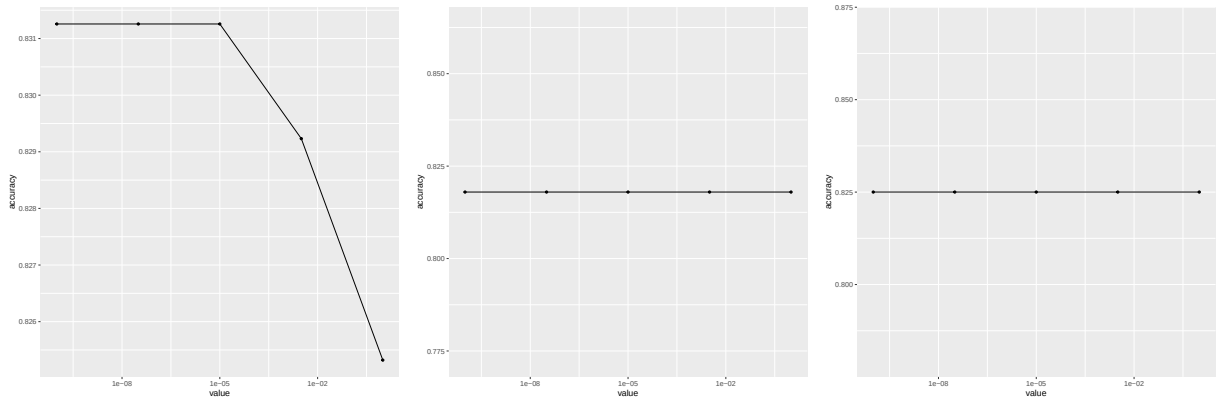

Figure 3.4: Linear discriminant analysis hyper parameter grid

Using a linear discriminant analysis in the dataset generates an accuracy of 82.45% with a penalty of 1e-10 and ridge regularisation. In the case of the standardised dataset, achieves an accuracy of 80.56% , also with a penalty of 1e-10 and ridge regularisation. The accuracy in the standardised principal components is 80.88% , once more with a penalty of 1e-10 and ridge regularisation.

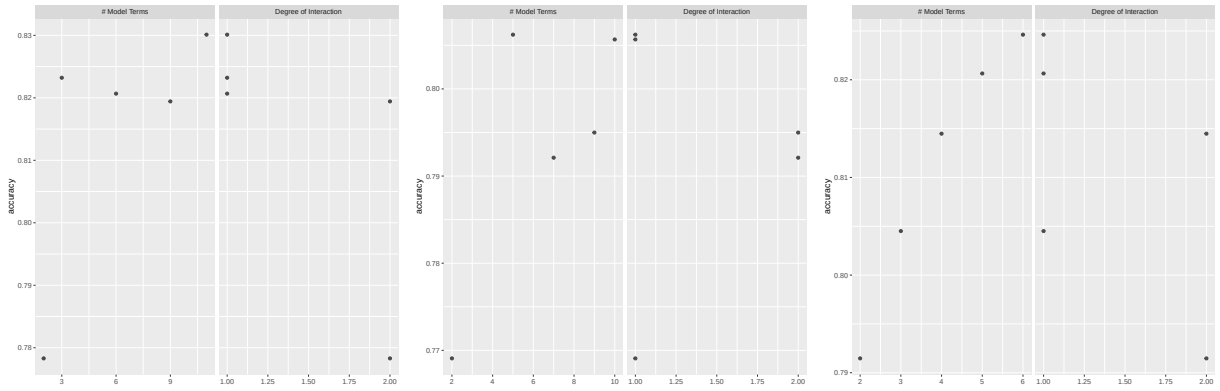

Figure 3.5: Flexible discriminant analysis hyper parameter grid

3.0.5 Flexible Discriminant Analysis

Applying a flexible discriminant analysis achieves an accuracy of 83.39% in the case of the dataset, with 11 terms retained in the model and a maximum degree of interaction of 1. In the standardised dataset, the model attains an acuracy of 79.31% , with 5 terms retained and also a maximum interaction degree of 1. Using the model in the standard principal components achieves an accuracy of 81.5% , with 6 terms retained and a maximum interaction degree of 1.

3.0.6 Comparison of the results

Table 3.1: Classification accuracy

	Logisic regression	Support Vector Machine with a polynomial kernel	Support Vector Machine with a radial kernel	Linear discriminant analysis	Flexible discriminant analysis
Filtered dataset	83.7%	80.56%	84.95%	82.45%	83.39%
Standardised filtered dataset	81.19%	82.13%	80.25%	80.56%	79.31%
Standardised filtered principal components	81.5%	80.56%	79.94%	80.88%	81.5%

The results in table 3.1 shows that the dataset leads to better testing accuracies in most cases and in particular using a support vector machines with a radial kernel, a logistic regression and also a flexible discriminant analysis. This is expected as looking into the bivariate density plots of both the dataset and the standardised principal components suggests a mixture of different distributions for the high and low quality classes so a non linear transformation will lead to a projection that will be able to better separate the classes.

4 Conclusion

The classification of wines in high and low qualities makes it easier for anyone to predict the quality of the wine and make better decisions regarding its consumption. Producers can also use these models and methodologies to direct their crops and transformation processes to wines that are able to attain a better quality class and retailers will be able to be more selective in the wines they sell, which with comparable prices lead to an improvement in sales, inventory turnover and net working capital.

Nevertheless, the analysis holds for green wines from the red variety, so the analysis should be repeated in datasets containing wines from other types of wine. Furthermore, the data was collected in a certain point in time, so new observations should be taken periodically in order to ensure that the validity of this analysis still holds.

A possible direction that can improve the results in future analysis is the use of non linear principal component, model bagging and stacking and also more recent methods of machine learning.

References

[1] P. Cortez et all. Wine quality data set. `https://archive.ics.uci.edu/ml/datasets/Wine+Quality`, 2009.

[2] Mia Hubert, Peter J Rousseeuw, and Karlien Vanden Branden. Robpca: A new approach to robust principal component analysis. *Technometrics*, 47(1):64–79, 2005.

[3] Cecil Hastings, Frederick Mosteller, John W. Tukey, and Charles P. Winsor. Low moments for small samples: A comparative study of order statistics. *Ann. Math. Statist.*, 18(3):413–426, 09 1947.

[4] Paulo Cortez, António Cerdeira, Fernando Almeida, Telmo Matos, and José Reis. Modeling wine preferences by data mining from physicochemical properties. *Decision Support Systems*, 47(4):547 – 553, 2009. Smart Business Networks: Concepts and Empirical Evidence.

List of Figures

List of Tables