Practical Data Science Assignment 2

Mitchell Susa s3601130

Bhargav Rele s3761977

Table of Contents

06/05/2019

Summary

The aim of this report was to develop a classification model that could enable us to predict the colour of wine, based on various attributes of wine. The two datasets used in the analysis refer to red and white variants of the Portuguese 'Vinho Verde' wine. Overall, six classification models were created (3 different types of splits over 2 different types of classification models). An evaluation of the performance measures of each classification model led us to choose a classification model that had a 99% accuracy in predicting wine colour based on wine attributes. It is recommended that the chosen classification model be used to evaluate whether or not the attributes of produced wine are suitable, or in other words, up to standard for its respective colour.

Introduction

The stark contrast between red and white wine can be observed in its various features. In particular, the features we can examine are fixed acidity, volatile acidity, citric acidity, residual sugar, chlorides, sulphur dioxide (free and total), density, pH, sulphates, alcohol and quality of wine. Each aforementioned feature plays a different role in the production process for different colours of wine i.e., the chemical components of wine vary across the various colours of wine [4]. The purpose of our classification model was to be able to accurately (to a certain degree) predict the colour of wine based on its chemical components, by utilising machine learning techniques. Thus, the classification model developed forms a small but relevant part of a larger study that focuses on how different colours of wine consist of different amounts of the same chemical components which in turn may provide varying health benefits.

Methodology

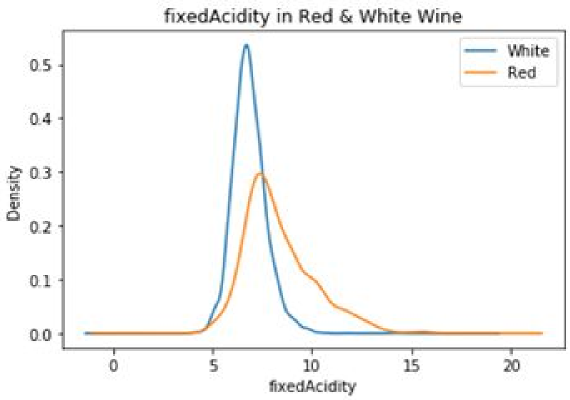
The two datasets used in the analysis refer to red and white variants of the Portuguese 'Vinho Verde' wine. The red wine dataset included 1,599 red wine observations while, the white wine dataset consisted of 4,898 observations. The aforementioned attributes amount to a total of 12 attributes for each wine colour [1]. The analysis included a data retrieval phase, data pre-processing phase, data exploration phase and a data modelling phase that was carried out using Python.

The *data retrieval and pre-processing phases* were conducted in order to prepare our dataset for analysis processes. This included the importation of a csv file into python, checking data types and ensuring the number of observations and attributes were as per the specifications provided earlier. In addition, any potential data entry errors (example; extra whitespace, capital letters, redundant/repeating data entries) were gotten rid of in order to 'clean' the data. The *data exploration phase* was then commenced in order to explore the individual attributes and pairs of attributes by using various graphical representations. The individual attributes of red wine were visualised with the same individual attributes of white wine in the same graph. This provided a means to easily compare the distribution of the same attribute for different wine colours. Pairs of attributes were then compared for white wine and red wine together. This was done by merging the two datasets such that we have a unified dataset consisting of red and white wine observations. Pairs of attributes were then graphed. This was done in order to observe the relation between two attributes regardless of wine colour. On observing various visualisations of wine attributes, we were ready to begin modelling our dataset. The *data modelling phase* was conducted with the purpose of classifying our observations into wine colours based on attributes the observation possessed. In order to do so, a column called 'wineColour' that consisted of levels 1 (for white wine) and 0 (for red wine), was added to our dataset. This column represented the variables (target variable) that we intend on predicting or, classifying our wine observations into. The models used for the classification process was the K-Nearest Neighbour(KNN) method and the Decision-Tree Classifier(DTC). The training-testing split used for classification included a 50%-50%, 60%-40% and 80%-20% training-testing split where, the former number is the size of the training sample and the latter number is the size of the testing sample. Each split was used to create a KNN model and a DTC model. The accuracy, confusion matrix, precision, recall and f1-score of each model was then reported to measure and compare the performance of each model.

Results

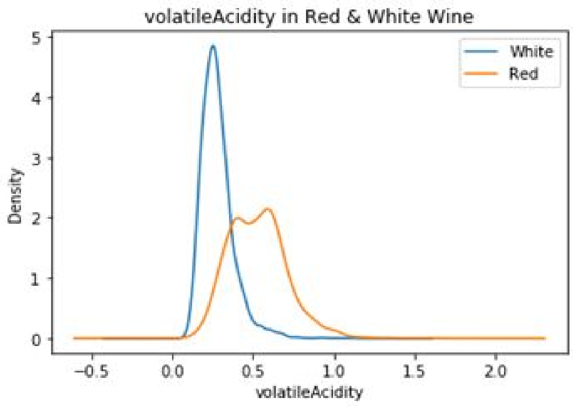
Individual Attribute Graphs

Fixed Acidity [figure 1]



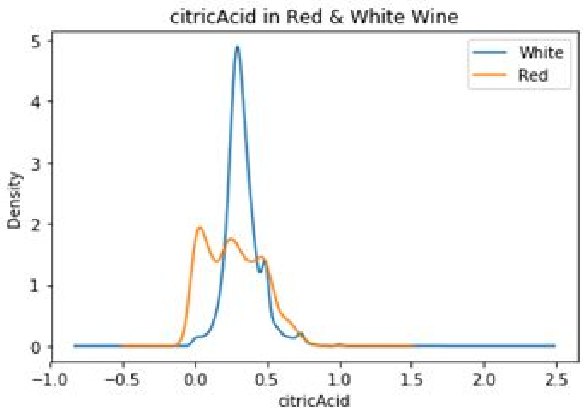
The results show that Fixed Acidity has a big peak over the 5-10 range. Therefore, showing that fixed acidity is heavily mid-range biased. This is even further accentuated in white wine compared to red wine. Red wine is shown to have a greater acidity on average.

Volatile Acidity [figure 2]



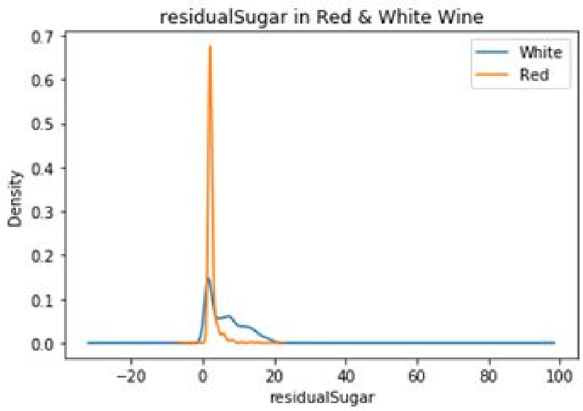
The results show a minor disparity between red and white wine in volatile acidity. Where white wine has a more defined peak of around 0.2, whereas red wine has a more balanced peak between 0.4 and 0.6. Overall, both lines show that there is a low amount bias, where a majority of the data is in the lower amounts.

Citric Acid [figure 3]



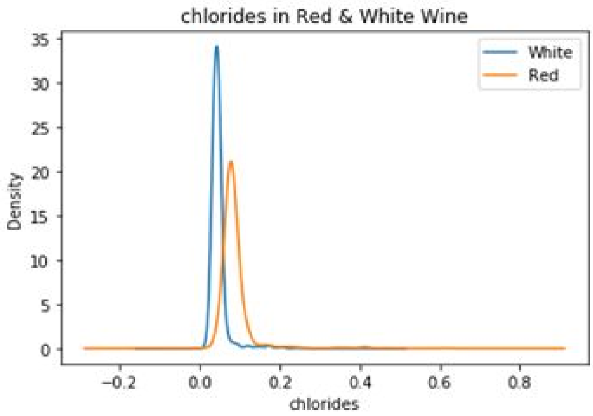
This graph shows how white wine has a more defined peak of around 0.2 to 0.4, and red wine has a more balanced spread between -0.1 and 0.5. Overall, the results for both wines hover between 0.0 and 0.5.

Residual Sugar [figure 4]



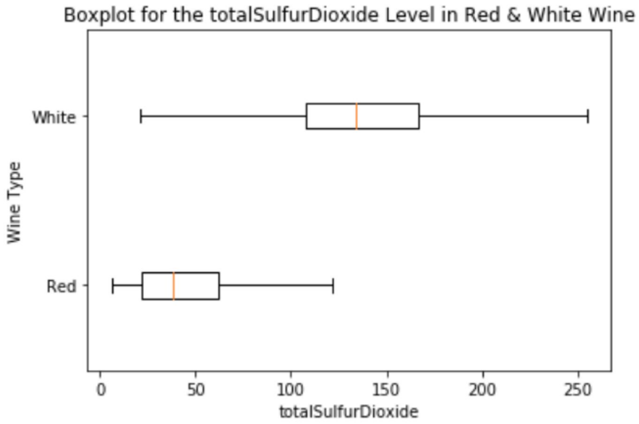
This graph has the biggest disparity in white and red wine compared to other results. Here we see red wine has a definite peak around 2, and red wine has a slight peak around 0 then goes downwards towards 20. Overall both wines hover around 5.

Chlorides [figure 5]



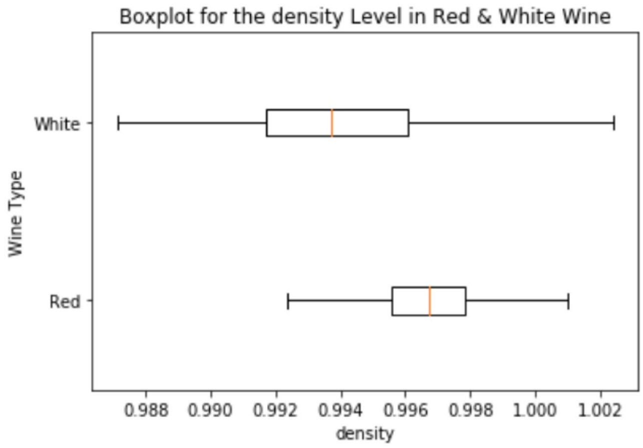
This graph has minor differences between white and red wine. White wine has a peak around 0-0.1, whereas red wine has a peak around 0.1-0.15. Overall, the results hover around 0-0.15.

Total Sulfur-dioxide [figure 6]



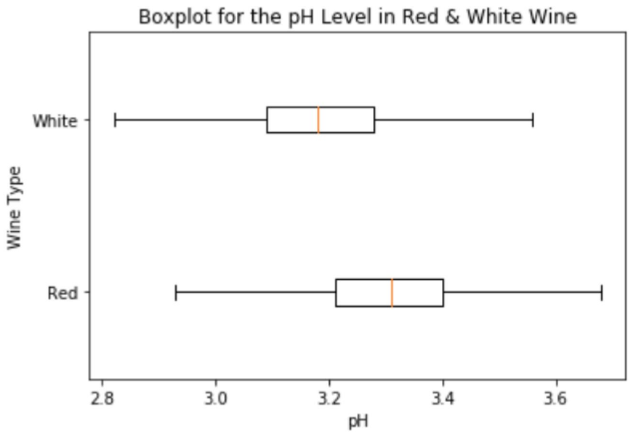
The boxplot for the total Sulfur-dioxide content of each wine tells us that the mean total Sulfur-dioxide content of red wine is lesser than that of white wine. Red wine also has a smaller interquartile range.

Density [figure 7]



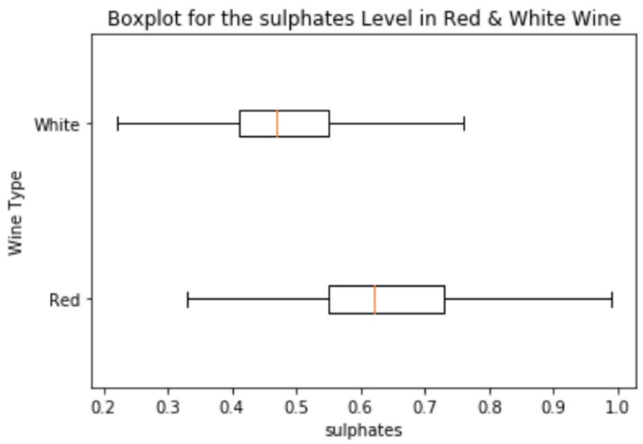
The boxplot for the density of each of wine tells us that the mean density of red wine is greater than that of white wine. Red wine also has a small interquartile range.

pH Level [figure 8]



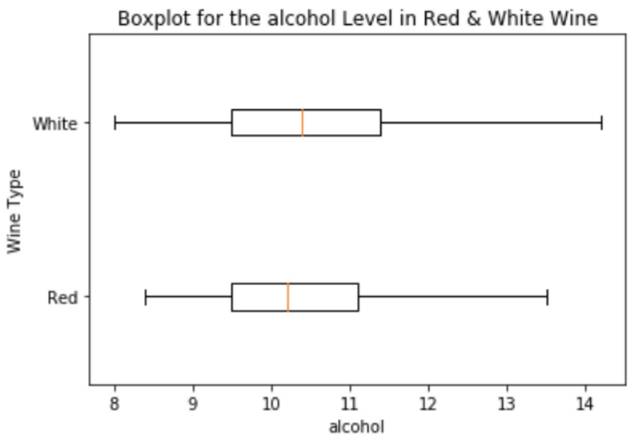
The boxplot for the pH levels of each of wine tells us that the mean pH levels of red wine is greater than that of white wine. Red wine and white wine have a similar interquartile range.

Sulphates [figure 9]



The boxplot for the sulphate content of each wine tells us that the mean sulphate content of red wine is greater than that of white wine. Red wine has a larger interquartile range.

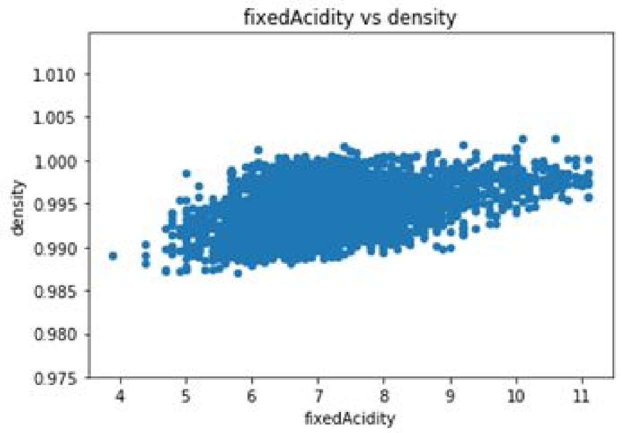
Alcohol [figure 10]



The boxplot for the alcohol levels in each wine tells us that the mean alcohol content of red wine is lower than that of white wine. Red wine and white wine have a similar interquartile range.

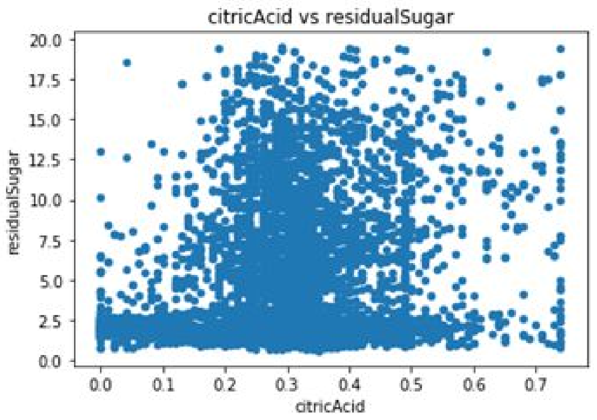
Pairs of Attributes Graphs

Fixed Acidity - Density [figure 11]



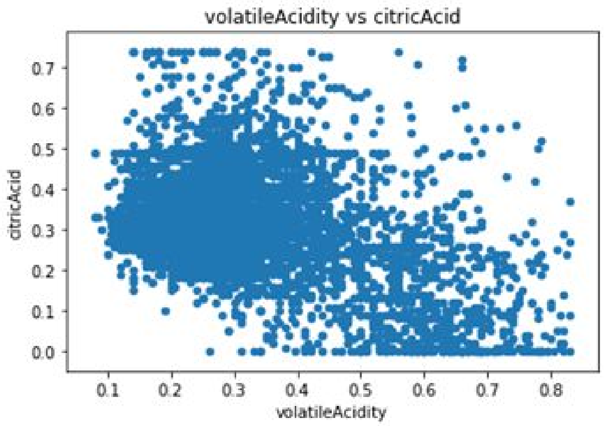
This graph shows an increase in density as fixed acidity increases, this is very minor compared to other results. Another thing we can get from this graph is that density has results that are very grouped together.

Citric Acid - Residual Sugar [figure 12]



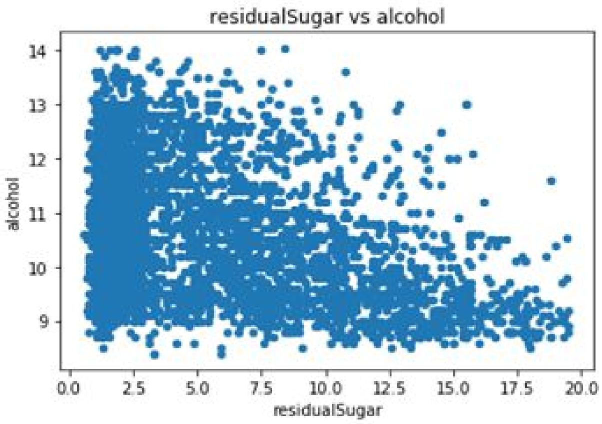
This graph shows that residual sugar has a big increase around the mid-range for citric acid, has a low when citric acid is lower, and is random at the highest levels of citric acid.

Volatile Acidity - Citric Acid [figure 13]



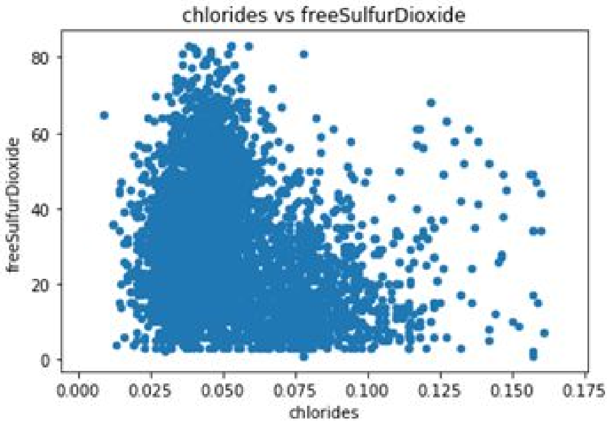
This graph shows a trend of decreasing values of citric acid as volatile acidity increases. It also shows a large grouping around lower values of volatile acidity.

Residual Sugar - Alcohol [figure 14]



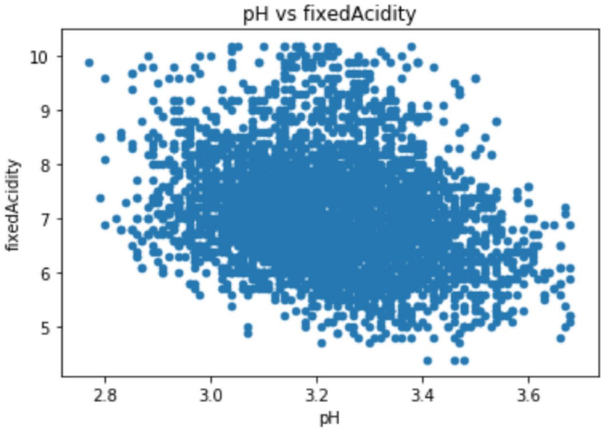
This graph shows that as residual sugar increases, alcohol decreases. The peak of alcohol is shown when residual sugar is at its lowest.

Chlorides - Free Sulfur Dioxide [figure 15]



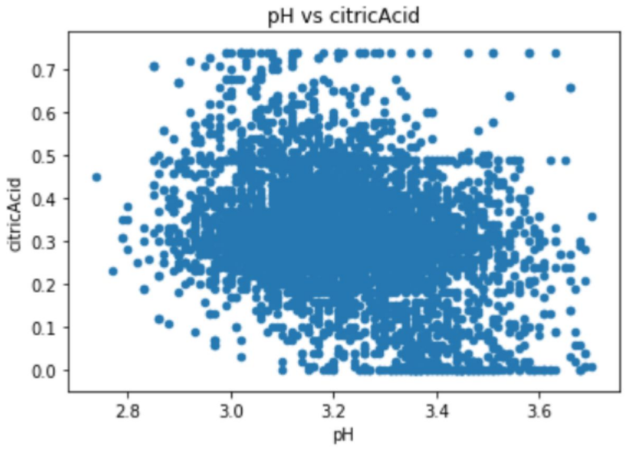
This graph shows a large decrease in free sulphur dioxide as chlorides increase. The peak of free sulfur dioxide is around 0.03 chlorides.

pH - Fixed Acidity [figure 16]



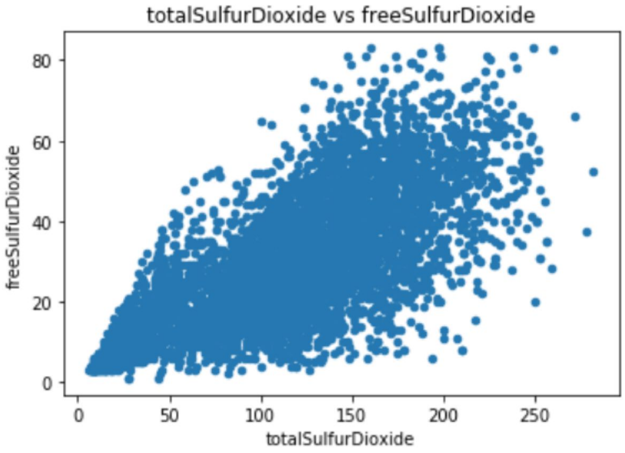
Hypothesis: Negative relation. A lower pH is associated with a higher fixed acidity [2]. As confirmed by the scatterplot, we observe low[high] fixed acidity for wines with higher[lower] pH levels.

pH - Citric Acid [figure 17]



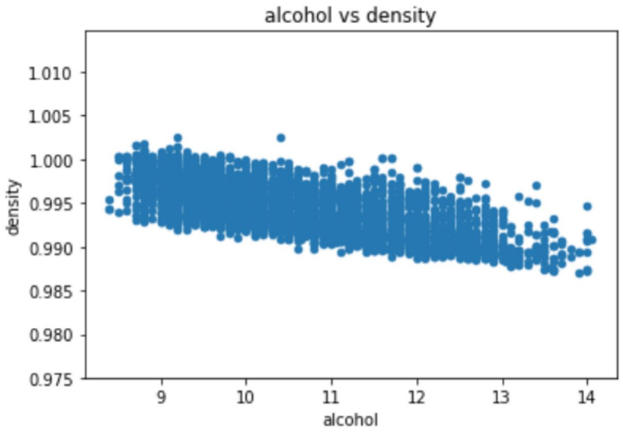
Hypothesis: Negative relation. Citric acid is a type of fixed acid [3]. Hence, if fixed acidity shares a negative relationship [figure 16] with the pH level of wine, citric acid would mirror this negative relationship. This can be observed in the scatterplot.

Total Sulfur-Dioxide - Free Sulfur-Dioxide [figure 18]



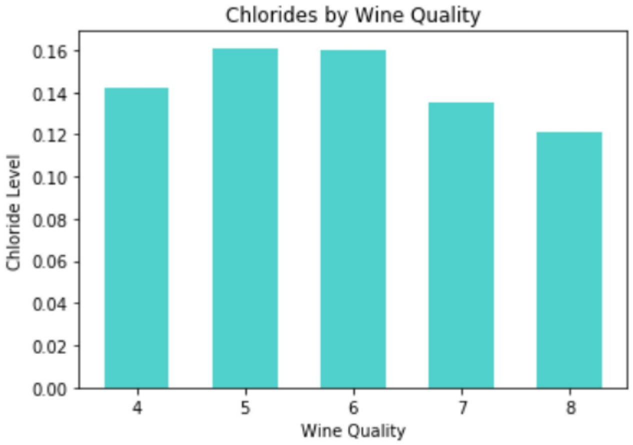
Hypothesis: Positive relation. Total Sulfur-dioxide content includes free Sulfur-dioxide, which is added by the producer to prevent wine from oxidising [4]. As confirmed by the scatterplot, we observe high[low] total Sulfur-dioxide content to be be associated with high[low] free sulfur-dioxide content.

Alcohol - Density[figure 19]



Hypothesis: Negative relation. A higher alcohol content is associated with a lower density as there is less sugar in wines that have a high alcohol content [5]. As confirmed by the scatterplot, we observe low[high] alcohol levels for wines with higher[lower] density.

Chlorides - Quality [figure 20]



Hypothesis: Chlorides in wine leave as undesirable salty taste [6]. This reduces the quality of the wine. Hence we expect high quality wines to have a lower chloride content. This can be observed in the bar plot. We can see that wine of highest quality has the lowest chloride content.

Data Modelling using Machine Learning Results (0 = Red Wine, 1 = White Wine)

K-Nearest Neighbour Classifier

50% Train, 50% Test

Accuracy Train vs Accuracy Test:

- Accuracy Train = 0.9919839679358717
- Accuracy Test = 0.9876460767946578
- Overfitting (as accuracy train > accuracy test)

Confusion Matrix:

```
[[ 611   20]
 [  17 2347]]
```

Classification Report:

	precision	recall	f1-score	support
0	0.97	0.97	0.97	631
1	0.99	0.99	0.99	2364

Performance Summary:

According to the *confusion matrix* of the model, red wine (denoted as 0) was predicted correctly 611 times whereas, white wine (denoted as 1) was predicted correctly 2,347 times. In addition, white wine was wrongly predicted as red wine 20 times whereas, red wine was wrongly predicted as white wine 17 times. According to the *classification report,* the f1-score for red wine is 0.97 and the f1-score for white wine is 0.99. A precision of 0.97 for red wine implies that out of 628 actual red wine observations, 97% were correctly predicted. A recall of 0.97 implies that out of 631 predicted red wine observations, 97% were in fact red wine observations. A precision of 0.99 for white wine implies that out of 2,367 actual white wine observations, 99% were correctly predicted. A recall of 0.99 implies that out of 2,364 predicted white wine observations, 99% were in fact white wine observations.

60% Train, 40% Test

Accuracy Train vs Accuracy Test:

- Accuracy Train = 0.9908154745338158
- Accuracy Test = 0.9908180300500835
- Not Overfitting (as accuracy train < accuracy test)

Confusion Matrix:

```
[[ 502   12]
 [  10 1872]]
```

Classification Report:

	precision	recall	f1-score	support
0	0.98	0.98	0.98	514
1	0.99	0.99	0.99	1882

Performance Summary:

According to the *confusion matrix* of the model, red wine (denoted as 0) was predicted correctly 502 times whereas, white wine (denoted as 1) was predicted correctly 1,872 times. In addition, white wine was wrongly predicted as red wine 12 times whereas, red wine was wrongly predicted as white wine 10 times. According to the *classification report,* the f1-score for red wine is 0.98 and the f1-score for white wine is 0.99. A precision of 0.98 for red wine implies that out of 512 actual red wine observations, 98% were correctly predicted. A recall of 0.98 implies that out of 522 predicted red wine observations, 98% were in fact red wine observations. A precision of 0.99 for white wine implies that out of 1,884 actual white wine observations, 99% were correctly predicted. A recall of 0.99 implies that out of 1,882 predicted white wine observations, 99% were in fact white wine observations.

80% Train, 20% Test

Accuracy Train vs Accuracy Test:

- Accuracy Train = 0.9908161135462326
- Accuracy Test = 0.994991652754591
- Not Overfitting (as accuracy train < accuracy test)

Confusion Matrix:

```
[[238   2]
 [  4 954]]
```

Classification Report:

	precision	**recall**	**f1-score**	**support**
0	0.98	0.99	0.99	240
1	1.00	1.00	1.00	958

Performance Summary:

According to the *confusion matrix* of the model, red wine (denoted as 0) was predicted correctly 238 times whereas, white wine (denoted as 1) was predicted correctly 954 times. In addition, white wine was wrongly predicted as red wine 2 times whereas, red wine was wrongly predicted as white wine 4 times. According to the *classification report,* the f1-score for red wine is 0.99 and the f1-score for white wine is 1.00. A precision of 0.98 for red wine implies that out of 242 actual red wine observations, 98% were correctly predicted. A recall of 0.99 implies that out of 240 predicted red wine observations, 99% were in fact red wine observations. A precision of 1.00 for white wine implies that out of 956 actual white wine observations, 100% were correctly predicted. A recall of 1.00 implies that out of 958 predicted white wine observations, 100% were in fact white wine observations.

Decision Tree Classifier

50% Train, 50% Test

Accuracy Train vs Accuracy Test:

- Accuracy Train = 0.9402137608550434
- Accuracy Test = 0.9409015025041736
- Not Overfitting (as accuracy train < accuracy test)

Confusion Matrix:

```
[[ 459  172]
 [   5 2359]]
```

Classification Report:

	precision	**recall**	**f1-score**	**support**
0	0.99	0.73	0.84	631
1	0.93	1.00	0.96	2364

Performance Summary:

According to the *confusion matrix* of the model, red wine (denoted as 0) was predicted correctly 459 times whereas, white wine (denoted as 1) was predicted correctly 2,359 times. In addition, white wine was wrongly predicted as red wine 172 times whereas, red wine was wrongly predicted as white wine 5 times. According to the *classification report,* the f1-score for red wine is 0.84 and the f1-score for white wine is 0.96. A precision of 0.99 for red wine implies that out of 464 actual red wine observations, 99% were correctly predicted. A recall of 0.73 implies that out of 631 predicted red wine observations, 73% were in fact red wine observations. A precision of 0.93 for white wine implies that out of 2,531 actual white wine observations, 93% were correctly predicted. A recall of 1.00 implies that out of 2,364 predicted white wine observations, 100% were in fact white wine observations.

60% Train, 40% Test

Accuracy Train vs Accuracy Test:

- Accuracy Train = 0.9473977177845812
- Accuracy Test = 0.9478297161936561
- Not Overfitting (as accuracy train < accuracy test)

Confusion Matrix:

```
[[ 396  118]
 [   7 1875]]
```

Classification Report:

	precision	recall	f1-score	support
0	0.98	0.77	0.86	514
1	0.94	1.00	0.97	1882

Performance Summary:

According to the *confusion matrix* of the model, red wine (denoted as 0) was predicted correctly 396 times whereas, white wine (denoted as 1) was predicted correctly 1,875 times. In addition, white wine was wrongly predicted as red wine 118 times whereas, red wine was wrongly predicted as white wine 7 times. According to the *classification report,* the f1-score for red wine is 0.86 and the f1-score for white wine is 0.97. A precision of 0.98 for red wine implies that out of 403 actual red wine observations, 98% were correctly predicted. A recall of 0.77 implies that out of 514 predicted red wine observations, 77% were in fact red wine observations. A precision of 0.94 for white wine implies that out of 1,993 actual white wine observations, 94% were correctly predicted. A recall of 1.00 implies that out of 1,882 predicted white wine observations, 100% were in fact white wine observations.

80% Train, 20% Test

Accuracy Train vs Accuracy Test:

- Accuracy Train = 0.9580463368816531
- Accuracy Test = 0.9632721202003339
- Not Overfitting (as accuracy train < accuracy test)

Confusion Matrix:

```
[[206  34]
 [ 10 948]]
```

Classification Report:

	precision	recall	f1-score	support
0	0.95	0.86	0.90	240
1	0.97	0.99	0.98	958

Performance Summary:

According to the *confusion matrix* of the model, red wine (denoted as 0) was predicted correctly 206 times whereas, white wine (denoted as 1) was predicted correctly 948 times. In addition, white wine was wrongly predicted as red wine 34 times whereas, red wine was wrongly predicted as white wine 10 times. According to the *classification report,* the f1-score for red wine is 0.90 and the f1-score for white wine is 0.98. A precision of 0.95 for red wine implies that out of 216 actual red wine observations, 95% were correctly predicted. A recall of 0.86 implies that out of 240 predicted red wine observations, 86% were in fact red wine observations. A precision of 0.97 for white wine implies that out of 982 actual white wine observations, 97% were correctly predicted. A recall of 0.99 implies that out of 958 predicted white wine observations, 99% were in fact white wine observations.
[Note: All ‘observations’ refer to the test observations and, all f1-scores, precisions and recall have been rounded up to two decimal places]

Discussion

For the comparison of attributes and wine colour we initially compared the same element in two different wine colours, and then combined the red wine dataset and white wine dataset to compare different attributes. For the prediction of wine colour based on machine learning techniques, we attempted to use the data modelling technique of classification and, explored the accuracy of each model given its model type and training-test split.

For our attribute comparison, a variety of attributes were compared for red and white wine due to which, this discussion will only be touching upon the broader findings. Firstly, we compared the three acidity attributes [figure 1,2,3], the conclusion we came to was that white wine had a lower range of acidity (proven by the large spikes in the data), whereas red wine had the tendency to have a wider range and higher acidity. This implies that the red wine observations inspected consisted of lighter reds (as heavier reds have less acidity) [7]. In addition, the visualisation gave less meaning to the theory that white wine is generally more acidic than red wine [7]. A high total Sulfur Dioxide (sulphite) level [figure 6] is required to extend the shelf life of wine[8]. Therefore, seeing on average that white wine had a larger count of Sulphites, it was implied that white wine without preservatives, had a lower shelf life in comparison to red wine without preservatives. Higher Chloride levels in wine brings forth a saltier taste [6]. When Chloride levels in red and white wine were compared, it was observed that red wine had a higher Chloride level which suggested, red wine had a saltier taste in comparison to white wine. Density was visualised for each wine in figure 7. If a wine has higher density, then it likely has more sugar and less alcohol (the opposite is also true) [9]. Hence, seeing that red wine had a higher density to white wine in our dataset, we can assume that red wine had a lesser alcohol content in comparison to white wine. This was further proven in the alcohol comparison [figure 10] where white wine was seen to have slightly higher alcohol content than red wine. Higher levels of residual sugars [figure 4] means that the wine is sweeter [10]. Therefore from the graph, we determined that white wine was on average, sweeter than red wine. Finally, higher pH levels generally means that the wine needs more Sulphates [figure 9] for the purpose of preservation [13]. This was observed in the graphs as red wine had higher levels of pH and Sulphates. Overall, the individual attributes and their visualisations were informative when comparing the chemical components comprised by wines of different colour. However, direct comparisons of the chemical components was also informative in comparing the relationship and correlations between each attribute, regardless of wine colour.

To compare pairs of attributes, scatter plots were used to display the data clearly and effectively. This is due to a scatter plot’s ability to visualise how an attribute behaves when another attribute in comparison either increases or decreases. Firstly, the discovery of attributes that were perceived to have very little expected relevance to each other were interesting to explore. For instance, fixed acidity and density [figure 11]. Figure 11 visualised only a slight but negligible increase in density with an increase in fixed acidity. This may have been because density is influenced by sugar and alcohol levels, not acidity [9]. A figure that further examined higher density and its positive association to sugar and negative association to alcohol content [9], was figure 14 and figure 19. It was observed that higher levels of residual sugar were associated with lower levels of alcohol (and vice-versa) and, higher levels of alcohol were associated with lower levels of density. Generally, as the producer increases the total Sulphite content, the free-Sulphite content also increases, so as to delay the oxidation process of the wine. This was observed in figure 18, where we saw that an increase in total Sulphite content was associated with an increase in free-Sulphite content. Obvious relations were also explored to test the validity [of] and observe the correlations between attributes within our dataset. For instance, figure 16 and 17; where it was observed that lower pH levels were associated with higher acidity levels. Citric acid was used for comparison in two figures; figure 12 and figure 13. In the former figure, we learned that extreme observations of citric acid lowered the amount of residual sugar dramatically. The second graph showed decreases in citric acid associated with an increase in volatile acidity.This may be because fixed acidity does not include citric acid as a component [3]. Figure 15 provided an interesting comparison of Chlorides and free Sulfur Dioxide. Higher levels of Sulphite leads to stronger flavours [14], and higher levels of Chlorides leads to saltier flavours [6]. The salty taste of wine contributes to the strength of the wine flavour in an unfavourable manner. This implied a positive relation between Chlorides and Sulphite. The saltiness of wine taste caused by Chlorides although makes the wine taste stronger, is not preferred. This implies that wines with higher Chloride content, are considered to be of lower quality [6]. This was observed in figure 20, where the highest quality wines have the lowest chloride content. It is to be noted that most of our data is centered around the mean quality of 5 and 6. Hence, the difference in quality caused by various attributes could not have been clearly examined and would not have bee too informative on examination.

After having prepared the data for the modelling process, we proceeded to create a classification model to be able to classify an observation of wine into two colour categories, based on the respective observations various attributes. It is to be noted that during the feature selection process, attributes that were correlated, could have been removed or adjusted by using a Boruta Algorithm [15]. However, the decision was made to not get rid of any attributes as that would have beyond the scope of this subject. The classification process for any given model (KNN or DTC) requires parameter specifications to be able to improve accuracies and fitting capabilities of the model. For each *type (i.e., K-nearest neighbours and Decision-tree classifier)* of model, the parameters were chosen based on trial and error. The performance measurements were examined under each combination of parameter specifications that were tried. The parameter specifications that had resulted in a model with higher testing accuracy than training accuracy *for all three* *splits*, was chosen to prevent overfitting. The same parameter specification were applied on each training-test split under each model type, in order to allow for a comparison of each split on mutual grounds. The specificity of the parameters are observable in the codes attached to this report.

Of all the performance measurements for any given model, we were most concerned with the f1-score of each category as opposed to the accuracy of the whole model. This is because the model was based on a sample of 4,898 white wine observations and only 1,599 red wine observations which in turns made our data imbalanced and asymmetrical (in terms of number of observations under each category)[11]. Looking at the accuracy of a model derived from an imbalanced dataset, may have led us to make wrong decisions. The precision and recall of a category, on the other hand gives us information on not only the number of observation that were correctly predicted but also on the number of observations that were wrongly predicted. Hence, it gives us the ability to examine the 'costliness' of misclassifying an observation[12]. As the f1-score of a given category is the combination of both the precision and recall of that category, it was a more reliable measurement than accuracy of the model as a whole, during the model-evaluation process.

Conclusion

Of all K-Nearest Neighbours and Decision-Tree models, the model with the highest f1-score and accuracy is a K-nearest neighbours model that has an 80% training sample and a 20% testing sample. The f1-score for the model is 99% for red wine and close to 100% for white wine and the accuracy of the model is 99.5%. Hence, the model has a 1% chance of misclassifying red wine observations and a close to 0% chance of misclassifying white wine observations. Therefore, the classifications of wine colour based on its attributes have a 0.5% chance of being incorrect. Hence, given features and attributes of a newly produced wine, we would be able to say with a certain degree of accuracy, whether it could be considered red wine or white wine. The classification of wine based on its attributes form a relatively small part of a larger study on the various health benefits associated with wines of different colours simply due to its chemical components.

References

[1] P. Cortez et. al., "Modeling wine preferences by data mining from physicochemical properties. In Decision Support Systems", Elsevier, vol. 47, no. 4, pp. 547-553, 2009.
[2] Wikipedia, Acids in Wine. [Online]. Accessed on: May 29, 2017. Available: https://en.wikipedia.org/wiki/Acids_in_wine
[3]D. Nierman, What's in Wine: Fixed Acidity, 2004. [Online]. Accessed on: May 29, 2019. Available: https://waterhouse.ucdavis.edu/whats-in-wine/fixed-acidity
[4] M. Carel, Wine From Here, Oct. 11, 2011. [Online]. Accessed on: May 29, 2019. Available: https://winobrothers.com/2011/10/11/sulfur-dioxide-so2-in-wine/
[5] Bartenderly, Alcohol Density Chart, May 8, 2013. [Online]. Accessed on: May 29, 2019. Available: https://bartenderly.com/tips-tricks/alcohol-density-chart/
[6] M. Coli et. al., "Food Science & Technology", Vol. 35, no. 1, Mar. 2015.
[7] Total Wine & More, Wine Acidity & Crispiness. [Online]. Accessed on: May 29, 2019. Available: https://www.totalwine.com/wine-guide/wine-acidity-crispness
[8] M.G. Mcadams, The Truth About Sulfites in Wine & the Myths of Red Wine Headaches, Nov. 12, 2009. [Online]. Accessed on: May 29, 2017. Available: https://www.thekitchn.com/the-truth-about-sulfites-in-wine-myths-of-red-wine-headaches-100878
[9] ETS Laboratory, Density Wine. [Online]. Accessed on: May 29, 2019. Available: https://www.etslabs.com/analyses/DEN
[10] Whicher Ridge, What is Residual Sugar in Wine?. [Online]. Accessed on: May 29, 2019. Available: https://whicherridge.com.au/what-is-residual-sugar-in-wine/
[11] R. Joshi, Accuracy, Precision, Recall & F1 Score: Interpretation of Performance Measures, Sep. 9, 2016. [Online]. Accessed on: May 29, 2019 Available: https://blog.exsilio.com/all/accuracy-precision-recall-f1-score-interpretation-of-performance-measures/
[12] K.P. Shung, Accuracy, Precision, Recall or F1?, Mar. 15, 2018. [Online]. Accessed on: May 29, 2019. Available: https://towardsdatascience.com/accuracy-precision-recall-or-f1-331fb37c5cb9
[13] Australian Wine Making Institute, Acidity and pH. Accessed on: May 29, 2019. [Online]. Available: https://www.awri.com.au/industry_support/winemaking_resources/frequently_asked_questions/acidity_and_ph/
[14] J. Miquel, *Sulphites (SO2) in Wine: Top 7 Facts,* Mar. 2, 2017. Accessed on: May 29, 2019. [Online]. Available: http://socialvignerons.com/2017/03/02/sulphites-so2-in-wine-top-7-facts/
[15] B. Kursa and R. Rudnicki, "The All Relevant Feature Selection using Random Forest", *University of Warsaw,* p. 4, Jun. 2011.