Developing an Artificially Intelligent Wine Critic

Project Report by Ben Thompson-Watson

Abstract

The determination of a wine's quality has been ambiguous for quite some time. This study demonstrates a system built for predicting the quality of new uncritiqued wines with precision. The purpose of this research is to investigate which features of wine affect its quality for a winery in Portugal to help optimise production. Using feature importance ranking, a subset of features was constructed for a wine dataset and was applied to the training of a Random Forests machine learning model. The amount of alcohol and total sulphur dioxide was found to be the most important features facilitating to an 87% prediction accuracy from the model. This project definitively details the effects of these two important features and their relation to wine quality. The quality proposed by the model was however generalised to low/high classification, further research should go into a less generalised predictive model.

Notebook URL

https://colab.research.google.com/drive/1K_7C1rmaVx6pQ6ELS4gZBAavii9gHq59?usp=sharing

What was Done and How

1. Data Preparation & Visualisation

By exploring the provided datasets 'winequality-red.csv' and 'winequality-white.csv', It was found best to concatenate[4] them together using a 'key' attribute. This differentiated the wines making data processing easier as it was all in one place. The datasets consisted of 1599 red wine and 4898 white wine records. This is unbalanced and accounted for when plotting distributions by setting the parameter 'common_norm' to false, normalizing the density.

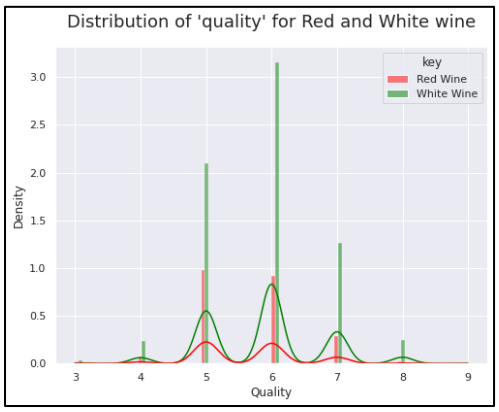

Figure 1.1

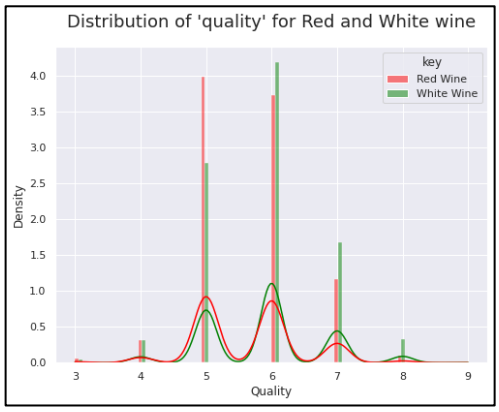

Figure 1.2

The distribution of quality for both datasets in Figure 1.2 exhibit white wine distributing more in higher quality classes in comparison to red wine. An assumption can be made that white wine is on average of higher quality than red wine. This indicates how it may be important to separate the wine types when training a machine learning model[3]. This is a visualisation improvement over Figure 1.1.

From figure 1.2's distribution of quality, it was found to be significantly unbalanced.

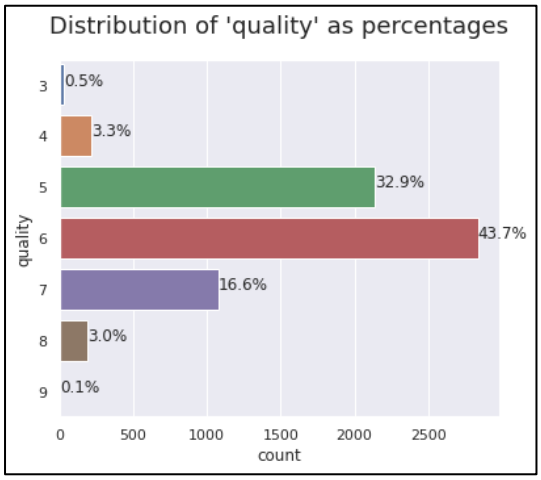

Figure 1.3

```
Frequency of records per quality level
----------------------------------------
Quality level 3 : 30
Quality level 4 : 216
Quality level 5 : 2138
Quality level 6 : 2836
Quality level 7 : 1079
Quality level 8 : 193
Quality level 9 : 5
```

Figure 1.4

In Figure 1.3 and 1.4, there are only 5 wines between both datasets that hold a quality value of 9 quality, covering only 0.1% of the datasets.

This would be problematic when training the machine learning model as predictions would skew towards predicting a quality of 5 or 6. To overcome this, the dataset will be adjusted preliminary to training the model using upsampling[10]. This should be done just before the training process because artificial records may affect some of the data preparation stages.

Outliers within datasets can greatly affect the accuracy of a model's predictions. It can confuse the algorithm to what a 'normal' value for that feature[1] is. To assist the selection of the bounds for each feature, the min and max distribution values were considered.

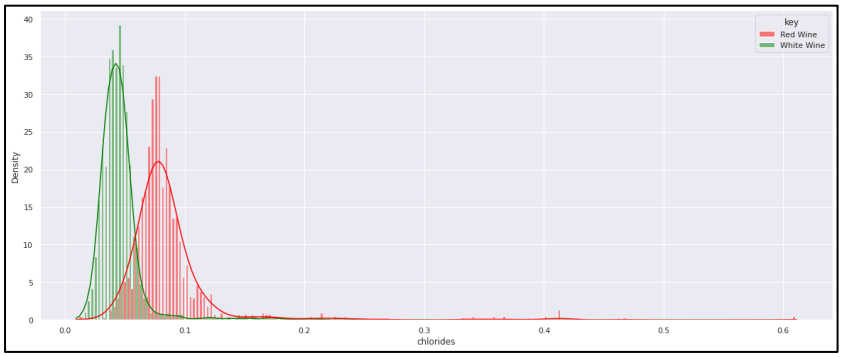

Figure 1.5

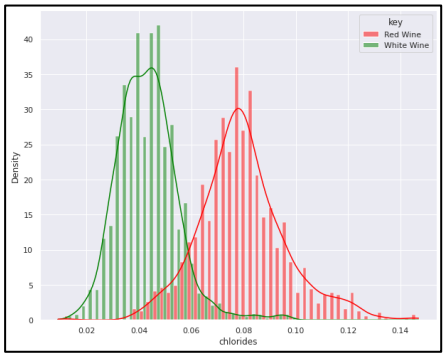

Figure 1.6

Figure 1.5 shows chloride levels for red wine reaching a maximum of 0.6. The distribution shows that this is for a very small number of records, therefore I have chosen to remove values reaching this magnitude. The bounds I set on chlorides for red wine is 0-0.15 and for white wine is 0-0.1. Figure 1.6 shows the resulting distribution, post outlier removal. All features with applicable outliers have been set with different upper and lower bounds.

Discretisation is a powerful method that can improve the accuracy of the model prediction. According to Gupta, ‘certain models may be incompatible with continuous data’ (Gupta, 2019). Continuous features would be recognised by models such as RF as multiclass and would decrease performance when training. I.e. alcohol with 111 different values between the datasets.

By discretising alcohol, using the mean and standard deviation (stdev), there are now three categories ‘low’, ‘mid’ and ‘high’. Anything below the mean – stdev falls into ‘low’, anything above the mean + stdev falls into high and anything in between falls into mid. The mean and stdev were calculated for each wine type separately.

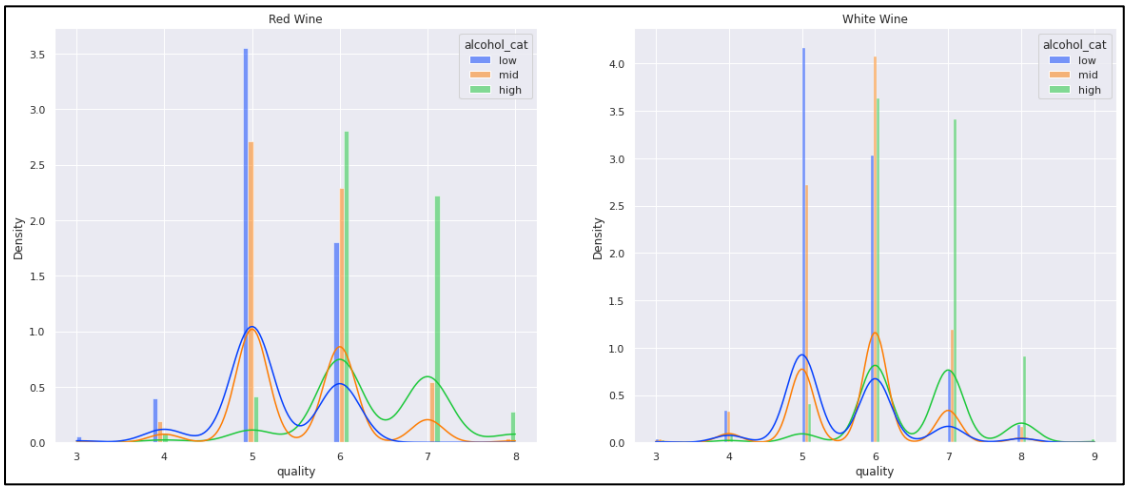

Figure 1.7

Following an analysis of figure 1.7, a relationship between alcohol levels and quality was derived, showing wines with a high value appearing frequently in quality levels 6+. This positive correlation appears in both wine types, suggesting that alcohol should be selected to train the model.

A second discretisation was to split the residual sugar feature into a binary classification. Note the threshold set does not accurately determine a sweet wine as according to Wine Folly it should be over 21 to be considered ‘sweet’ (*Sugar in Wine Chart (Calories and Carbs)*, 2015). The threshold actually set needed to split the distribution equally. Not relying entirely on separate wine machine learning models, a third threshold was set on a concatenated dataset. Later in the machine learning stage, each dataset model will be compared to find a successor. Figure 1.8 outlines the results:

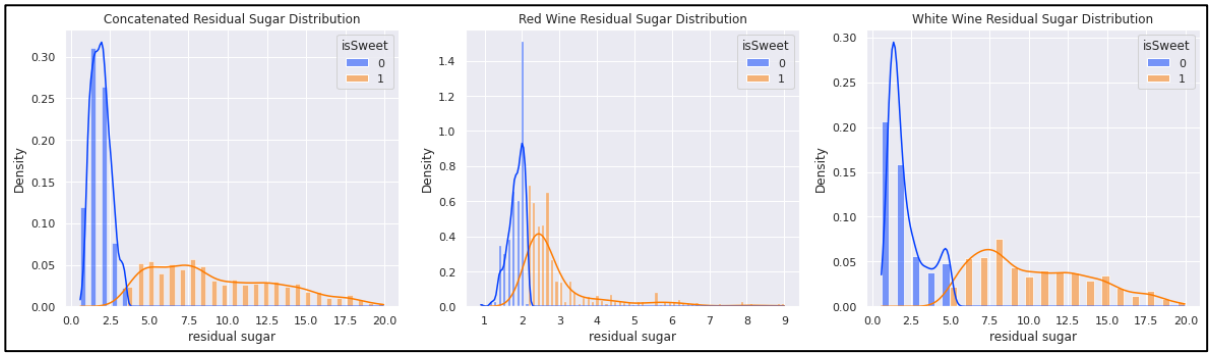

Figure 1.8

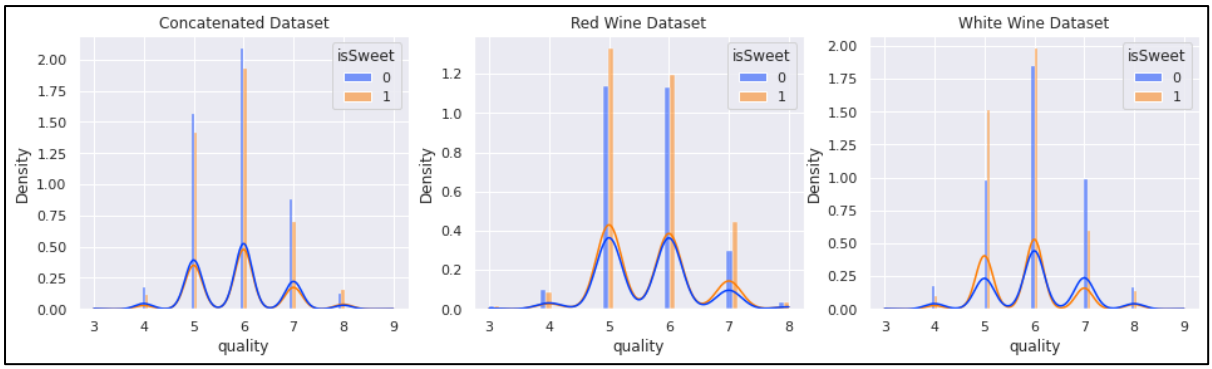

Figure 1.9

Figure 1.9 renders the comparison between quality and isSweet for each dataset. There is no significant correlation between these variables as both sweet and not sweet wines appear to be distributed evenly across all quality levels. This is insufficient evidence to include this feature in the training subset.

2. Correlation Analysis and Feature Selection

When selecting the best features for a machine learning model, it is useful to analyse a correlation matrix over all the features in the dataset. This highlights relationships between pairs of values and helps determine which one of the two should be redundant. For this analysis, the Kendall metric was chosen. Favoured over Pearson by being able to work well with nominal data such as alcohol_cat and over Spearman which "is more sensitive to error and discrepancies in the data" (kjytay, 2019).

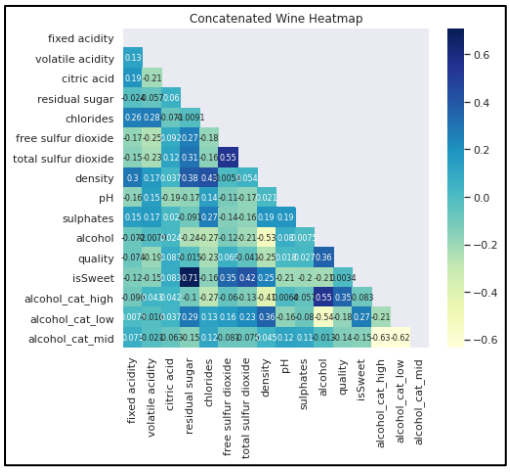

Figure 2.1

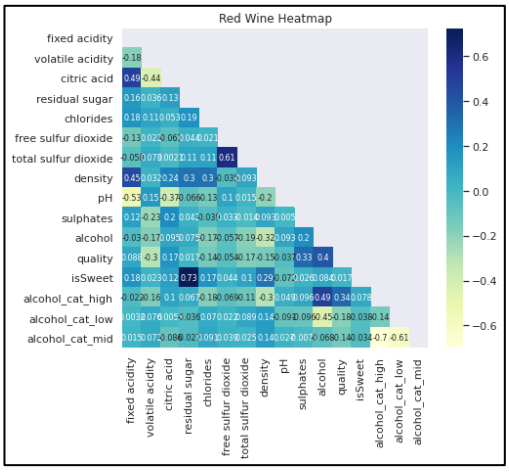

Figure 2.2

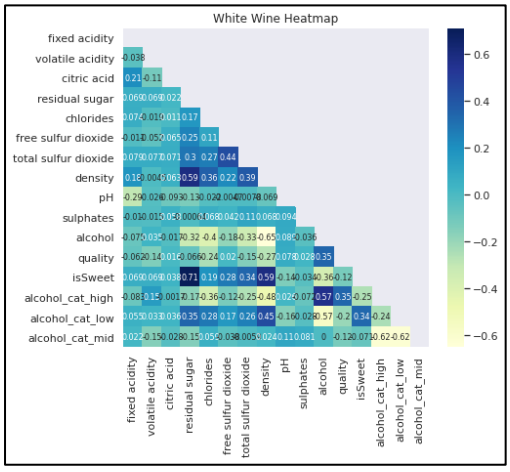

Figure 2.3

Figures 2.1 – 2.3 display a correlation matrix in the form of a heat map for each of the datasets to be tested. For all three datasets, the inclusion of the feature alcohol in the training subsets is reinforced as it showcases a strong correlation with quality. A high correlation between a feature and the target[2] will improve the accuracy of the model.

For both the concatenated and white wine dataset free sulphur dioxide and density correlate with numerous other features. This makes these variables redundant and less useful in their training subsets.

Another useful metric to consider is the variance of each feature. If a feature has a very small variance (small distribution) it could suggest that this feature is not useful because the record values do not change significantly. This can make it difficult for a model to detect patterns in a dataset.

```
Three of the lowest variances for features in concatenated dataset
-----------------------------------------------------------------------
density      0.000009
chlorides    0.000412
sulphates    0.018336
dtype: float64
-----------------------------------------------------------------------

Three of the lowest variances for features in red wine dataset
-----------------------------------------------------------------------
density      0.000003
chlorides    0.000277
sulphates    0.017984
dtype: float64
-----------------------------------------------------------------------

Three of the lowest variances for features in white wine dataset
-----------------------------------------------------------------------
density             0.000008
chlorides           0.000131
volatile acidity    0.008930
dtype: float64
-----------------------------------------------------------------------
```

Figure 2.4

Figure 2.4 reveals the three lowest variance features for each dataset captured from the variance analysis section of the studies notebook.

The features density and chlorides appear in all three datasets with significantly low variances.

These features would then be considered as potential exclusions from the training subset.

Granted that there are benefits of examining correlation and variance in the feature selection stage, it only looks at individual or pairs of features. Machine learning algorithms use an entire training dataset to build a model, building links between multiple features.

The scikit-learn library includes a feature selection class called SelectKBest. This class can then be used with a selected statistical test to rank each feature. The test ‘chi2’ is used as datasets contain both continuous and categorical data and according to Brownlee a ‘common correlation measure for categorical data is the chi-squared test’. (Brownlee, 2019).

After the SelectKBest class is trained, the ‘k-scores’ are retrieved and ranked for each feature. The highest being the most useful and the lowest being the least useful.

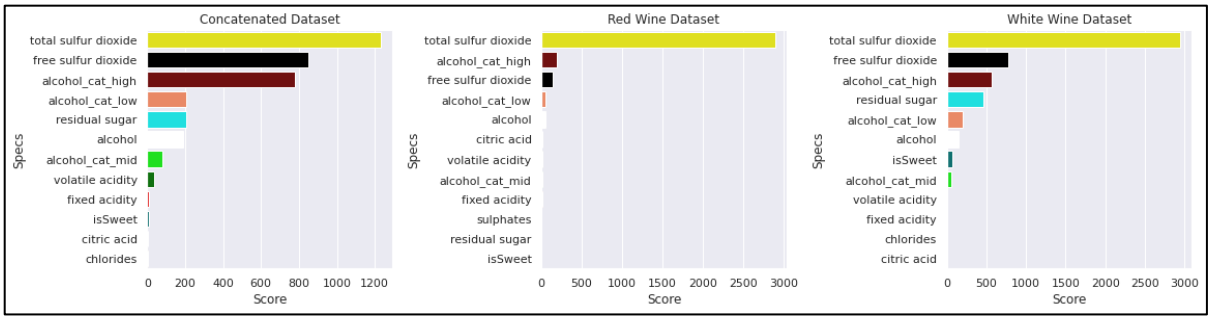

Figure 2.5

After applying SelectKBest, the resulting ‘k-scores’ were transformed into bar charts, ranking from highest to lowest as shown in Figure 2.5. The datasets show a pattern with the top three features all being total sulphur dioxide, free sulfur dioxide and alcohol_cat_high. This indicates the importance of these features to be included in the training datasets.

Using the analysis in part 2 of this document the training subsets are:

Concatenated Dataset	Red Wine Dataset	White Wine Dataset
▪ Alcohol ▪ Density ▪ Volatile Acidity ▪ Chlorides ▪ Alcohol_cat_low ▪ Alcohol_cat_mid ▪ Alcohol_cat_high ▪ Total Sulphur Dioxide ▪ Residual Sugar	▪ Alcohol ▪ Citric Acid ▪ Alcohol_cat_low ▪ Alcohol_cat_high ▪ Total Sulphur Dioxide ▪ Free Sulphur Dioxide ▪ Residual Sugar	▪ Alcohol ▪ Density ▪ Volatile Acidity ▪ Chlorides ▪ Alcohol_cat_low ▪ Alcohol_cat_mid ▪ Alcohol_cat_high ▪ Total Sulphur Dioxide

Table 1.1

These subsets were standardised using a MinMaxScaler, ensuring the feature values were between 0 and 1. This removes noise[5] when training the model to reduce overfitting.

3. Machine Learning Approaches and Evaluation

A mentioned in section 1, upsampling is used to balance the data for quality to improve the accuracy of the model. After upsampling and comparing results with an unbalanced dataset, it was regarded as successful due to improved model accuracy.

Using the established data subsets in table 1.1, three classification models and two regression models have been created.

The Classification Problem

The target quality was represented using 7 labels: integers 1 to 7. On the condition that a model is performing a multi-class classification, there is a trade-off between the number of labels and the accuracy of the model. A solution to this problem would be to reduce the number of these ‘labels’ by mapping new values to the ones in the dataset using thresholds to split the old values. The chosen mapping was binary classification.

The thresholds selected were tested and tweaked due to some thresholds resulting in overfitting the model. This overfitting occurred when a threshold disproportionately split the data. In the diagram below, red is low and blue is high.

Figure 3.1

Figure 3.2

The threshold decided was to rate quality ‘high’ if it greater than 6, otherwise rate as ‘low’ as seen in figure 3.2. Overfitting occurred when setting the threshold as 4, seen in figure 3.1. This overfitting would rate most wines as ‘high’ as there was more diverse data in the ‘high’ class to train the model.

Beginning with Logistic Regression, the hyperparameters[8] were tuned to ensure efficient usage of the dataset when training. This was done for each dataset as they required different tuned parameters. The red wine dataset performed best on this model, achieving an accuracy of 0.84 on the test set with a cross-validation score[9] (CVS) of 0.88. The CVS is the best measure of accuracy as it compares folds (subsets) of a dataset as a test set and calculates the average accuracy.

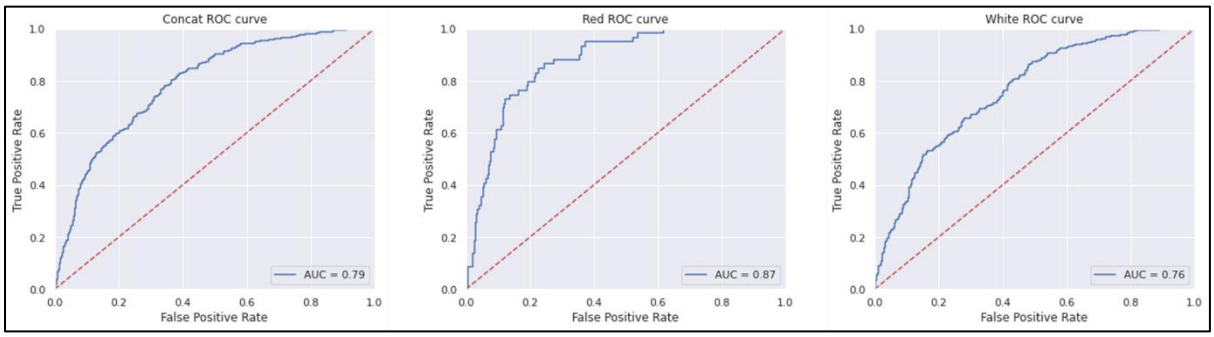

Figure 3.3

In figure 3.3, red wine stands out as the best performing with an AUC[7] of 0.87. This study assumes the red wine dataset performed better due to it having the least amount of features and having fewer records to train the model, reducing the susception of noise. (*Classification: ROC Curve and AUC*, 2020)

The next algorithm is Random Forests. Also hyperparameter tuned, the concatenated dataset performed the best on this model, achieving an accuracy of 0.87 on the test set with a CVS of 0.96. This was followed closely by the white wine dataset which also had a CVS of 0.96.

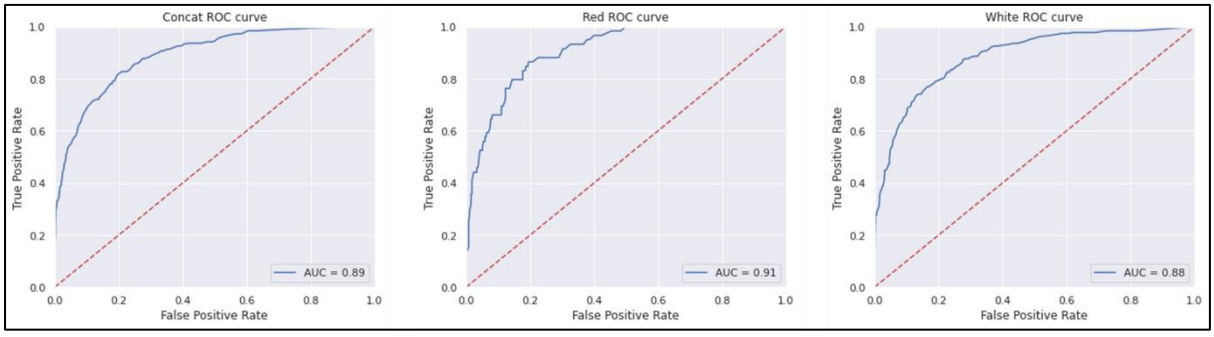

Figure 3.4

In figure 3.4, red wine showcases a higher AUC, however is jagged, suggesting there are inconsistencies in the true positive rate. The concatenated dataset has a similar AUC and follows a smoother curve, assuring a consistent true positive rate. This reinforces the use of the concatenated dataset with this model.

The final classification model used was K-Nearest Neighbours. Also hyperparameter tuned, this algorithm performed significantly well across all the datasets, each having a CVS of 0.96.

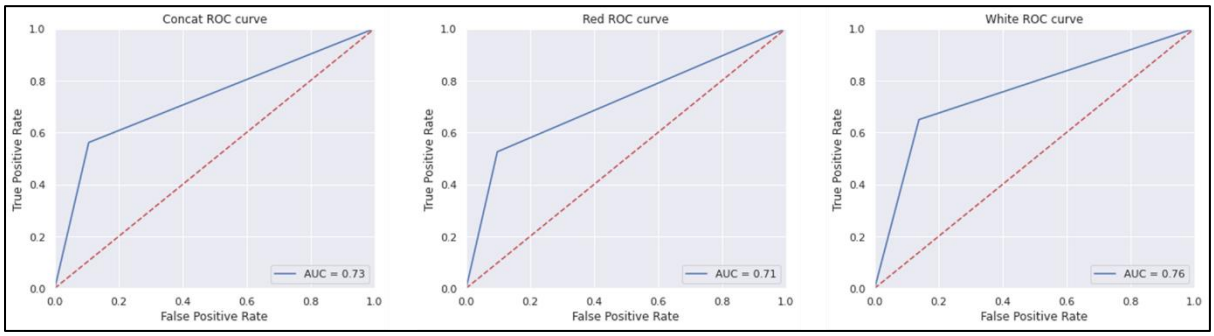

Figure 3.5

From figure 3.5, I can deduce the best performing dataset to be the white wine dataset with the highest AUC and an accuracy of 0.81 on the test set.

The Regression Problem

The problem that arises from a regression algorithm is the attempt to predict multiclass labels as they predict in a continuous format. This means when trying to predict a wine that is of quality 7, it may return 7.30 or 6.80. It is difficult to evaluate these models using traditional techniques. For this evaluation, the root mean squared error (RMSE), the mean absolute error (MAE) and scatter plots are used for evaluation. (Glen, 2021)

Figure 3.6

The first algorithm tested was Linear Regression. This algorithm did not perform as well as expected with the smallest RMSE on the test data being 0.91 and the MAE being 0.72. The closer to 1 the worse the accuracy is.

Figure 3.6 shows a positive correlation, indicating there is some prediction accuracy however a high spread of results for each quality level, reduced the confidence in using this model.

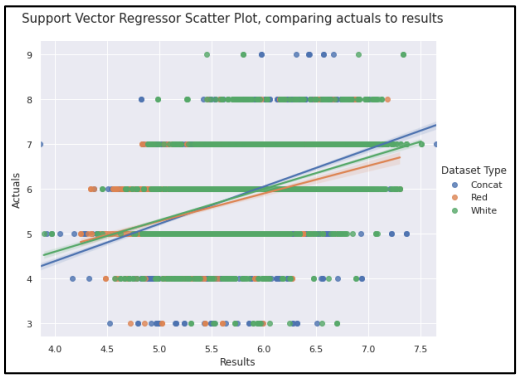

Figure 3.7

The final regression algorithm tested was a Support Vector Regressor. It performed slightly better than the Linear Regression algorithm with the smallest RMSE on the test data being 0.89 and the MEA being 0.70.

Figure 3.7 shows how the concatenated dataset performed moderately better than the other datasets. This is still insufficient evidence to approve the use of this model in quality prediction.

This work is important to the wider computing industry. It reflects the benefits of applying machine learning models to save time and increase profits for a wine company. Further applications of these models, could be 'the usefulness of a medicine' or the 'tire damage to a car' based on existing attributes of the provided object of interest. The applications are endless.

Conclusion

Part 1's analysis successfully revealed important features that would be included in the models and further highlighted some suggestions to make to the Winery regarding what determines the quality for 'Vinho Verde' wines. The correlation analysis indicated that high levels of alcohol insinuate a higher quality. Furthermore, feature importance results indicated that levels of total sulphur dioxide were important for high-quality wine and after consulting the correlation matrix, these levels should be low. After selecting features in part 2 and comparing each of the model's results with all the features and with the selected features there were significant improvements and these subsets were approved. The machine learning in part 3 was interesting to experiment with. 5 models were produced, 3 being very successful at predicting a binary classification of quality. A multiclass model was also built however the accuracy was not good enough to be used in practice, therefore binary classification was used instead.

The most successful model was Random Forests, trained using best using the concatenated dataset. It achieved a cross-validation score of 0.96 and in light of this evidence, it should accurately predict if a wine is of good or bad quality. The high cross-validation score suggests that there is very little overfitting and suggests the model should work well on different wine datasets for quality prediction. The least successful model was the Linear Regression, followed closely by the support vector regressor. The evidence suggests this was because regression algorithms fit results between bounds. For example, my results for quality would be between 3 and 9, however with regression algorithms they would never hit those bounds, only fill values between. This seriously affected the accuracy of the model.

In future using different subsets for classification and regression would result in stronger accuracy for regression. Furthermore, the imbalance in data for quality levels 3 and 9, predicting those labels was very difficult. Future datasets used should have a uniform distribution for the target class, even with upsampling, the results were for those classes were still inaccurate.

For me, the project went very well and the most useful experience I got was from it was to experiment with all the different machine learning algorithms that scikit-learn provides. Before this project, I had never experienced using machine learning, but it had always been a huge interest. It was the reason I came to University to study Computing. I have built significant skills in data analysis and manipulation which I know will be useful in stage 3 of my degree. In future, I would like to work on my experience using regression algorithms as they were unsuccessful in this project and I want to know why.

References

Brownlee, J. (2019) ‘How to Choose a Feature Selection Method For Machine Learning’, *Machine Learning Mastery*, 26 November. Available at: https://machinelearningmastery.com/feature-selection-with-real-and-categorical-data/ (Accessed: 11 May 2021).

Classification: ROC Curve and AUC (2020) *Google Developers*. Available at: https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc (Accessed: 12 May 2021).

Glen, S. (2021) *RMSE, Statistics How To*. Available at: https://www.statisticshowto.com/probability-and-statistics/regression-analysis/rmse-root-mean-square-error/ (Accessed: 13 May 2021).

Gupta, R. (2019) *An Introduction to Discretization in Data Science, Medium*. Available at: https://towardsdatascience.com/an-introduction-to-discretization-in-data-science-55ef8c9775a2 (Accessed: 4 May 2021).

kjytay (2019) ‘Spearman’s rho and Kendall’s tau’, *Statistical Odds & Ends*, 8 July. Available at: https://statisticaloddsandends.wordpress.com/2019/07/08/spearmans-rho-and-kendalls-tau/ (Accessed: 10 May 2021).

Sugar in Wine Chart (Calories and Carbs) (2015) *Wine Folly*. Available at: https://winefolly.com/deep-dive/sugar-in-wine-chart/ (Accessed: 5 May 2021).

Appendices

Appendix A: Glossary

1. **Feature**: A column in the dataset used for analysis.
2. **Target**: A feature that is to be predicted by the machine learning model.
3. **Model**: A machine learning object that can be trained using a dataset and predict a target.
4. **Concatenate**: To join two or more objects together to form a new ‘concatenated’ object.
5. **Noise**: Irrelevant data that can significantly affect the accuracy of a model.
6. **ROC**: A graph of the relationship between true positives rates and false-positive rates.
7. **AUC**: The area under the ROC, determining the area covered by the true-positive rate.
8. **Hyperparameter Tuning**: Tuning a machine learning model for optimum performance.
9. **Cross-Validation Score**: The average accuracy of a model after testing on many folds.
10. **Upsampling**: To create artificial records with values in the desolate feature labels. E.g. balancing quality to have more records in quality labels 3 and 9.

Appendix B: Figures

Appendix C: Tables