Wine Quality

Penny Cookson

28/12/2021

Executive Summary

Introduction

This report documents an analysis of wine quality data, describes the methodology used to create a quality prediction model, and reports on the results obtained.

The data is sourced from https://archive.ics.uci.edu/ml/machine-learning-databases which provides data sets for study purposes. The data set used was created by Paulo Cortez (Univ. Minho), Antonio Cerdeira, Fernando Almeida, Telmo Matos and Jose Reis (CVRVV) @ 2009 (1).

The wine attributes are physical and chemical properties, and the quality is a rating provided by wine experts, with each wine subject to at least three expert evaluations.

The aim of the analysis is to provide a prediction for a wine's quality given its physical and chemical properties. This could be very useful to a restaurant's wine buyer that does not have access to a wine expert.

Data is available for both red and white wines. This analysis was restricted to the red wine data. The wine is a red variant of the Portuguese "Vinho Verde" wine.

Key Steps

The following steps were performed:

1. Data loading
   Data was loaded from the given source.
2. Data quality checking
   Quality checking included empty and invalid values.
3. Data exploration and visualisation
   Visualisation of data was used to evaluate the attributes which could be used to predict quality.
4. Model creation and evaluation.
   The initial model used the whole data set. Random Forest, K-Means and Support Vector Machine models were used to predict quality.

   The visual analysis indicated that outliers may affect the results and the training and prediction for some models were performed after stripping outliers.
5. Final model evaluation
   All models were evaluated against the test data set and the final model selected.

Results Summary

The data is not balanced and has few values for very low and high quality wines. There are no wines at all with a quality rating of 0-2 , or 9-10. Prediction for wines in the mid quality was accurate, however it was not possible to positively predict wines at the low quality levels of 3 and 4. The best prediction model was a

Random Forest which provided an overall accuracy of 0.7236025. Combining the Random Forest with other models did not improve the result.

There were a few instances of distant outliers which were removed before training some models. Not all the attributes were useful and those with less effect were ignored.

It is recommended that the analysis be repeated with more data.

Analysis

Data Source

The data in the form of a csv file was downloaded and the local file loaded into R.
Two minor data errors were corrected ,and the data types of the columns were changed to numbers for all the attributes and a factor for the target (quality).

Data Quality Checking/Cleaning

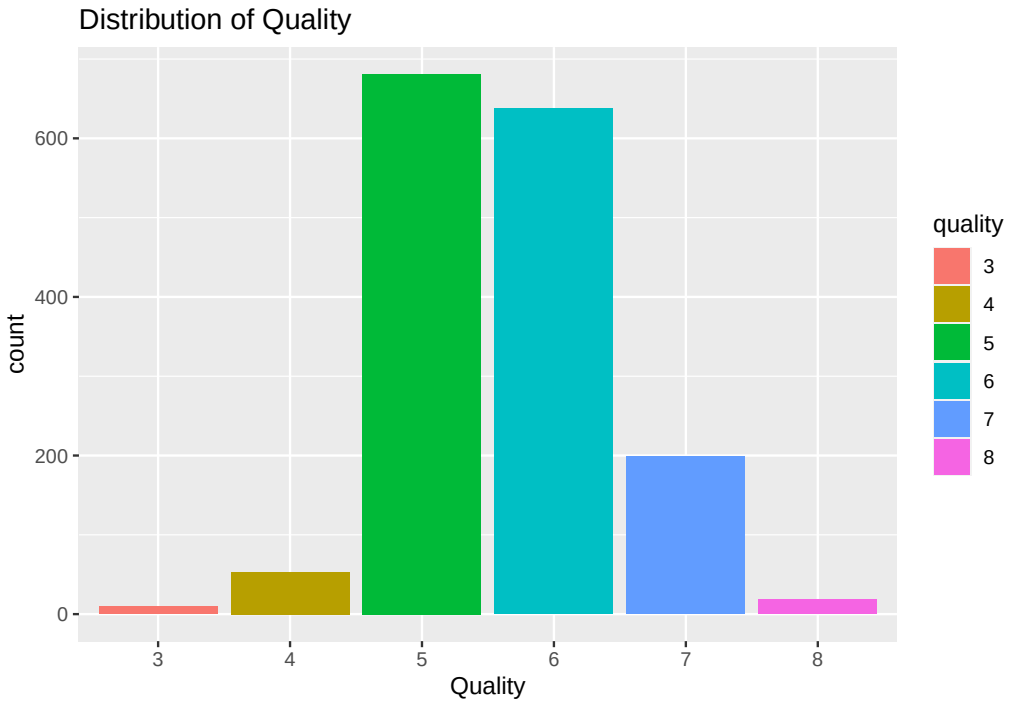
Structure The structure of the data , following these fixes is as follows:

```
## spec_tbl_df [1,599 x 12] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ fixedAcidity      : num [1:1599] 74 78 78 112 74 74 79 73 78 75 ...
##  $ volatileAcidity   : num [1:1599] 0.7 0.88 0.76 0.28 0.7 0.66 0.6 0.65 0.58 0.5 ...
##  $ citricAcid        : num [1:1599] 0 0 0.04 0.56 0 0 0.06 0 0.02 0.36 ...
##  $ residualSugar     : num [1:1599] 19 26 23 19 19 18 16 12 2 61 ...
##  $ chlorides         : num [1:1599] 0.076 0.098 0.092 0.075 0.076 0.075 0.069 0.065 0.073 0.071 ...
##  $ freeSulfurDioxide : num [1:1599] 11 25 15 17 11 13 15 15 9 17 ...
##  $ totalSulfurDioxide: num [1:1599] 34 67 54 60 34 40 59 21 18 102 ...
##  $ density           : num [1:1599] 0.998 0.997 0.997 0.998 0.998 ...
##  $ pH                : num [1:1599] 351 32 326 316 351 351 33 339 336 335 ...
##  $ sulphates         : num [1:1599] 0.56 0.68 0.65 0.58 0.56 0.56 0.46 0.47 0.57 0.8 ...
##  $ alcohol           : num [1:1599] 94 98 98 98 94 94 94 10 95 105 ...
##  $ quality           : Factor w/ 6 levels "3","4","5","6",..: 3 3 3 4 3 3 3 5 5 3 ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   `fixed acidity` = col_number(),
##   ..   `volatile acidity` = col_character(),
##   ..   `citric acid` = col_character(),
##   ..   `residual sugar` = col_number(),
##   ..   chlorides = col_character(),
##   ..   `free sulfur dioxide` = col_number(),
##   ..   `total sulfur dioxide` = col_double(),
##   ..   density = col_character(),
##   ..   pH = col_number(),
##   ..   sulphates = col_character(),
##   ..   alcohol = col_number(),
##   ..   quality = col_double()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

Data Quality There are 0 rows with invalid or empty values.

Counts There are 1599 rows of data in the red wine data set.

In order to assess how balanced the data is, plot the number of wines for each level of quality:

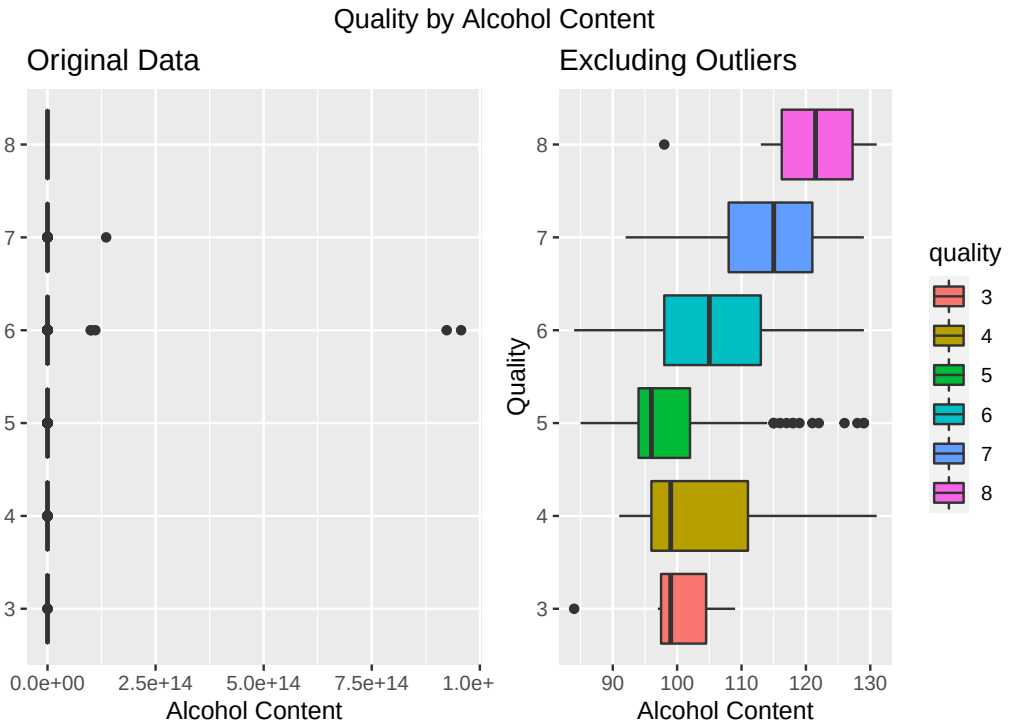


The plot above shows that the data is not balanced. It contains few records where the wine is rated very highly or very low. This is to be expected since most wine would be likely to fall within an acceptable quality range. There are no wines at all at the 0,1,2 ,9,10 quality levels.

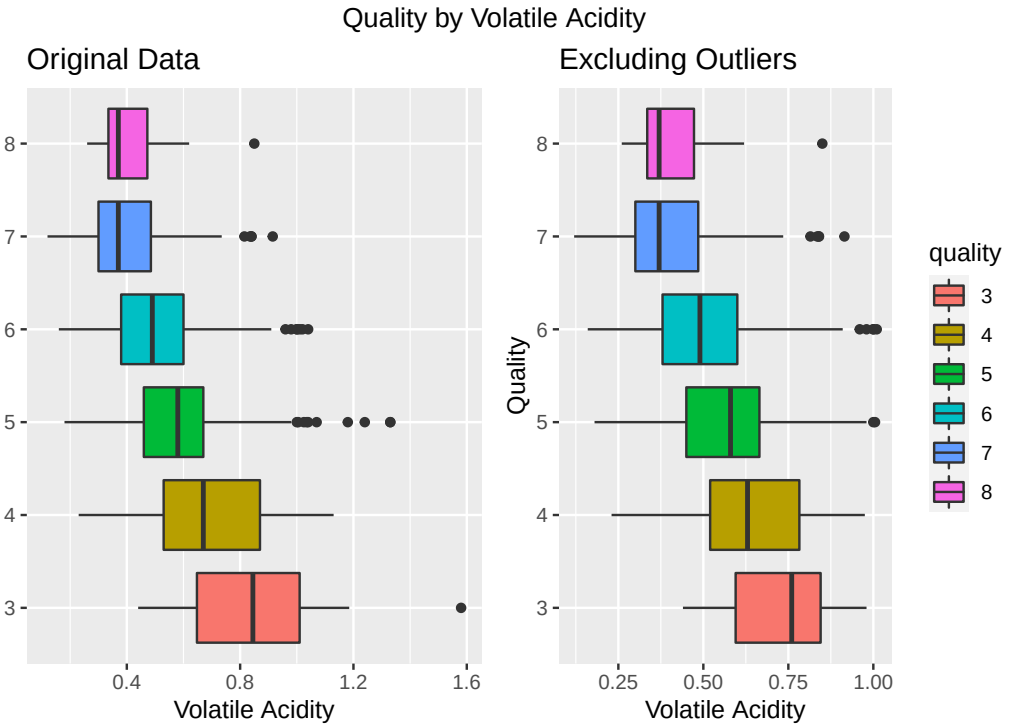
Exploration and Visualisation

The data was investigated graphically. Each attribute was plotted against quality. Since many of the attributes have outliers, each attribute is plotted with and without the outliers.

(The plots have been included in order of attribute relevance determined during the Random Forest model training.)

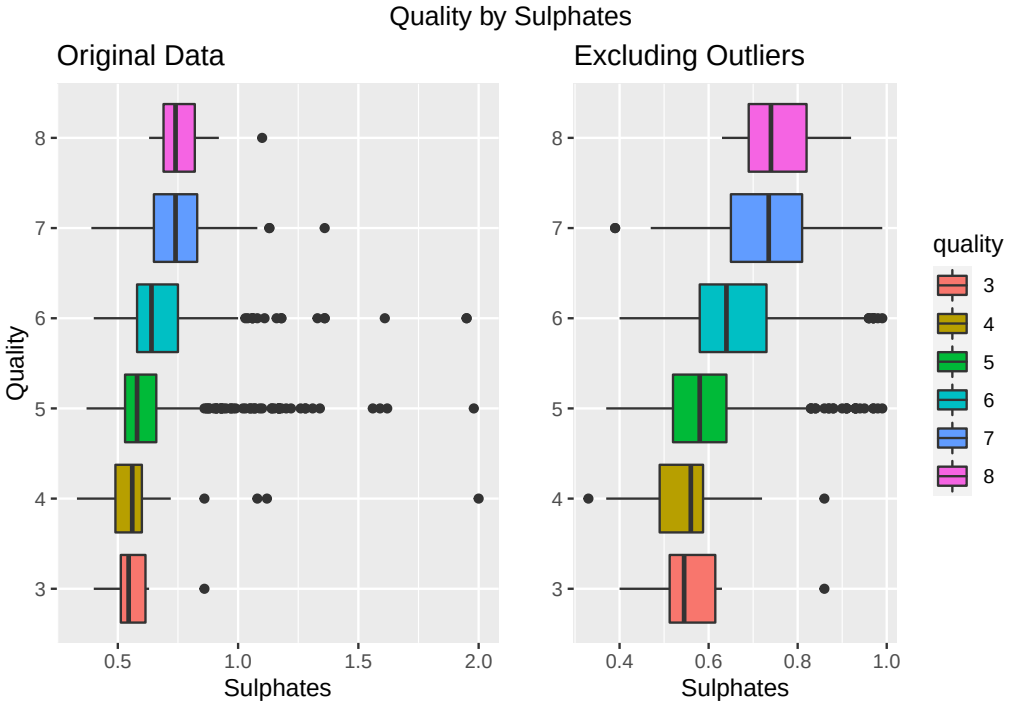


The alcohol data is skewed by the existence of a few wines with very high alcohol content. Since these wines do not have a very high or low quality rating, the outliers are unlikely to contribute to the model and will be removed before model training.
Wines with a high alcohol content are generally rated higher than those with a low alcohol content. The alcohol content is expected to be a good indicator of high quality wines.

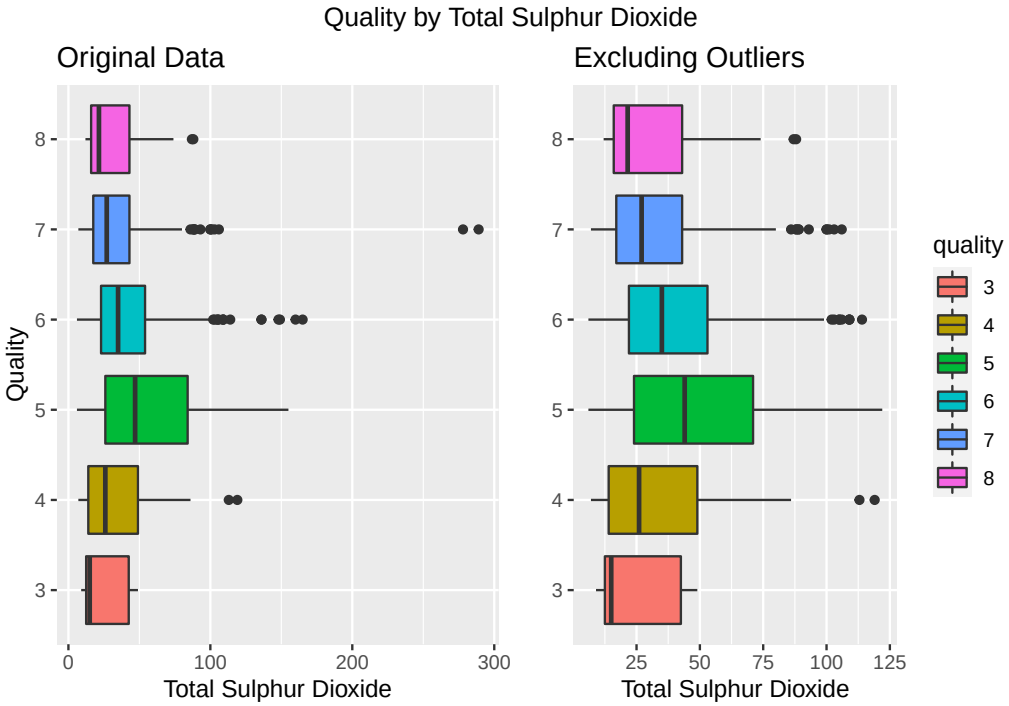


There are only a few outliers for Volatile Acidity, and their range is not as extreme as the alcohol content. The quality of wine decreases as the Volatile Acidity increases, and it is likely that high Volatile Acidity

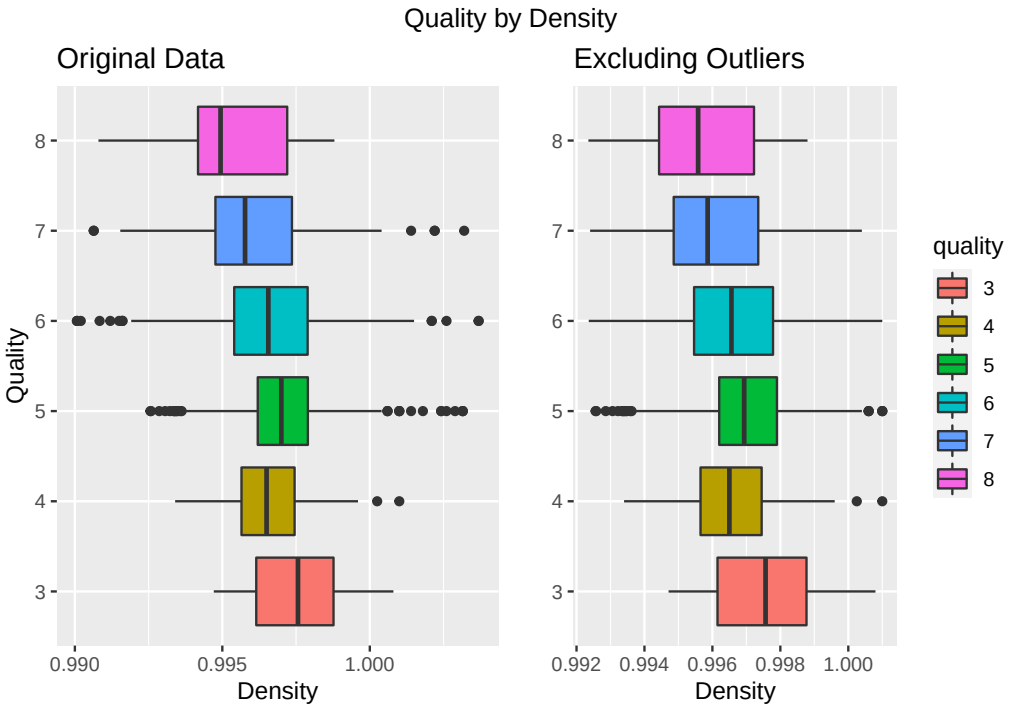
would be a good indicator of a low quality wine.



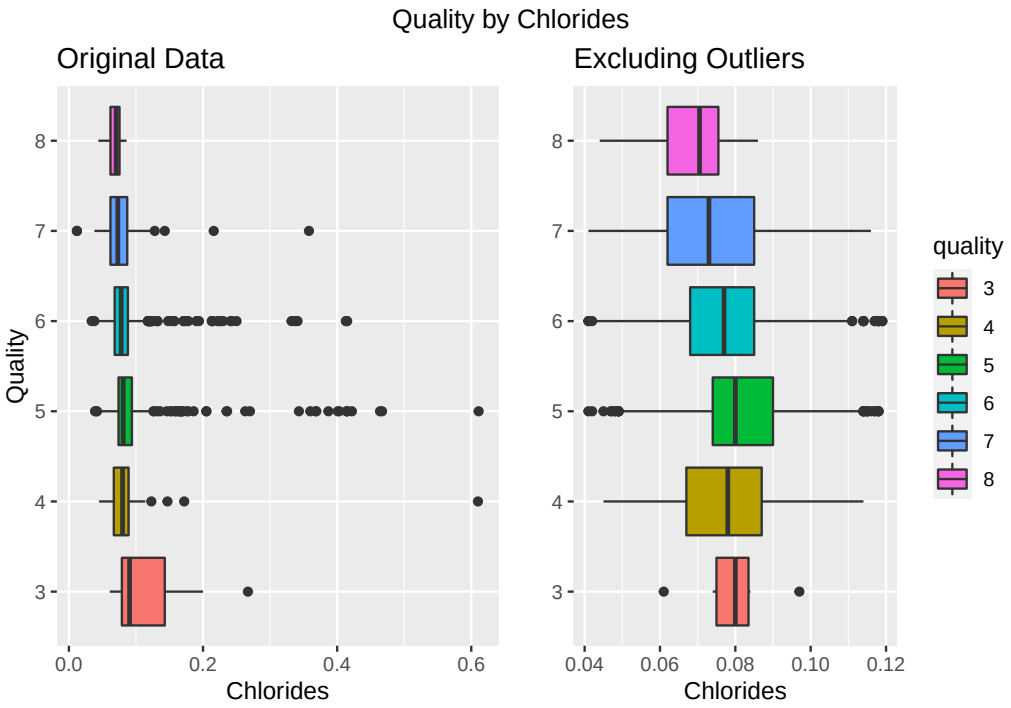
There are only a few outliers for Sulphates, and their range is again not as extreme as the alcohol content. Wines with a high Sulphates content are generally rated higher than those with a low Sulphates content. The Sulphates content is expected to be a good indicator of higher quality wines.



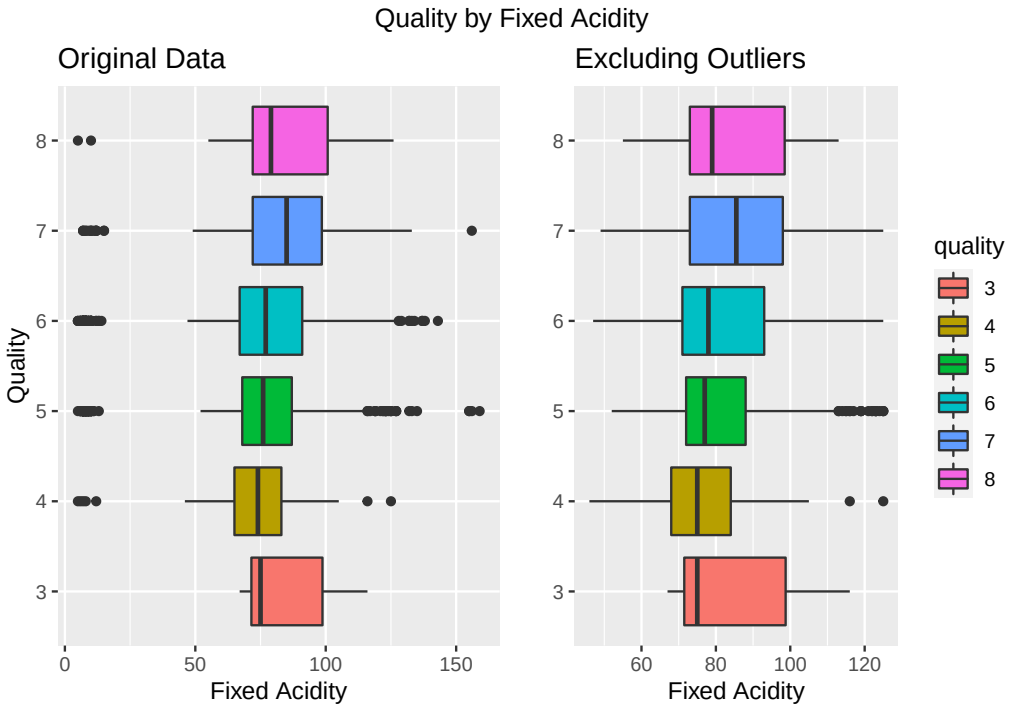
There are two outliers for Total Sulphur Dioxide, both of which have a much higher value than the mean. The quality of wines is generally mid range for wines with a higher Total Sulphur Dioxide content. Wines with a low Total Sulphur Dioxide content are either highly rated or in the low classifications.



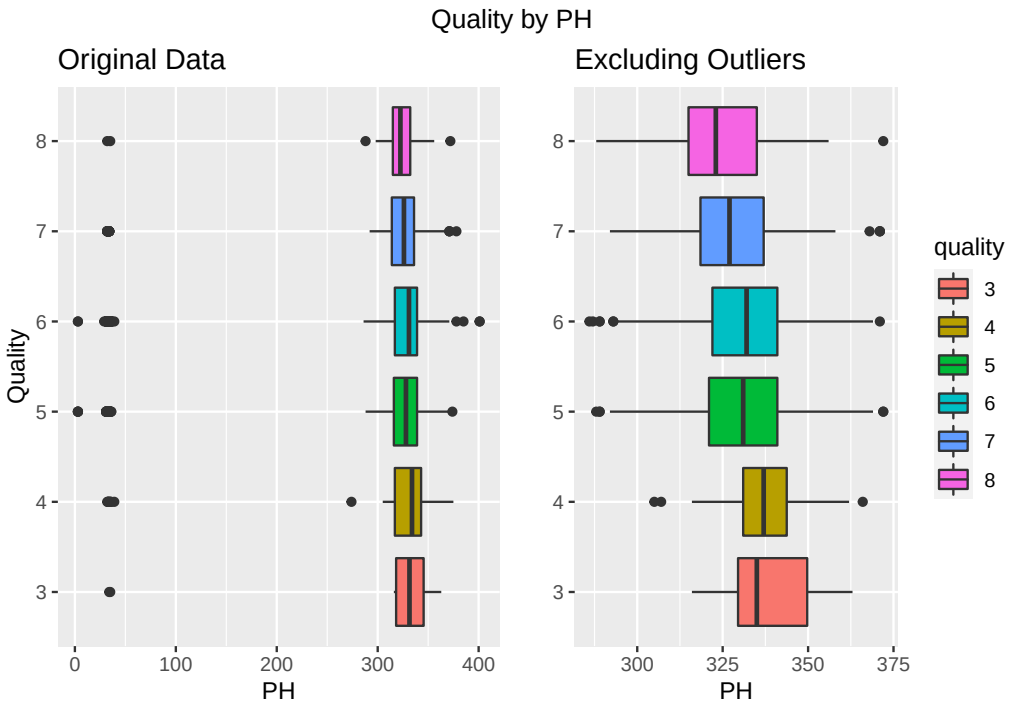
There are a few outliers for density, the values are not as far from the mean as in the alcohol data or Total Sulphur Dioxide data. Wines with a low density are generally rated higher than those with a hugh density content.



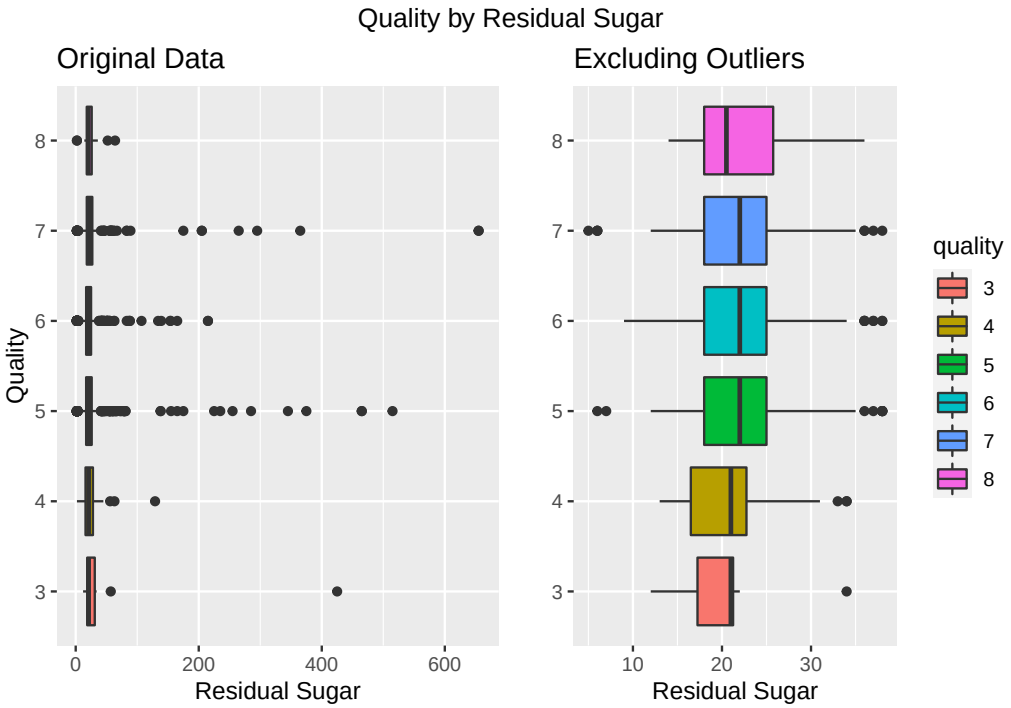
The Chlorides data is skewed by the existence of a few wines with very high Chlorides content. Since these wines do not have a very high or low quality rating, the outliers are unlikely to contribute to the model and will be removed before model training.
There is a slight decrease in quality for wines with a higher chlorides content.



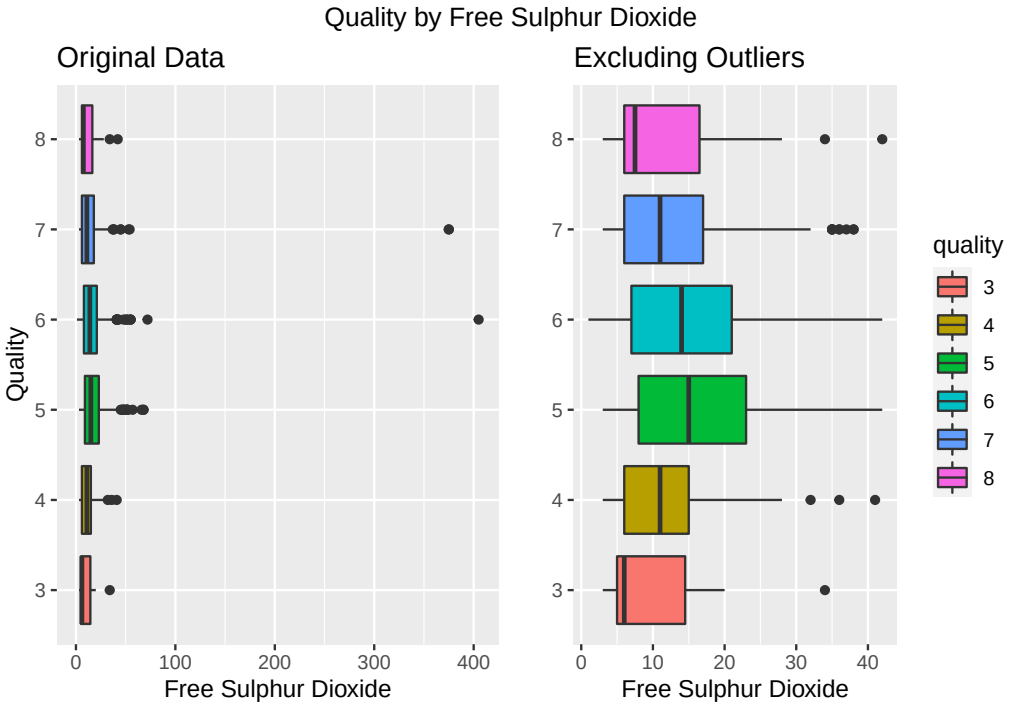
The Fixed Acidity data is skewed by the existence of a few wines with very low Fixed Acidity. These outliers occur for all quality levels, are unlikely to contribute to the model and will be removed before model training. There is a slight decrease in quality for wines with a lower Fixed Acidity content.



The PH data is skewed by the existence of a few wines with very low PH content. These outliers occur for all quality levels, are unlikely to contribute to the model and will be removed before model training.
There is a slight decrease in quality for wines with a higher PH.



The Residual Sugar data is skewed by the existence of a few wines with high Residual Sugar content. These outliers occur for all quality levels. The Residual Sugar does not appear to provide a good indicator of quality.



The Free Sulphur Dioxide data is skewed by the existence of a few wines with high Free Sulphuur Dioxide content. The Free Sulphur Dioxide does not appear to provide a good indicator of quality. Higher levels are associated with mid range wines.

Model Creation and Evaluation

Train and Test data sets

The data provided was divided into a train data set (80% of the orginal data set), and a test data set (20% of the original data set). The test data set was used only to provide an evaluation of the models created.

There are 1277 rows in the train set.

There are 322 rows in the test set.

List of the number of rows of each quality in the train set

```
## # A tibble: 6 x 2
##   quality number
##   <fct>    <int>
## 1 3            8
## 2 4           42
## 3 5          544
## 4 6          510
## 5 7          159
## 6 8           14
```

```
## # A tibble: 6 x 2
##   quality number
##   <fct>    <int>
## 1 3            2
## 2 4           11
## 3 5          137
## 4 6          128
## 5 7           40
## 6 8            4
```

Both data sets include wines with quality in each level between 3 and 8.

Model 1 - Random Forest

A Random Forest was used as the first model. The initial model used all attributes in the data set. The caret package was allowed to provide the tuning parameters nodesize and mtry. The resulting best tuning parameters were as follows:

```
##   mtry
## 1    2
```

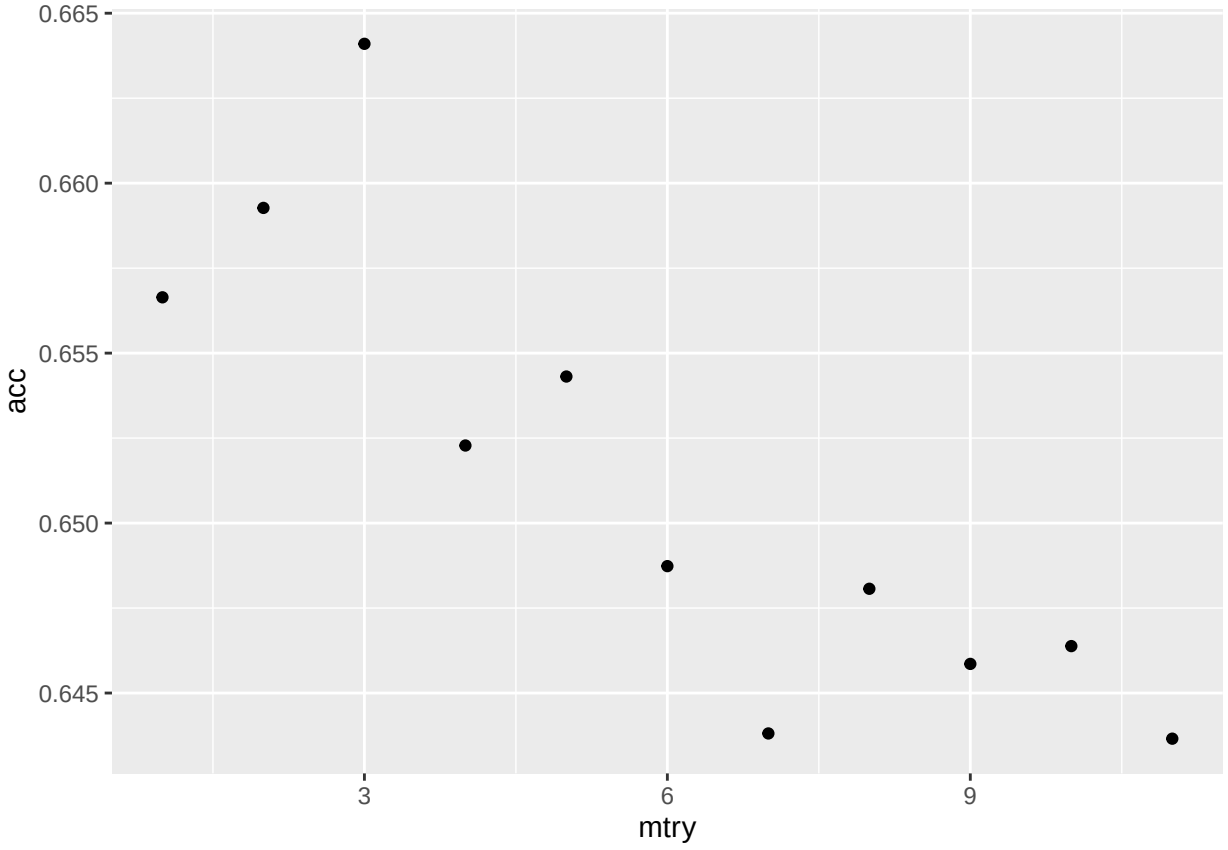
```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   3   4   5   6   7   8
##          3   8   0   0   0   0   0
##          4   0  42   0   0   0   0
##          5   0   0 544   0   0   0
##          6   0   0   0 510   0   0
##          7   0   0   0   0 159   0
##          8   0   0   0   0   0  14
##
## Overall Statistics
##
##                Accuracy : 1
##                  95% CI : (0.9971, 1)
##     No Information Rate : 0.426
##     P-Value [Acc > NIR] : < 2.2e-16
```

```
##
##                   Kappa : 1
##
##  Mcnemar's Test P-Value : NA
##
## Statistics by Class:
##
##                      Class: 3 Class: 4 Class: 5 Class: 6 Class: 7 Class: 8
## Sensitivity          1.000000  1.00000    1.000   1.0000   1.0000  1.00000
## Specificity          1.000000  1.00000    1.000   1.0000   1.0000  1.00000
## Pos Pred Value       1.000000  1.00000    1.000   1.0000   1.0000  1.00000
## Neg Pred Value       1.000000  1.00000    1.000   1.0000   1.0000  1.00000
## Prevalence           0.006265  0.03289    0.426   0.3994   0.1245  0.01096
## Detection Rate       0.006265  0.03289    0.426   0.3994   0.1245  0.01096
## Detection Prevalence 0.006265  0.03289    0.426   0.3994   0.1245  0.01096
## Balanced Accuracy    1.000000  1.00000    1.000   1.0000   1.0000  1.00000

## Accuracy
##        1
```

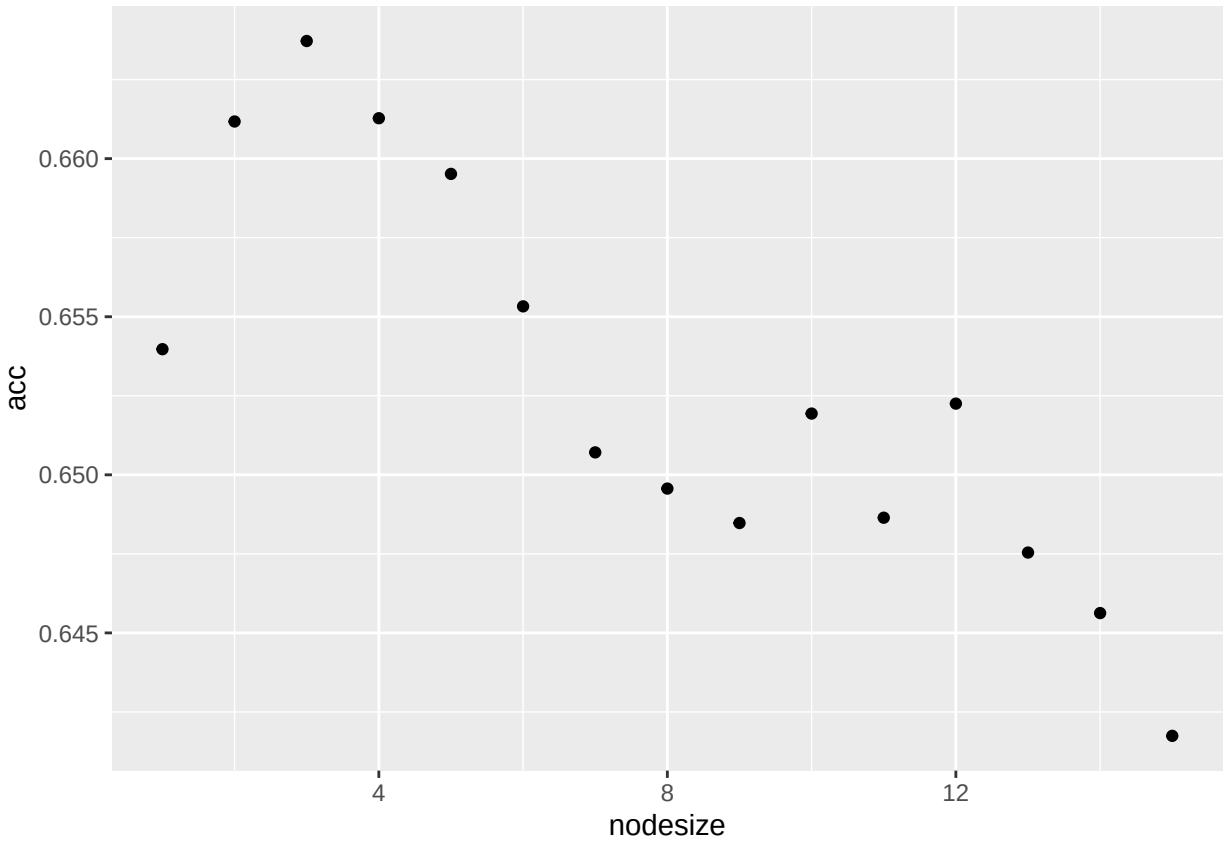
It is assumed that the relatively small number of rows in the data set allow the random forest model to handle all possible combinations and hence the accuracy is 1.

The tuning parameters were then determined manually. First the nodesize was set to 1 and the mtry parameter tuned.



The optimum value of mtry is 3.

Then the nodesize parameter is tuned.



The optimum value of mtry is 3.

These tuning parameters are then used for the random forest model (rather than the self tuning parameters). As expected the result is the same, however these parameters will be used when running the model on the test data set.

```
## Accuracy
##        1
```

The variable importance is determined from the model. The following lists variables in order of importance:

```
## rf variable importance
##
##                     Overall
## alcohol            100.000
## sulphates           63.531
## volatileAcidity     53.435
## totalSulfurDioxide  49.886
## density             32.681
## chlorides           21.995
## fixedAcidity        15.138
## pH                  13.056
## citricAcid          12.466
## residualSugar        7.605
## freeSulfurDioxide    0.000
```

Alocohol, Sulphates and Volatile Acidity provide the best indicators of quality. Residual Sugar and Free Sulphur Dioxide provide little value.

Model 2 - k-nearest neighbours

Before training the knn model the outliers identified in the data visualisation section were removed (Alcohol, Sulphur Dioxide, Chlorides, Fixed Acidity, Ph). Residual Sugar and Free Sulphur Dioxide were ignored as they appear to have minimal value.)

761 rows remain in the train set.

The k-nearest neighbours model was trained on the adjusted train set, (without outliers) and the results on the train set predicted.

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   3   4   5   6   7   8
##          3   0   0   0   0   0   0
##          4   0   1   2   0   0   0
##          5   5  32 438 195  32   2
##          6   3   8  98 277  69   7
##          7   0   1   6  38  58   5
##          8   0   0   0   0   0   0
##
## Overall Statistics
##
##                Accuracy : 0.6061
##                  95% CI : (0.5787, 0.633)
##     No Information Rate : 0.426
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.3543
##
##  Mcnemar's Test P-Value : NA
##
## Statistics by Class:
##
##                      Class: 3  Class: 4 Class: 5 Class: 6 Class: 7 Class: 8
## Sensitivity          0.000000 0.0238095   0.8051   0.5431  0.36478  0.00000
## Specificity          1.000000 0.9983806   0.6371   0.7588  0.95528  1.00000
## Pos Pred Value            NaN 0.3333333   0.6222   0.5996  0.53704      NaN
## Neg Pred Value       0.993735 0.9678179   0.8150   0.7141  0.91360  0.98904
## Prevalence           0.006265 0.0328896   0.4260   0.3994  0.12451  0.01096
## Detection Rate       0.000000 0.0007831   0.3430   0.2169  0.04542  0.00000
## Detection Prevalence 0.000000 0.0023493   0.5513   0.3618  0.08457  0.00000
## Balanced Accuracy    0.500000 0.5110950   0.7211   0.6510  0.66003  0.50000
```

```
##  Accuracy
## 0.6061081
```

The high prevalence of classes 5 and 6 impact the results. The model has not been able to predict levels 3 or level 8 quality. Balanced accuracy is $> .65$ for quality classes 5,6 and 7, but there is insufficient data for the other levels.

Model 3 - Support Vector Model

The SVM model is trained on the adjusted train set, (without outliers) and the results on the train set predicted.

```
## Confusion Matrix and Statistics
```

```
##
##           Reference
## Prediction   3   4   5   6   7   8
##          3   0   0   0   0   0   0
##          4   0   0   0   0   0   0
##          5   8  37 472 257  46   4
##          6   0   5  71 240  62   6
##          7   0   0   1  13  51   4
##          8   0   0   0   0   0   0
##
## Overall Statistics
##
##                Accuracy : 0.5975
##                  95% CI : (0.57, 0.6245)
##     No Information Rate : 0.426
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.3272
##
##  Mcnemar's Test P-Value : NA
##
## Statistics by Class:
##
##                      Class: 3 Class: 4 Class: 5 Class: 6 Class: 7 Class: 8
## Sensitivity          0.000000  0.00000   0.8676   0.4706  0.32075  0.00000
## Specificity          1.000000  1.00000   0.5198   0.8123  0.98390  1.00000
## Pos Pred Value            NaN      NaN   0.5728   0.6250  0.73913      NaN
## Neg Pred Value       0.993735  0.96711   0.8411   0.6976  0.91060  0.98904
## Prevalence           0.006265  0.03289   0.4260   0.3994  0.12451  0.01096
## Detection Rate       0.000000  0.00000   0.3696   0.1879  0.03994  0.00000
## Detection Prevalence 0.000000  0.00000   0.6453   0.3007  0.05403  0.00000
## Balanced Accuracy    0.500000  0.50000   0.6937   0.6414  0.65233  0.50000
```

```
##  Accuracy
## 0.5974941
```

The high prevalence of classes 5 and 6 impact the results. The model has not been able to predict levels 3,4 or 8 quality. Balanced accuracy is $> .64$ for quality classes 5,6 and 7, but there is insufficient data for the other levels.

Results

Model 1 - Random Forest Results (Test Data Set)

The model was evaluated against the test data.

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   3   4   5   6   7   8
##          3   0   0   0   0   0   0
##          4   0   0   0   0   0   0
##          5   1   8 107  23   1   0
##          6   1   3  30 102  16   3
##          7   0   0   0   3  23   0
##          8   0   0   0   0   0   1
```

```
##
## Overall Statistics
##
##                Accuracy : 0.7236
##                  95% CI : (0.6713, 0.7717)
##     No Information Rate : 0.4255
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.5495
##
##  Mcnemar's Test P-Value : NA
##
## Statistics by Class:
##
##                      Class: 3 Class: 4 Class: 5 Class: 6 Class: 7 Class: 8
## Sensitivity          0.000000  0.00000   0.7810   0.7969  0.57500 0.250000
## Specificity          1.000000  1.00000   0.8216   0.7268  0.98936 1.000000
## Pos Pred Value            NaN      NaN   0.7643   0.6581  0.88462 1.000000
## Neg Pred Value       0.993789  0.96584   0.8352   0.8443  0.94257 0.990654
## Prevalence           0.006211  0.03416   0.4255   0.3975  0.12422 0.012422
## Detection Rate       0.000000  0.00000   0.3323   0.3168  0.07143 0.003106
## Detection Prevalence 0.000000  0.00000   0.4348   0.4814  0.08075 0.003106
## Balanced Accuracy    0.500000  0.50000   0.8013   0.7618  0.78218 0.625000
```

```
##  Accuracy
## 0.7236025
```

The overall accuracy was 0.7236025.

The high prevalence of classes 5 and 6 impact the results. The model has not been able to predict levels 3 or 4. It has predicted level 8, but only with 25% sensitivity. Balanced accuracy is $> .76$ for quality classes 5,6 and 7, but there is insufficient data for the other levels.

This model would be expected to differentiate between wines of mid quality, assigning them a quality of between 5 and 7. This covers most of the wines in this data set.

Model 2 - k-nearest Neighbours Results (Test Data Set)

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   3   4   5   6   7   8
##          3   0   0   0   0   0   0
##          4   0   0   0   0   0   0
##          5   1   7 104  49   6   0
##          6   1   4  31  71  20   2
##          7   0   0   2   8  14   2
##          8   0   0   0   0   0   0
##
## Overall Statistics
##
##                Accuracy : 0.587
##                  95% CI : (0.531, 0.6413)
##     No Information Rate : 0.4255
##     P-Value [Acc > NIR] : 4.132e-09
##
```

```
##                   Kappa : 0.3229
##
##  Mcnemar's Test P-Value : NA
##
## Statistics by Class:
##
##                      Class: 3 Class: 4 Class: 5 Class: 6 Class: 7 Class: 8
## Sensitivity          0.000000  0.00000   0.7591   0.5547  0.35000  0.00000
## Specificity          1.000000  1.00000   0.6595   0.7010  0.95745  1.00000
## Pos Pred Value            NaN      NaN   0.6228   0.5504  0.53846      NaN
## Neg Pred Value       0.993789  0.96584   0.7871   0.7047  0.91216  0.98758
## Prevalence           0.006211  0.03416   0.4255   0.3975  0.12422  0.01242
## Detection Rate       0.000000  0.00000   0.3230   0.2205  0.04348  0.00000
## Detection Prevalence 0.000000  0.00000   0.5186   0.4006  0.08075  0.00000
## Balanced Accuracy    0.500000  0.50000   0.7093   0.6279  0.65372  0.50000
```

```
##  Accuracy
## 0.5869565
```

The overall accuracy was 0.5869565.

The high prevalence of classes 5 and 6 impact the results. The model has not been able to predict levels 3,4 or 8. Balanced accuracy is $> .62$ for quality classes 5,6 and 7, but there is insufficient data for the other levels.

Combining the Random Forest with this model did not improve results and this model has been discarded for this data set.

Model 3 - Support Vector Model Results (Test Data Set)

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   3   4   5   6   7   8
##          3   0   0   0   0   0   0
##          4   0   0   0   0   0   0
##          5   2   9 115  57  13   2
##          6   0   2  22  68  18   1
##          7   0   0   0   3   9   1
##          8   0   0   0   0   0   0
##
## Overall Statistics
##
##                Accuracy : 0.5963
##                  95% CI : (0.5404, 0.6503)
##     No Information Rate : 0.4255
##     P-Value [Acc > NIR] : 5.332e-10
##
##                   Kappa : 0.323
##
##  Mcnemar's Test P-Value : NA
##
## Statistics by Class:
##
##                      Class: 3 Class: 4 Class: 5 Class: 6 Class: 7 Class: 8
## Sensitivity          0.000000  0.00000   0.8394   0.5312  0.22500  0.00000
## Specificity          1.000000  1.00000   0.5514   0.7784  0.98582  1.00000
```

```
## Pos Pred Value            NaN      NaN   0.5808   0.6126  0.69231      NaN
## Neg Pred Value       0.993789  0.96584   0.8226   0.7156  0.89968  0.98758
## Prevalence           0.006211  0.03416   0.4255   0.3975  0.12422  0.01242
## Detection Rate       0.000000  0.00000   0.3571   0.2112  0.02795  0.00000
## Detection Prevalence 0.000000  0.00000   0.6149   0.3447  0.04037  0.00000
## Balanced Accuracy    0.500000  0.50000   0.6954   0.6548  0.60541  0.50000
```

```
##  Accuracy
## 0.5962733
```

The overall accuracy was 0.5962733.

The high prevalence of classes 5 and 6 impact the results. The model has not been able to predict levels 3,4 or 8 quality. Balanced accuracy is $> .60$ for quality classes 5,6 and 7, but there is insufficient data for the other levels.

Combining the Random Forest with this model did not improve results and this model has been discarded for this data set.

Conclusion

Findings

The model provides a good prediction for the quality of mid range wines. The main indicators were high alcohol content and low volatile acidity.

The Random Forest model was the chosen model since combining it with the other models did not improve the results.

This model would be expected to differentiate between wines of mid quality, assigning them a quality of between 5 and 7. This covers most of the wines in this data set. A more balanced data set would be required to create a model which could handle very high and low quality wines.

Limitations

The prevalence of mid range wines makes it difficult to obtain valid results for low and high quality wines. It would also be expected that rating a wine as very high quality is a more subjective decision, based on personal preference and may be hard to predict.

Recommended Future Work

Future studies including more data would be valuable. In addition it would be interesting to extend the investigation to wines of various types, and assess whether the model remains valid. Some outlier identification techniques could be investigated to identify the very low quality wines.

It is likely that some of the attributes are related, and a further analysis of the relationship between attributes, and into the relative contribution of the attributes could result in an effective model with fewer measurements of the wine required.

Technical Details

The code takes around 10 minutes to run. It was executed on a MacBook Pro with 32G memory.

References

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