Data Science Challenge

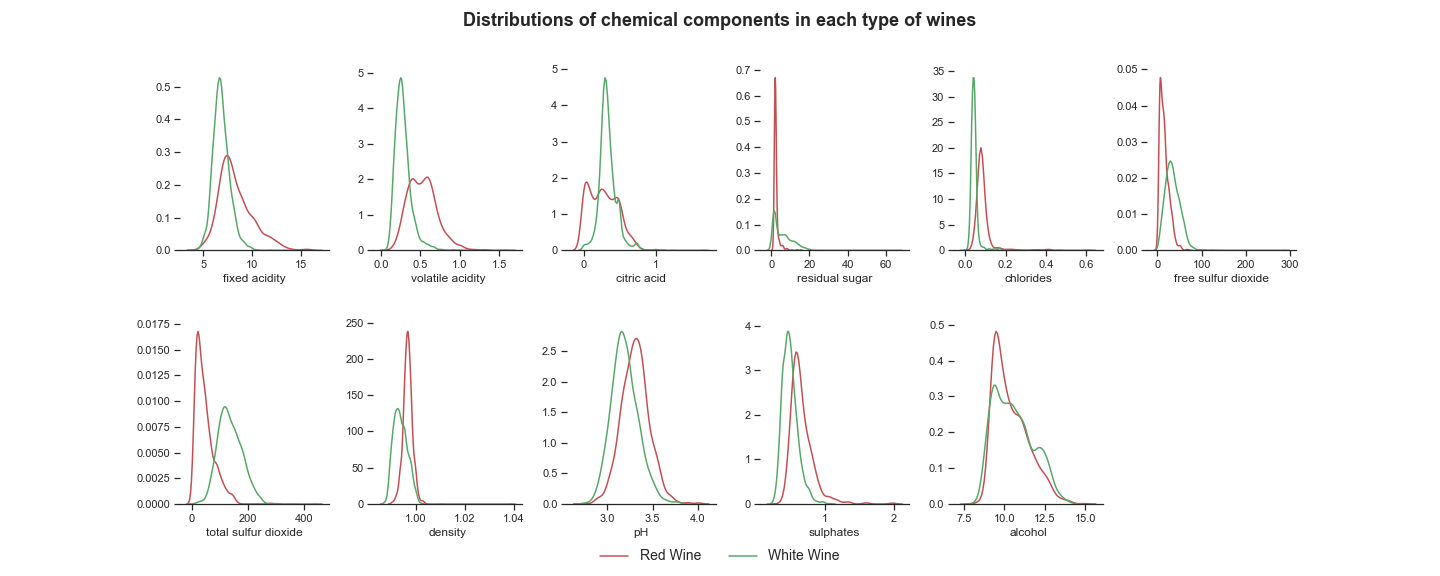
*Classifying different types of wines and predicting wine quality*

Key Summary

|  |  |
| --- | --- |
| White wine has better quality on average. | 5 different groups of wines were found. |
| Unbalanced classes makes predicting wine quality unreliable. | 99% accuracy classifying white or red wine. |

Sample statistics

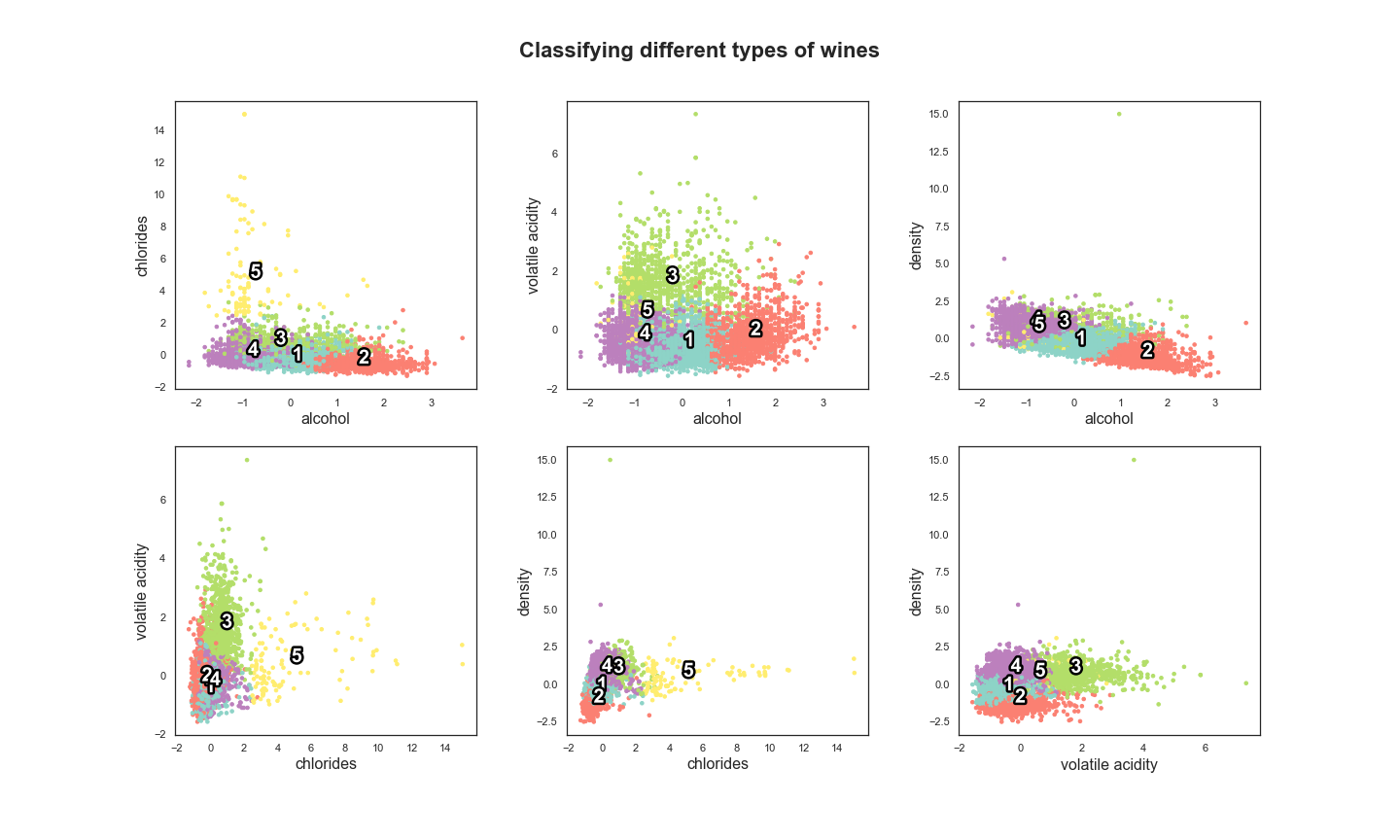
|  |  |
| --- | --- |
| * After duplicate entries were removed, there are **1359** samples of red wines and **3961** samples of white wine. * Initial analysis shows that:   + White wines tend to have a higher concentration of *sulfur dioxide* and *residual sugar*.   + The maximum level of *residual sugar* found in white wines is over **4 times** as the maximum found in red wines.   + The average amount of *chlorides* in red wine is twice the amount found in white wine. | * The average white wine has a *quality* of 6, one rating higher than the average red wine. * A significant *t-value* (-8.75) were found when analysing the mean difference in quality. * The minimum quality rating for both types of wines is 3. * The maximum rating for white wines is 9 while it is 8 for red wines. * The distribution of the *quality* rating is very unbalanced. Most of the wines have a rating of 5 or 6. * The wines are divided into **3 categories** according to their ratings. |



|  |  |
| --- | --- |
| * Most of the distributions are left skewed and clustered around a single, sharp peak. * There is a clear difference between the components of red and white wines. | * The peaks of the distributions of *residual sugar,* *sulphur dioxide, density* in red wine is much higher. * The distribution of different types of acid in red wines are more spread out than in white wines. |

Classifying different types of wines

Using K-means, 5 different groups of wines were classified using the four chemicals that affected wine quality the most: **Alcohol**, **Chlorides**, **Volatile** **Acidity**, and **Density.**



|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Alcohol | Chlorides | Volatile Acidity | Density |
| Group 1 | Medium | Low | Low | Medium |
| Group 2 | High | Low | Low | Low |
| Group 3 | Medium | Medium | High | High |
| Group 4 | Low | Medium | Low | High |
| Group 5 | Low | High | Medium | High |

Predicting type of wines and wine quality

* Using Random Forest Classifier, the model was able to classify the type of wine (*red* or *white*) at a 99% accuracy rate on the testing set. The features that the model found to be the most informative are: **Volatile acidity**, **Chlorides**, **Total sulfur dioxide**, **Density**, and **Sulphates.**
* To classify the categories of quality of wines, different machine learning algorithms were trained on the original dataset as well as an *oversampled* and *undersampled* dataset.

|  |  |
| --- | --- |
| * The best performing algorithm is Random Forest Classifier with an 81% testing accuracy score. * The most informative feature is the level of **alcohol.** * Due to the heavily unbalanced class distribution, the model has a very low recall score on classifying low quality wines, i.e. the model is not good at identify the low quality wines. * Even when the dataset is *oversampled* or *undersampled*, the recall score only improved slightly. * Removing the outliers also did not improve model performance. | C:\Users\BUI\AppData\Local\Microsoft\Windows\INetCache\Content.Word\Feature importances.png |