Wine Quality Prediction

Claudia Valdeavella

7 February 2019

Introduction

The data for this project are the wine quality datasets[1] from the UCI Machine Learning Repository. The observations were gathered from wine samples of red and white variants of the Portuguese "Vinho Verde" wine. The inputs are objective tests and the output is based on sensory data, that is, the median of at least 3 evaluations made by wine experts. Each wine expert graded the wine quality anywhere from 0 (very bad) to 10 (excellent).

The objective of this project is to predict the quality of the wine, which is a score between 0 and 10, based on the following attributes

- fixed acidity
- volatile acidity
- citric acid
- residual sugar
- chlorides
- free sulfur dioxide
- total sulfur dioxide
- density
- pH
- sulphates
- alcohol

This is a multiclass classification problem. The author will apply Principal Component Analysis (PCA), Support Vector Machines (SVM), and Random Forests (RF) to the datasets. Separate models will be built for the red and white variants.

Analysis

We will start our analysis by examining the range of values of each of the variables. The summary for the red wine dataset is shown below.

```
##  fixed.acidity   volatile.acidity  citric.acid    residual.sugar
##  Min.   : 4.60   Min.   :0.1200   Min.   :0.000   Min.   : 0.900
##  1st Qu.: 7.10   1st Qu.:0.3900   1st Qu.:0.090   1st Qu.: 1.900
##  Median : 7.90   Median :0.5200   Median :0.260   Median : 2.200
##  Mean   : 8.32   Mean   :0.5278   Mean   :0.271   Mean   : 2.539
##  3rd Qu.: 9.20   3rd Qu.:0.6400   3rd Qu.:0.420   3rd Qu.: 2.600
##  Max.   :15.90   Max.   :1.5800   Max.   :1.000   Max.   :15.500
##    chlorides       free.sulfur.dioxide total.sulfur.dioxide
##  Min.   :0.01200   Min.   : 1.00       Min.   :  6.00
##  1st Qu.:0.07000   1st Qu.: 7.00       1st Qu.: 22.00
##  Median :0.07900   Median :14.00       Median : 38.00
##  Mean   :0.08747   Mean   :15.87       Mean   : 46.47
##  3rd Qu.:0.09000   3rd Qu.:21.00       3rd Qu.: 62.00
##  Max.   :0.61100   Max.   :72.00       Max.   :289.00
##     density             pH          sulphates         alcohol
```

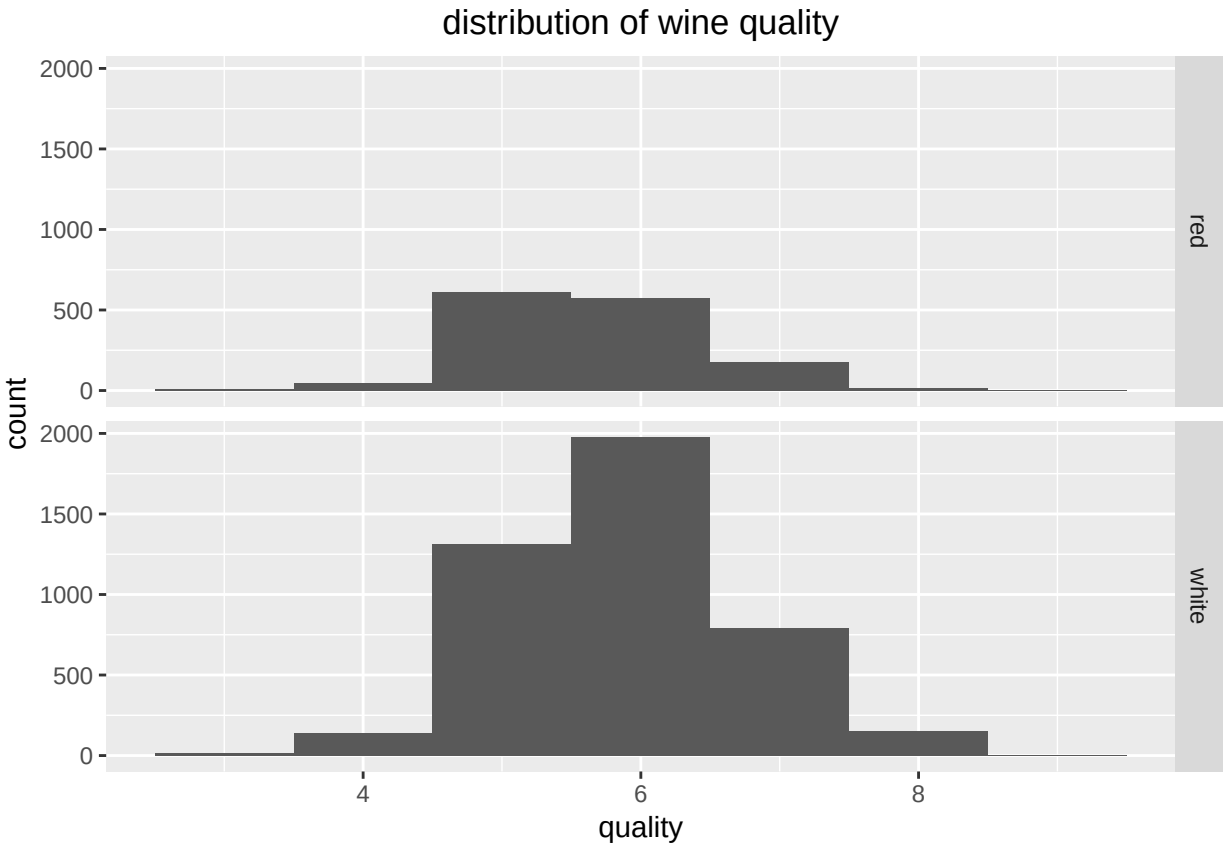
```
##  Min.   :0.9901   Min.   :2.740   Min.   :0.3300   Min.   : 8.40
##  1st Qu.:0.9956   1st Qu.:3.210   1st Qu.:0.5500   1st Qu.: 9.50
##  Median :0.9968   Median :3.310   Median :0.6200   Median :10.20
##  Mean   :0.9967   Mean   :3.311   Mean   :0.6581   Mean   :10.42
##  3rd Qu.:0.9978   3rd Qu.:3.400   3rd Qu.:0.7300   3rd Qu.:11.10
##  Max.   :1.0037   Max.   :4.010   Max.   :2.0000   Max.   :14.90
##     quality
##  Min.   :3.000
##  1st Qu.:5.000
##  Median :6.000
##  Mean   :5.636
##  3rd Qu.:6.000
##  Max.   :8.000
```

Similary, we obtain the following summary for the white wine dataset.

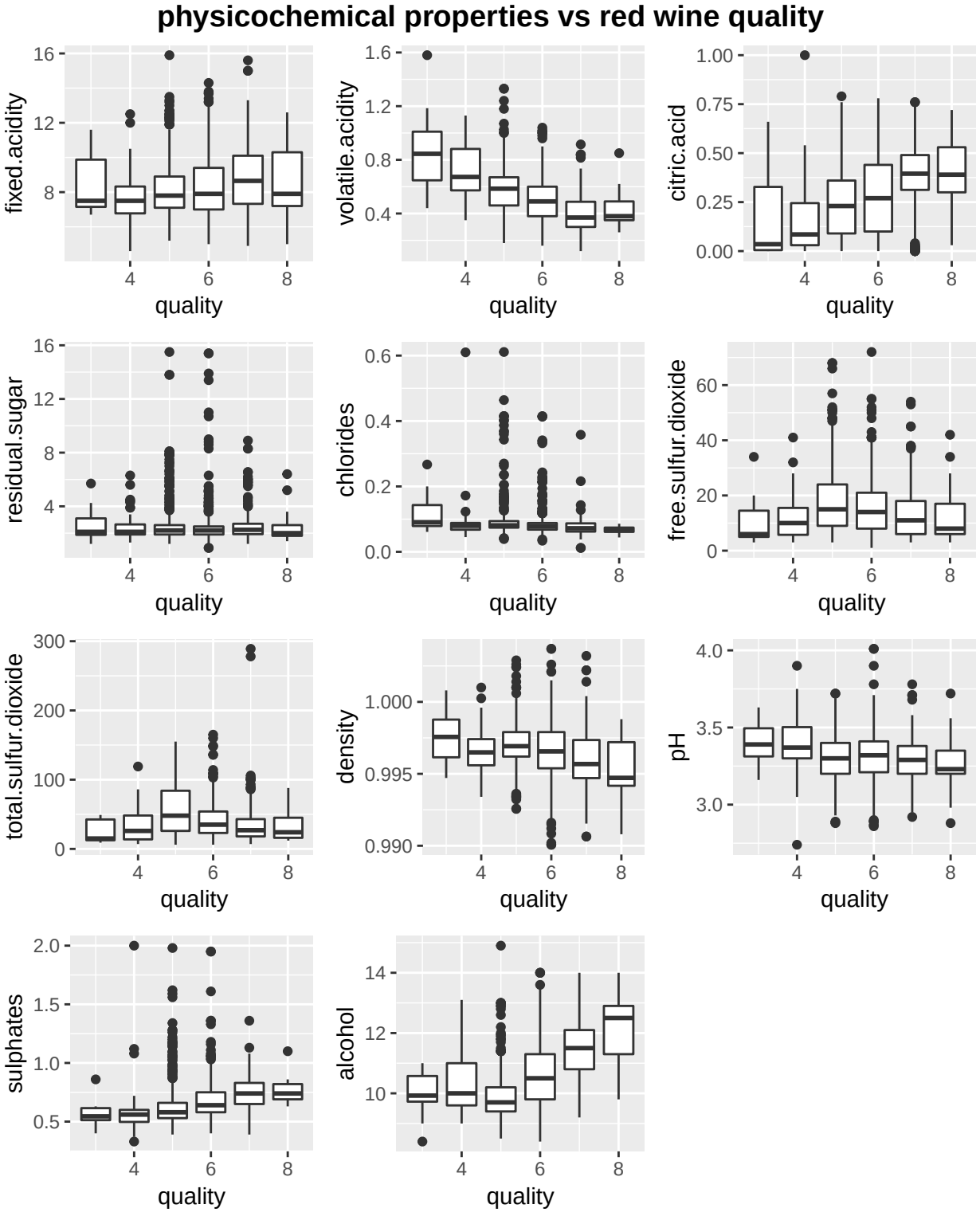
```
##  fixed.acidity    volatile.acidity  citric.acid     residual.sugar
##  Min.   : 3.800   Min.   :0.0800   Min.   :0.0000   Min.   : 0.600
##  1st Qu.: 6.300   1st Qu.:0.2100   1st Qu.:0.2700   1st Qu.: 1.700
##  Median : 6.800   Median :0.2600   Median :0.3200   Median : 5.200
##  Mean   : 6.855   Mean   :0.2782   Mean   :0.3342   Mean   : 6.391
##  3rd Qu.: 7.300   3rd Qu.:0.3200   3rd Qu.:0.3900   3rd Qu.: 9.900
##  Max.   :14.200   Max.   :1.1000   Max.   :1.6600   Max.   :65.800
##    chlorides       free.sulfur.dioxide total.sulfur.dioxide
##  Min.   :0.00900   Min.   :  2.00      Min.   :  9.0
##  1st Qu.:0.03600   1st Qu.: 23.00      1st Qu.:108.0
##  Median :0.04300   Median : 34.00      Median :134.0
##  Mean   :0.04577   Mean   : 35.31      Mean   :138.4
##  3rd Qu.:0.05000   3rd Qu.: 46.00      3rd Qu.:167.0
##  Max.   :0.34600   Max.   :289.00      Max.   :440.0
##     density             pH          sulphates         alcohol
##  Min.   :0.9871   Min.   :2.720   Min.   :0.2200   Min.   : 8.00
##  1st Qu.:0.9917   1st Qu.:3.090   1st Qu.:0.4100   1st Qu.: 9.50
##  Median :0.9937   Median :3.180   Median :0.4700   Median :10.40
##  Mean   :0.9940   Mean   :3.188   Mean   :0.4898   Mean   :10.51
##  3rd Qu.:0.9961   3rd Qu.:3.280   3rd Qu.:0.5500   3rd Qu.:11.40
##  Max.   :1.0390   Max.   :3.820   Max.   :1.0800   Max.   :14.20
##     quality
##  Min.   :3.000
##  1st Qu.:5.000
##  Median :6.000
##  Mean   :5.878
##  3rd Qu.:6.000
##  Max.   :9.000
```

From the above summaries of the data, we observe that there is a wide range in the values of the variables, the total sulfur dioxides are in the hundreds while the chlorides are in the order of 0.01, hence it is necessary to scale the values prior to creating the models. There are no missing values in the datasets.

Next we observe that there is an imbalance in the quality of the wines, that is, majority of the wine samples are average in quality. Few excellent or poor quality wines are included in the dataset.

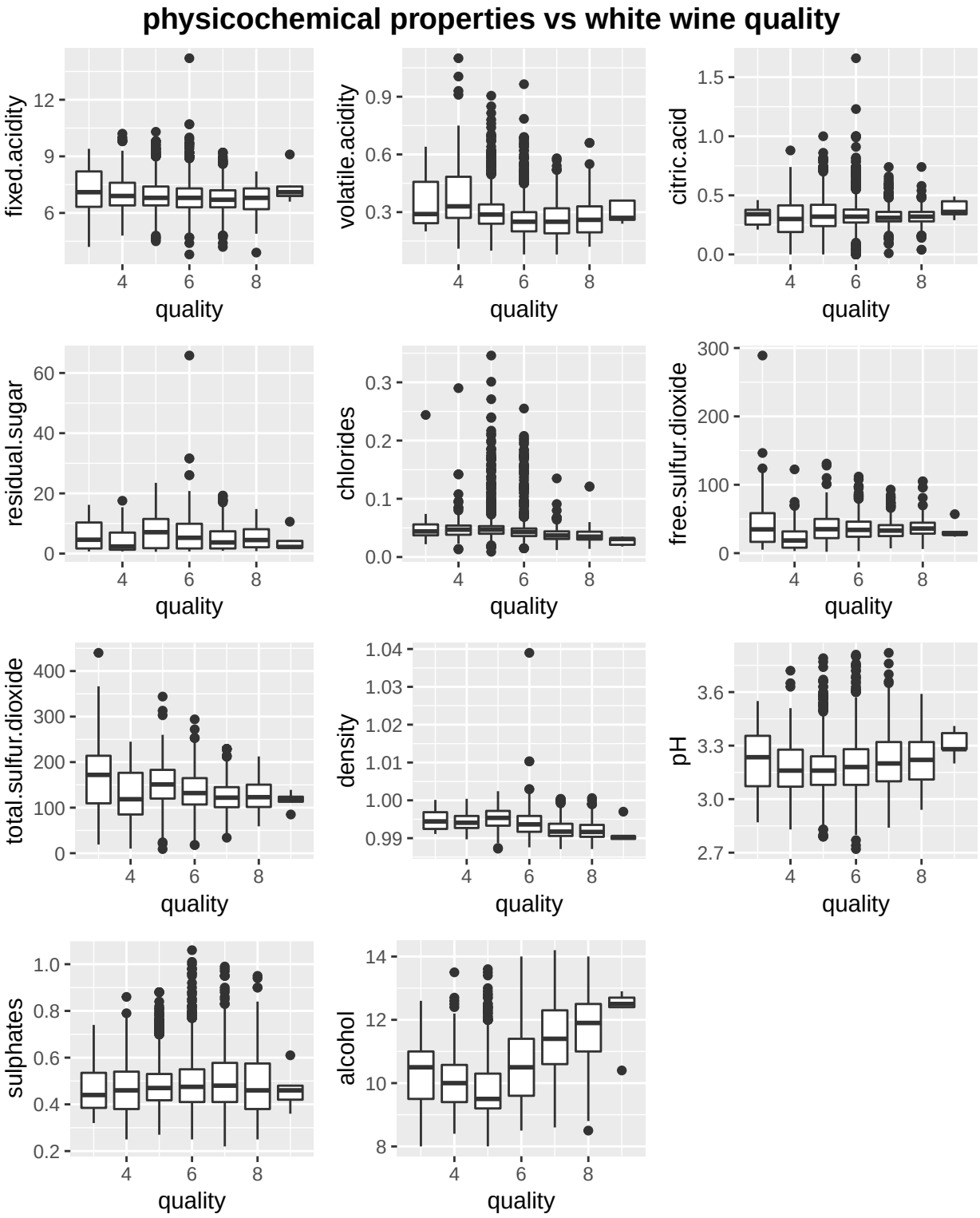


Prior to modeling, we will examine how the wine quality varies with each predictor variable.



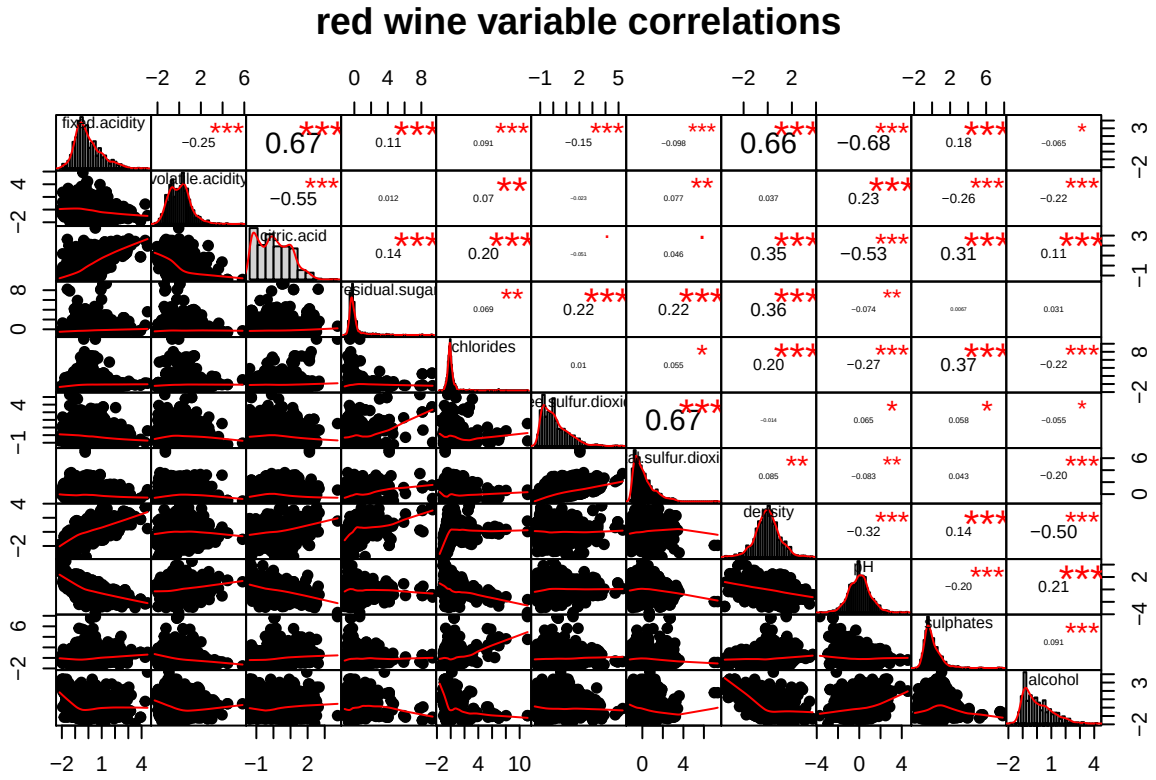
For red wines, the levels of alcohol, citric acid and sulfates are higher in excellent wine samples. In contrast, volatile acidity, density and pH levels are lower in excellent samples.

For white wines, the alcohol levels are high in excellent wine samples. Besides the trend in alcohol levels, we don't observe clear trends in the other variables.

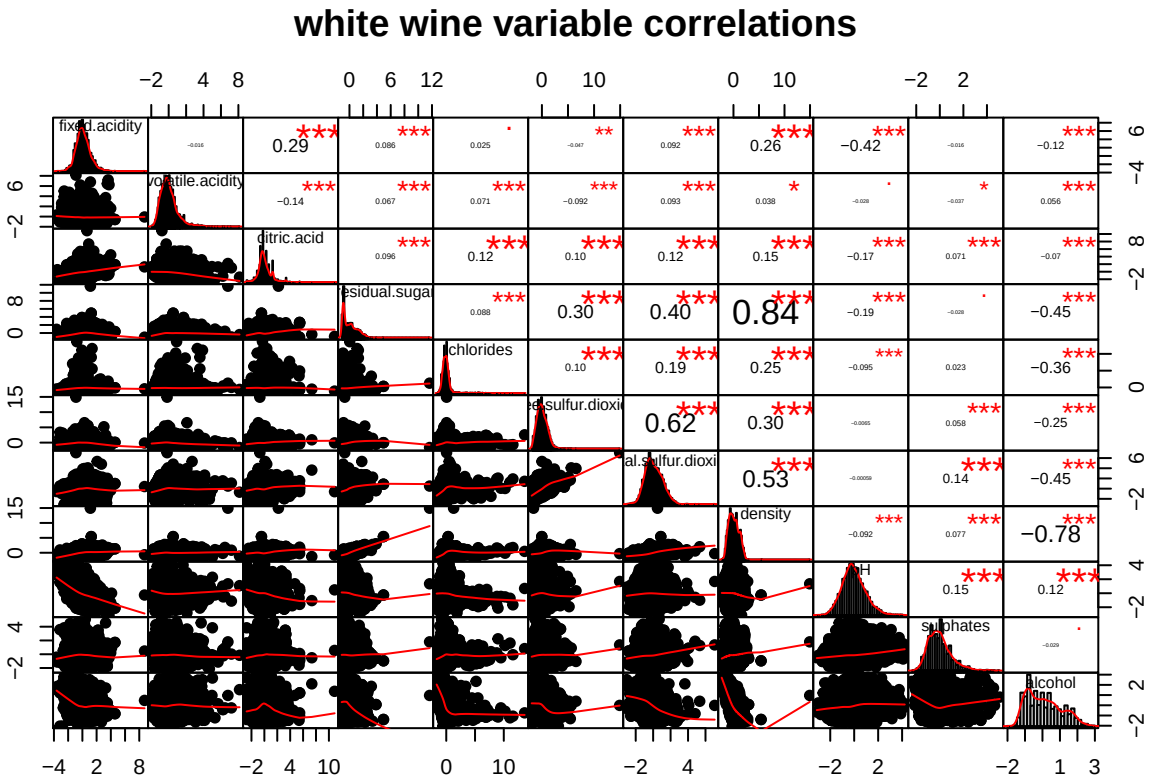


Are there correlations among the 11 predictor variables? If so, is it possible to reduce the dimensionality of the problem?

In the red wine dataset, we see from the following plot that certain pairs of variables are highly correlated, fixed acidity and citric acid, free sulfur dioxide and total sulfur dioxide, and fixed acidity and pH, to name a few.



We see high correlations between pairs of variables in the white wine dataset also. Examine the correlations between residual sugar and density, density and alcohol, and free sulur dioxide and total sulfur dioxide.

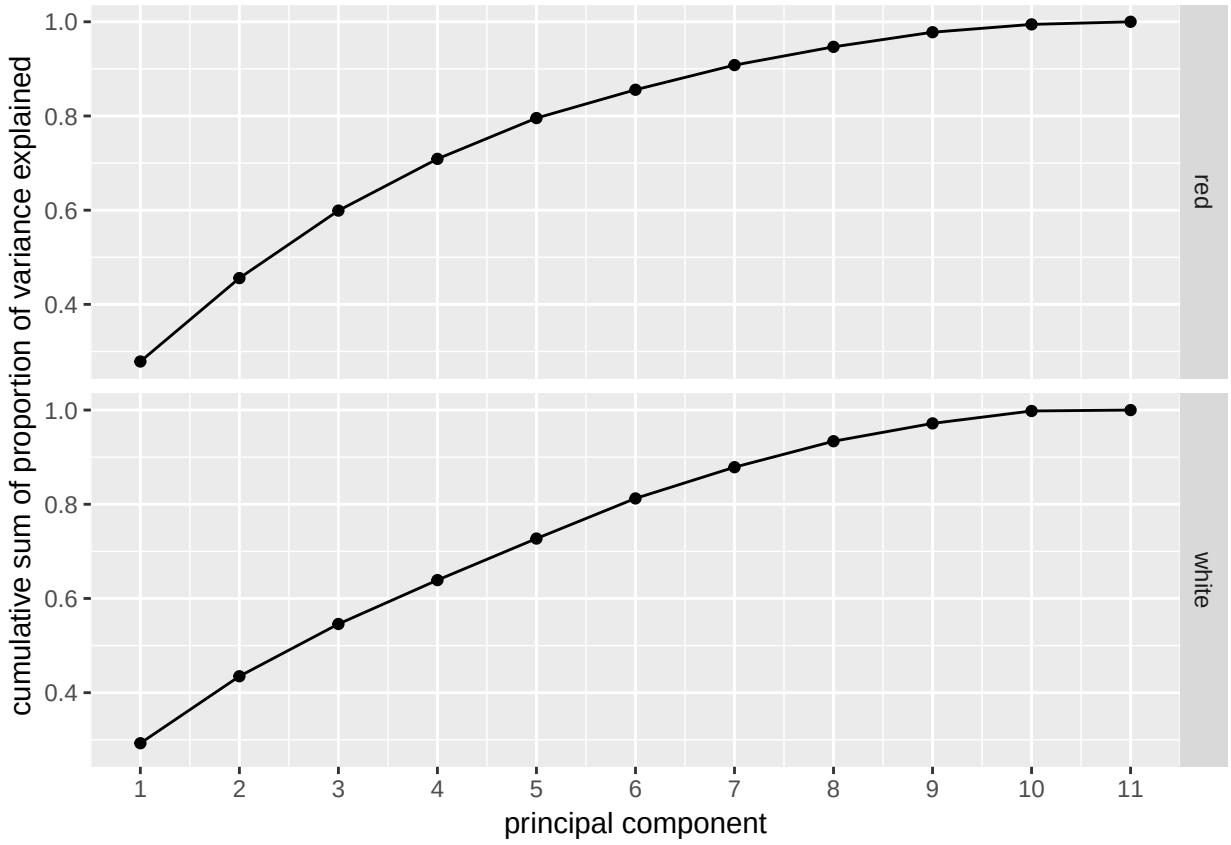


In this project, separate models were built for the red and white wines because the quality ratings are based on sensory data, that is, ratings by wine experts, and the author thinks that, in general, expectations are different for red and white wines.

Principal Component Analysis

When faced with a large set of correlated variables, PCA allows us to summarize this set with a smaller number of representative variables that explain most of the variability in the original set. Each of the principal components found by PCA is a direction in feature space which is a linear combination of the original features, 11 in the wine quality datasets. These are ordered such that the first principal component has the largest variance.

The contribution of each of the principal components to the variance in the data can be gleaned from the following graphs.



From the above analysis, we see that we need PC1 through PC8 to account up to 95% and PC1 through PC10 to account up to 99% of the variance in the data. A similar analysis of the white wine dataset reveals that PC1 through PC8 accounts for up to 93% and PC1 through PC10 accounts for up to 99.8% of the variance in the dataset. Given the observed behavior, it doesn't seem like PCA will allow us to reduce the dimensionality of the problem by much in both the red and white wine datasets.

We will use three approaches for our multiclass classification problem

- SVM
- PCA + SVM
- Random Forests

SVM classifies data by constructing hyperplanes in multidimensional space that separates observations according to their class labels. It can be applied to nonlinear data distributions with the use of kernels. These are essentially mathematical functions that map the data to higher dimensions where they will hopefully be linearly separable.

In the second approach, we use the principal components as the predictors in the SVM model in place of the original variables.

The third approach, Random Forests, involves producing multiple decision trees that are then combined to yield a single consensus prediction. The use of a large number of trees improve prediction accuracy.

Support Vector Machines

SVM analysis for the red and white wine datasets involved the following steps

1. Tune nonlinear SVM using 10-fold cross validation to find the best values for the cost and gamma parameters
2. Build the model using the training set and the best estimates for cost and gamma
3. Apply the model to the test set to predict the quality

PCA+SVM

The second approach that was pursued is a combination of PCA and SVM, that is, subsets of the principal component vectors were used as variables for the SVM model.

The rationale for doing this is two-fold. Pre-processing the data with PCA

- reduces the dimensionality of the data
- removes correlations between the predictors

both of which are expected to facilitate the downstream SVM analysis.

Models were built for the red wine data, first using PC1-PC8, then using PC1-PC10. These vectors accounted for 95% and 99%, respectively, for the variance in the data as observed previously. For the white wine dataset, recall that PC1-PC8 accounts for 95% and PC1-PC10 accounts for 99.8% for the variance in the data set.

The first step is the projection of the observations in the training and test sets onto the principal component vectors. The rest of the steps are as described in the previous section for SVM.

Random Forests

As in the previous sections of this report, the red wine dataset was analyzed followed by the white wine dataset.

The following parameters were tuned for this model

- nTree is the number of trees to grow
- predFixed is the number of randomly selected parameters for splitting
- minNode is the minimal node size

These were tuned using the train() function and the Rborist package. After tuning the parameters, the model was used to predict the quality of the wine samples in the test set. Based on the following accuracy values obtained during the training, we used 700 trees for subsequent analysis steps.

nTree	predFixed	minNode	red	white
500	2	5	0.6514	0.6387
700	2	5	0.6641	0.6389
1000	2	5	0.6630	0.6384

Results

The author used accuracy as the metric in evaluating the models. A comparison of the results from the different models is shown in the following table

method	red	white
svm 11 var	0.7081	0.6837
pca (1-8) + SVM	0.6584	0.6367
pca (1-10) + SVM	0.6708	0.6878
random forests	0.7081	0.6551

The SVM model and the Random Forests model predicted the quality of the red wine samples with equal accuracy (0.7081), but the combination of PCA and SVM gave the most accurate prediction (0.6878) for the white wine samples.

For the white wine data, combining PCA and SVM gives better results than SVM alone. This is a consequence of the modeling being done in the transformed space where the predictor varibles (the principal components) are orthogonal to each other.

Including more principal components into the subsequent SVM model increased the accuracy of the prediction. This is the expected behavior for this dataset because we knew from the PCA analysis that the PC9 and PC10 principal components still have significant contributions to the variance in the data.

According to the literature, correlations among the predictor variables can be handled by the use of custom kernels within SVM. This would render the use of principal components unnecessary.

The author did not use principal components in conjunction with Random Forests. The rationale is that Random Forests is a correlation robust algorithm, hence, pre-processing the data to remove correlations is not necessary prior to modeling.

One distinctive feature of the datasets in this project is the imbalance in the distribution of the samples into the classes. There are very few excellent wines or really poor quality wines. The lack of samples at the two extremes means that there isn't enough data from which to discern the pattern that makes for an excellent wine, for example.

The author noted that some analyses of the red wine dataset published on the web simplified the classification problem by binning the observations into 3 groups (bad, average, or good wine) instead of the 0-10 quality rating. The accuracy in this case can be as high as 0.9. However, this number can be an artifact of starting with a dataset which is disproportionately populated with average wine.

Conclusion

The author found that the best predictors of red wine quality are SVM with all 11 predictor variables and Random Forests. However, the combination of PCA with 10 component vectors and SVM gave the predictions with the highest accuracy for the white wine dataset followed by SVM alone.

High correlations were observed between pairs of variables in the dataset. PCA was done to address this, and although PCA in conjuntion with SVM gave good predictions of wine quality, the PCA approach for this particular dataset was not able to reduce the dimensionality of the problem by much.

References

1. P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis. 2009. Modeling wine preferences by data mining from physicochemical properties. In Decision Support Systems, Elsevier, 47(4):547-553. ISSN: 0167-9236.