Advanced Analytics and Machine Learning Assignment 1

Terry McElroy

40126429

10th April 2020

Contents

Word Count: 3450

Introduction

The alcohol industry has a global revenue of 1.59 trillion dollars with wine being a quarter of those sales (Statista, 2020a). Unlike most of the other products in the alcohol sector the branding of wine is not as important when coming to selling the wine (Nowak et al. 2006). The emotion that the consumer experiences whilst consuming the product is more important thus making the quality of this experience and by extension the quality of the wine itself to be more important when selling wine (Nowak et al. 2006). This then brings around the problem with selling wine; What defines good quality wine? It is from this point that a wine expert is needed to taste the wine and give their opinion. However, with the wine industry producing 24.6 billion litres of wine in 2019 (Statista, 2020b) there is too much for one person to attribute their opinion to the quality of the wine. Furthermore, as mentioned previously wine is an emotive experience, this then leaves the scale to be biased towards one style or region of wine due to the testers own personal memories that have been associated with the wine. This presents an interesting machine learning problem of can opinionated sensory data match scientific data from physicochemical tests. The aim of this project is to use analytics to develop a machine learning model to determine the quality of a Portuguese wine. With this we can define the business problem to be;

> *Is the opinion of a wine taster able to be proven by scientific exploration of the wine's chemical properties? And can these models be of a sufficient accuracy to replace an expert taster saving the winery money in their production costs?*

Methodology

In this analysis several different techniques have been used to best determine the quality of the wine. This section will discuss what each method is and how it helps in determining the wine quality. It is important to note that the target variable is categorical data and that the techniques that have been used in this analysis are ones for classification problems. Since category is being predicted models will use the probability that it belongs to a certain category (James et al., 2013).

Logistic Regression

This is one of the more rudimental classification techniques but nonetheless still useful in modelling. It uses a logistic function to describe the probability that a series of data points belongs to a certain category (Equation 1), this can then be plotted (figure 1) and from there it is possible to determine which category the data series belongs to.

$$P(y) = \frac{1}{1 + e^{-(\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_n x_n}}$$

Equation 1.The equation that represents the general formula for logistic regression. B_0 is the intercept for the y axis, x are values from the data for different variables and $B_{1, 2, n}$ is the weight of how a certain variable affect the model.

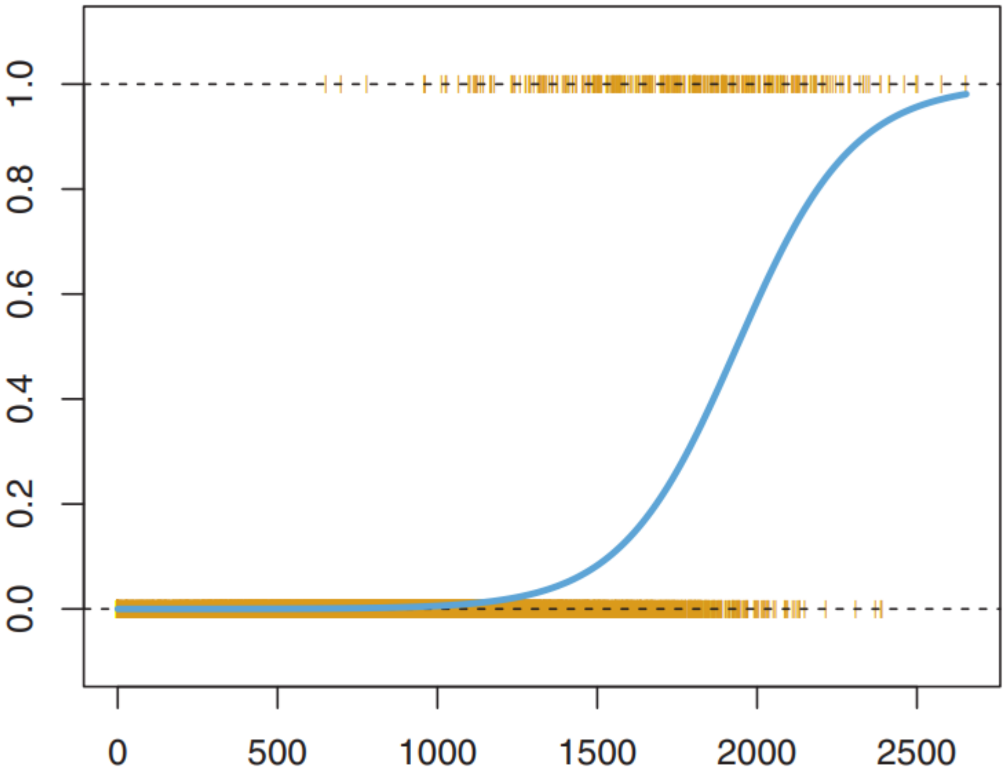


Figure 1.This shows how the function works. If the probability is greater than 0.5 then it will predict the outcome to be true (James et al., 2013)

The cut off figure can be edited to make the model better at predicting the outcome you want. However, the main downfall of this method is that it only accounts for a binary classifier. The data that is being used in this project is a multinomial classification. This method can be used to develop a very rudimental model to determine which wine is good or bad. This will be done by creating a centre point in the quality category and determining all wine below this mark is bad and all wine above is good.

Linear Discriminant Analysis

This method is very similar to the logistic regression method. This method estimates the distribution of responses with respect to each category. This is a parametric method as we then assume that this distribution is normal and then use Bayes theorem to establish the probability that will be associated with each class (James et al., 2013). The main objective is to segregate the data so that a certain value will account for a certain category (figure 2). It does this by maximising the distance between the means for each category whilst simultaneously reducing the spread for each category (equation 2). By doing so this will reduce the dimensionality of the problem and confine certain predictor values to a certain category.

$$\frac{(\mu_1 - \mu_2)^2}{(s_1^2 + s_2^2)} = \frac{\sum d_n^2}{\sum s_n^2}$$

Equation 2. This is the equation that satisfies the two criteria for LDA. μ is the average of each category, s is the scatter of the values around this category, d is the distance between the means when it is only a two category problem, when there is more than two classes then d is the distance to a point for each category.

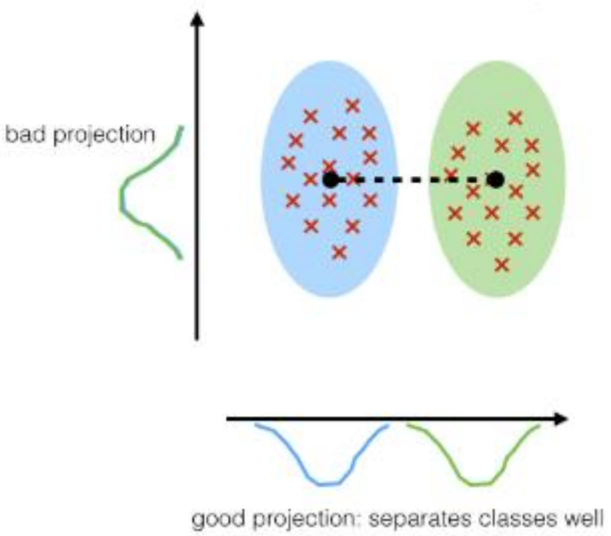


Figure 2. This show how LDA maximises the separation between the different classes (Raschka, 2014).

Linear discriminant analysis uses a linear discriminant function (equation 3) to calculate the category that it belongs too. To determine the best variables used for discriminating into categories the larger the coefficients maximises the distance between the means thereby making them good at predicting the class.

$$D = v_1X_1 + v_2X_2 + \cdots + v_iX_i + a$$

Equation 3. This is the discriminant function, D, that LDA uses, v is the coefficient, X is the variable value, and a is a constant

Tree Classification

Tree based methods work very similarly to human logical thinking. It will start with a root node that will choose a variable that best splits the data into either a true or false camp. After this it will quiz the data again and try to find how it can best split the data into the categories. This is recursively repeated until the data cannot be separated anymore. When there cannot be anymore it is called a leaf node, this node will hold which category the data corresponds too (Loh, 2011). The divisions in the data are then easily displayed with a tree (Figure 3).

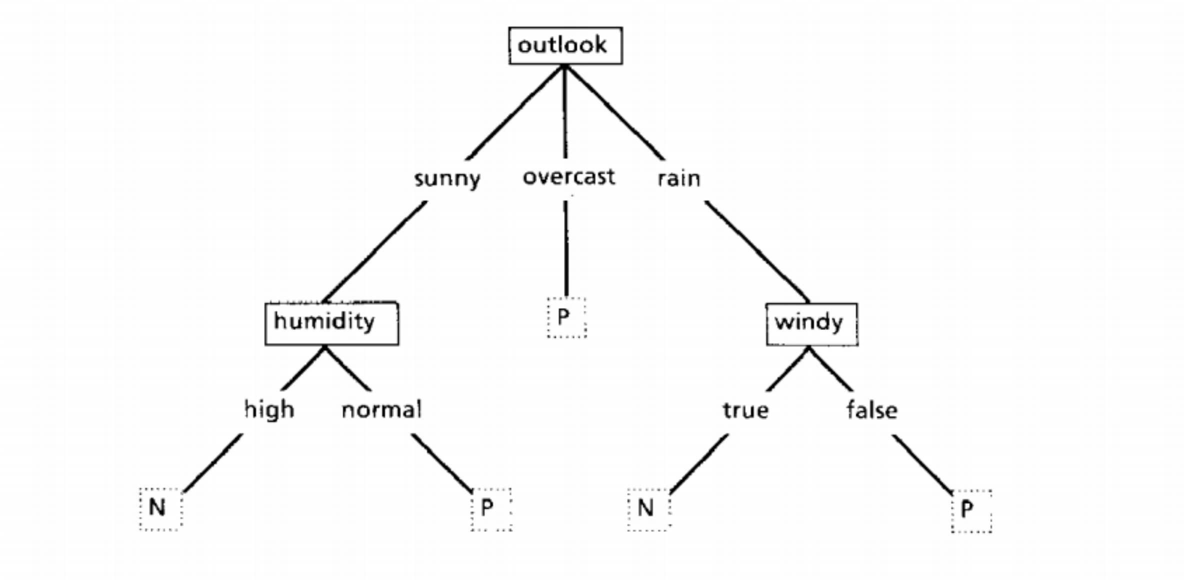
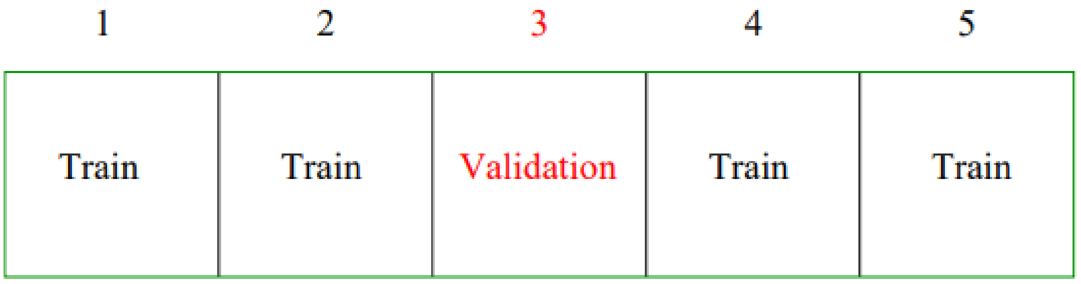


Figure 3. An example of a decision tree that is determining whether to play golf depending on what the weather is (Quinlan, 1986)

Cross Validation

Cross validation is a resampling method. By using this technique, the data that is currently had can be sampled several times over and be fed through a model. The reason for doing such an approach is to improve the model's predictability of the classes whilst not having to source new data to train and test the model. The method that will be used is K-fold cross validation. This method works by randomly arranging the dataset and then splitting the dataset into k groups. One of these groups will be used as the testing group. The other groups are then used to train the model by combining the groups through as if it were new data (figure 4). The test group is then used to evaluate the model accuracy. The test group is then changed to another group and the process will repeat until each group has been the testing group for the model. By doing this method it results in a less biased result that what would be achieved using a simple test/train split of the data (Hastie et al., 2009; James et al., 2013)



1	2	3	4	5
Train	Train	Validation	Train	Train

Figure 4. This is showing how a dataset has been split into five groups with group 3 being used to validate the model that will be developed by accumulated data from the other four groups (Hastie et al., 2009).

Data Quality Report

Before developing the models to predict the wine quality it is important to look at the data given to ensure that the data is sensible and that the outliers are dealt with in an efficient manner. This section will discuss the dataset characteristics and will use visualisations to show potential outliers.

Data Characteristics

The data provided is data that is composed of physiochemical and sensory data on 4898 white *Vinho Verde* from a region in Portugal. The objective of the analysis is to try and establish a connection between the eleven physiochemical variables and the one sensory variable. Table 1 below shows summary statistics for each of the physicochemical variables.

Variable	Minimum Value	Mean Value	Maximum Value
Fixed Acidity	3.80	6.855	14.2
Volatile Acidity	0.08	0.3342	1.10
Citric Acid	0.00	0.3342	1.66
Residual Sugar	0.60	6.391	65.80
Chlorides	0.009	0.0457	0.346
Free SO_2	2.00	35.31	289.00
Total SO_2	9.00	138.40	440.00
Density	0.987	0.994	1.04
pH	2.72	3.188	3.82
Sulphates	0.22	0.49	1.08
Alcohol	8.00	10.51	14.20

Table 1. Summary statistics of physicochemical variables

The quality variable has been defined in the data dictionary to range from 0 to 10, however, the data that has been presented does not have data for all wine categories. Category 0, 1, 2, and 10 are absent from the selected wines, therefore it will not be possible to predict wine that corresponds to these categories. Furthermore, of the wine that is present over 40% of the wine is attributed to the one category, 6. This will cause an imbalance amongst the classes and could make it tricky to correctly identify the wine that doesn't belong to this category.

Dealing with problem data and outliers

From table 1 it is possible to identify problem variables for the analysis, such as residual sugar where the maximum value is much greater than the mean. To efficiently identify all problem data and outlier's algorithm 1 is used to display each variable in a histogram. This function allows the comparison of the variable with and without the problem data. Algorithm 2 is then used to perform a clamp transformation on the data. A clamp transformation sets the upper and lower bounds that are acceptable for a list of data. If the data is smaller or greater than these threshold point, then the data will be rewritten to the accepted lower or upper threshold respectively. The reason this method is used is to prevent the outliers being set to missing making predictions more difficult. The upper and lower bounds were set to be ±3 x Standard Deviation, this is to ensure that only the extreme 0.27% of data is affected. Figure 5 below shows how these functions are used to visually identify the outliers in the data and table 2 is an updated summary of the variables after they have been tested for outliers.

```
compare <- function(x1,x2 = NULL, title = NULL){
  p1 <- hist(x1)
  p2 <- hist(x2)
  plot(p1, col = "orange", main = paste("Comparison of ",
  title), xlab = title)
  plot(p2, col = "green", add = T)
}
```

Algorithm 1. This compare function is used to plot two histograms on the one axis to compare how it looks with and without the outliers

```
clamped <- function(x){
  clamp(x, lower = (mean(x, na.rm = T) - 3*sd(x, na.rm = T)),
        upper = (mean(x, na.rm = T) + 3*sd(x, na.rm = T)))

}
```

Algorithm 2. This is the clamped transformation function to remove any outliers in the data

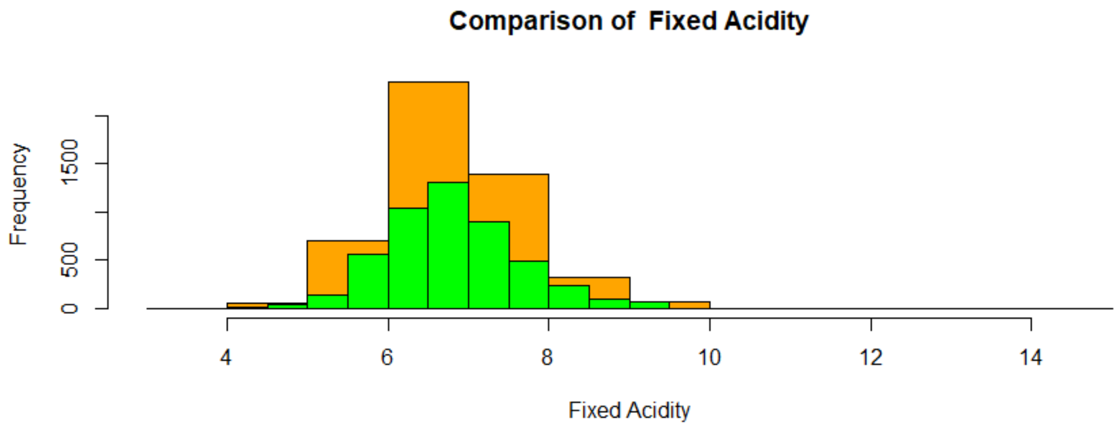


Figure 5. A comparison of the fixed acidity before (orange) and after (green) the clamp transformation. It is possible to know that there is an outlier here because R has developed the axis upto 14 with the rest of the data being 10 or less

Variable	Minimum Value	Mean Value	Maximum Value	Change?
Fixed Acidity	4.323	6.851	9.356	Yes
Volatile Acidity	0.08	0.2766	0.580	Yes
Citric Acid	0.00	0.3327	0.6973	Yes
Residual Sugar	0.60	6.376	21.608	Yes
Chlorides	0.009	0.04465	0.11132	Yes
Free SO_2	2.00	35.15	86.33	Yes
Total SO_2	10.87	138.25	265.85	Yes
Density	0.987	0.994	1.003	Yes
pH	2.735	3.188	3.641	Yes
Sulphates	0.22	0.49	1.08	No
Alcohol	8.00	10.51	14.20	No

Table 2. This is the updated summary statistics with the outliers removed

Results

With the business problem to establish if these chemical tests can account for quality and thereby extension its price, the analysis was split up into two major sections. The first section was to predict if a wine was defined as good or bad, and the second was to establish if it is possible to create three sub categories of quality and how each subsection can differs to the other in so doing allowing a price structure to be developed for the wine.

Good wine vs bad wine

Before beginning it is important to define what I am defining as good or bad wine. From the data quality is marked on a 0 to 10 scale, in this section of the analysis I am defining good wine to be of quality 6 or greater with the remaining defaulting to bad wine. By doing this it leaves the quality variable to be a binary decision.

Logistic Regression

As wine quality has been reduced into a good and bad category it is possible to preform logistic regression. Before this is carried out it is important to test whether the variable is significant to the target variable. This is done by running a logistic regression model that includes all variables. If the p-values are greater than 0.01 then the variable is not significant and will be removed from the model, the variables that have been kept will then be tested to check if they work well in predicting quality. The results of this approach are shown in table 3;

Variable	Model 1 p-value	Keep?	Model 2 p-value	Keep?
Fixed Acidity	0.0786	No		
Volatile Acidity	<2x10^{-16}	Yes	<2x10^{-16}	Yes
Citric Acid	0.6860	No		
Residual Sugar	7.91x10^{-14}	Yes	<2x10^{-16}	Yes
Chlorides	0.8961	No		
Free SO_2	3.15x10^{-5}	Yes	5.88x10^{-7}	Yes
Total SO_2	0.4272	No		
Density	1.15 x10^{-7}	No*		
pH	8.27 x10^{-5}	Yes	0.04239	No
Sulphates	7.52 x10^{-8}	Yes	0.00055	Yes
Alcohol	6.47 x10^{-10}	Yes	<2x10^{-16}	Yes

*Table 3. This is the result of testing which variables are significant in predicting quality, * symbolises that this variable was removed because it was highly correlated with other variables in the model*

After this has been done it is now possible to build a model based on these variables giving the formula; that quality will be modelled with respect to Volatile acidity, Residual sugar, Free Sulphur dioxide, Sulphates, and Alcohol content. Once the model has been built it is now possible to use it for predictability. To do this the data should be partioned into a training and testing datasets (75% - 25%). This will allow the model to developed under the training and the testing to check how accurate it is. From the formula, deduced from above, the accuracy was 78.6% with the cut off probability at 0.46. This is quite an accurate model, from this model the variables that are key to good wine are Sulphate, Alcohol content, Residual Sugar, and free sulphur dioxide this leaves Volatile acidity as an indicator of poor-quality wine.

Linear Discriminant Analysis

The data used in this section will be the same as that in the logistic regression. The main difference of LDA is that the p-values are not of concern so all variable can be used to predict the outcome. Doing so will give an accuracy of 76.54% in predicting the correct class for the wine. Table 4, below, displays the coefficients and gives an insight as to which variables are responsible for the classification of the wine. The variables that indicate good wine are the ones with positive coefficients and the ones that are negative are indicating poor wine quality.

Variable	Coefficients
Fixed Acidity	0.217
Volatile Acidity	-5.95
Citric Acid	-0.27
Residual Sugar	.176
Chlorides	.426
Free SO_2	.0106
Total SO_2	$2.72x10^{-4}$
Density	-353.03
pH	1.88
Sulphates	1.707
Alcohol	0.454

Table 4. This shows the coefficients that affect the class of the wine

Modelling wine quality on a 3-part scale

This section is to determine what the quality of wine on three levels of low, mid, and high-quality wine, the definition of the quality is classified in table 5. To do this analysis I will start with determining whether the wine belongs to mid or other and then wine that is identified as either low or high quality will be examined. After this I have attempted to split find the exact quality for the mid-range wines.

New Quality	Old Quality
Low	3,4
Mid	5,6,7
High	8,9

Table 5. This is the new qualities that will be the target of the models in this section

Mid-range vs Other wine qualities

To make this model the qualities of low and high were merged into one category called other. This is now a binary problem; it is possible to use logistic regression. This time it is not necessary to remove any variables from the model as most of the wine will be in the mid-range category and we are trying to determine what qualities make that happen. Using the same partitioning method as before to test the model we get an accuracy of 91.9% with the variables most responsible for this classification in table 6. From this we can see that the mid quality wine is predicted based on citric acid, sulphates, Residual sugar, and sulphur dioxide content. Since the accuracy is so high and the data is mostly made up of mid-range wines there is no need to retest with a different method.

Variable	Coefficient
Intercept*	5.60E+52
Citric Acid	2.14E+00
Sulphates	1.99E+00
Residual Sugar	1.06E+00
Total SO_2	1.00E+00
Free SO_2	1.00E+00
Fixed Acidity	8.82E-01
Alcohol	7.12E-01
pH	3.54E-01
Chlorides	6.21E-02
Volatile Acidity	1.90E-02
Density	2.47E-49

*Table 6. This is an ordered table for which variables are most significant in the wine being mid-range, *intercept is part of the output of the model and not a new variable*

High vs Low quality wine

Since the majority of the wine is mid quality it is important to identify wine that is to either extreme. To do this I am testing what is the difference between high quality and low-quality wine. To do this I will be using the LDA and logistic regression to model this split. The LDA model can now be developed with all variables in it and is tested with data partitioning this then gives a model accuracy of 84.6% with table 7 which variables are the most discriminant.

Variable	Coefficient
Chlorides	5.50E+00
pH	2.86E+00
Sulphates	1.12E+00
Fixed Acidity	4.31E-01
Residual Sugar	3.17E-01
Alcohol	1.48E-01
Free SO_2	2.20E-02
Total SO_2	-2.11E-03
Citric Acid	-5.57E-01
Volatile Acidity	-3.53E+00
Density	-5.98E+02

Table 7. LDA Coefficients for low vs high quality wine

This model is then compared to a logistic regression model with the same formula as in LDA. This resulted in a dramatic boost in the accuracy of the model with it now being at 90.1%. This model is therefore chosen over the LDA model. When determining if a wine is high or low quality, we can determine through two logistic regression models that will give an overall accuracy of 82.83%.

Trying to Determine the mid-range qualities

This section is focusing on the three qualities that make up the mid-range quality. Since this is more than a binary output logistic regression cannot be used. The methods that will be used are LDA, QDA, tree, and cross validation. Once the data has been formatted to only have mid-range qualities, LDA, QDA and tree methods were used to try and conclude what determined these classes. The accuracy of each model is shown in table 8. The accuracy, of the three methods mentioned, is poor compared to other models. Since, LDA has the highest accuracy a cross validation approach was used to try to improve the model accuracy. This slightly improved the model, to further improve the model the formula was changed for the LDA method to the formula used to determine good or bad from earlier. With this new formula and a cross validation approach the accuracy of the model rose again to 68%. Therefore, to determine where a wine ranks in mid quality two models are used. The first model checks if the wine would be in mid class range and then a cross validated model is used to determine where it is in the this mid class range, this gives an overall accuracy to be 62.8%.

Method	Accuracy
LDA	58.5%
QDA	57.1%
Tree	57.2%
LDA CV	66%
LDA CV new formula	68%

Table 8. The accuracy of the models for mid-range qualities

Conclusion

The task of this analysis was to determine if it would be possible to replace the opinionated sensory data that is used to grade wines with a machine learning program that is able to use physicochemical data to determine the quality. From the analysis we can make general statements about the wine quality, we can determine whether the wine will be good or bad, we are also able to determine if the wine is of high quality or low quality. However, the sensory data was split into 11 groups and trying to model these individual levels of quality proved to be challenging even when determine between the three most populated classes. With this knowledge, the question posed in the introduction, *can these models be sufficiently accurate to replace an expert tester,* the answer would have to be no. The models are only able to make sweeping statements about the wine and not be able to pick up on the nuance of the different levels. Furthermore, the quality of wine is specific to this variety of the Portuguese wine itself, so if this model was applied to data of Argentine Malbec it is unlikely to classify them accurately. To further this study, it wold be interesting to get more wine to have the classes more balanced so it could potentially be more accurate in predicting different wine qualities and could allow for analysis between Low, Mid, and High-quality wines.

References

Hastie, T., Tibshirani, R. and Friedman, J. (2009) *The Elements of Statistical Learning Data Mining, Inference, and Prediction*, 2nd edn., New York: Springer.

James, G., Witten, D., Hastie,T. and Tibshirani, R. (2013) *An Introduction to Statistical Learning with Applications in R*, New York: Springer.

Loh, W.Y., 2011. Classification and regression trees. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *1*(1), pp.14-23.

Nowak, L., Thach, L. and Olsen, J.E., 2006. Wowing the millennials: creating brand equity in the wine industry. *Journal of Product & Brand Management*.

Quinlan, J.R., 1986. Induction of decision trees. *Machine learning*, *1*(1), pp.81-106.

Raschka, S. (2014) *Linear Discriminant Analysis – Bit by Bit,* Available at: *https://sebastianraschka.com/Articles/2014_python_lda.html#%E2%80%93-bit-by-bit* (Accessed: 8th April 2020).

Statista (2020a) *Alcoholic Drinks: Worldwide,* Available at: *https://www.statista.com/outlook/10000000/100/alcoholic-drinks/worldwide* (Accessed: 8th April 2020).

Statista (2020b) *Wine: Worldwide,* Available at: *https://www.statista.com/outlook/10030000/100/wine/worldwide* (Accessed: 8th April 2020).