Predicting Wine quality via it's Chemical Properties

Samuel Brege (A15476506) and Mani Amani (A15401744)

November 2020, ECE 196 FA 2020

1 Introduction

1.1 Motivation

Wine testing and ranking is an old tradition that still continues strongly to this day. However, with advances in technology comes advances in skepticism of the old. There are many strong arguments implicating that a lot of what Expert wine tasters experience is merely Placebo effect [4]. This raises a Question: is what the Tasters taste in their Heads? Or in their Wine?

1.2 Problem Statement

Using a data set from the UCI Machine Learning repository [2], we hope to determine whether there is a relation between a Wine's chemical properties and it's perceived quality by a wine taster. We plan to evaluate this relationship through the use of ML Classification techniques. The more accurately our model predicts a wine's quality using its chemical properties, the more likely a wine's quality is based off of said properties. However, should we be unable to predict quality at all, or very poorly, that would imply that the quality is more Placebo based.

1.3 The Data Set

M↓

```
whiteData.head()
```

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
0	7.0	0.27	0.36	20.7	0.045	45.0	170.0	1.0010	3.00	0.45	8.8	6
1	6.3	0.30	0.34	1.6	0.049	14.0	132.0	0.9940	3.30	0.49	9.5	6
2	8.1	0.28	0.40	6.9	0.050	30.0	97.0	0.9951	3.26	0.44	10.1	6
3	7.2	0.23	0.32	8.5	0.058	47.0	186.0	0.9956	3.19	0.40	9.9	6
4	7.2	0.23	0.32	8.5	0.058	47.0	186.0	0.9956	3.19	0.40	9.9	6

M↓

```
redData.head()
```

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
0	7.4	0.70	0.00	1.9	0.076	11.0	34.0	0.9978	3.51	0.56	9.4	5
1	7.8	0.88	0.00	2.6	0.098	25.0	67.0	0.9968	3.20	0.68	9.8	5
2	7.8	0.76	0.04	2.3	0.092	15.0	54.0	0.9970	3.26	0.65	9.8	5
3	11.2	0.28	0.56	1.9	0.075	17.0	60.0	0.9980	3.16	0.58	9.8	6
4	7.4	0.70	0.00	1.9	0.076	11.0	34.0	0.9978	3.51	0.56	9.4	5

Figure 1: Preview of Wine data

The Data set [2] is of White and Red Wines from Portugal, with 4898 White Wines and 1599 Red Wines recorded. As shown by Figure 1, each wine has 11 attributes and is assigned a quality between 3 and 9, wines with higher scores being better. These assigned qualities are somewhat normally distributed (See Figure 2). The attributes are the wines' chemical properties.

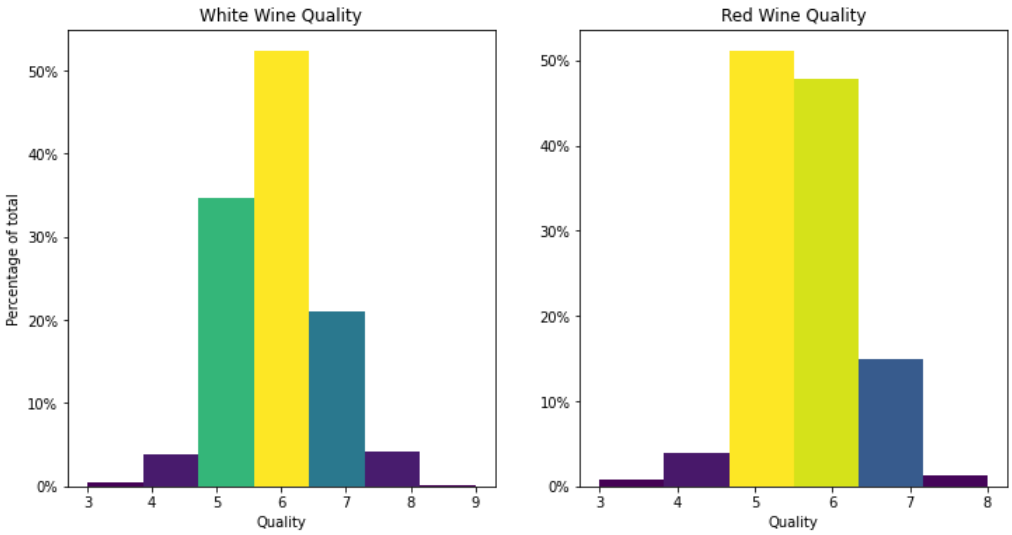

Figure 2: Quality Distributions

2 Finding a Model

We tested out three different classifying techniques: k-Nearest Neighbor (kNN), Random Forest Classifier (RFC), and Support Vector Machine (SVM). 70% of our data was used for training, with the remaining 30% being used for testing. As there is more data on White Wine than Red Wine, we expect a slightly higher accuracy with the Whites. We felt this data was not so suitable for regression and decided to stick to the classifiers.

2.1 k-Nearest Neighbor

kNN is the simplest of the classification algorithms we are using. It attempts to cluster the data by comparing the 'distance' of its attributes from one another. In the case of this data set, it would attempt to cluster the chemical properties around the quality.

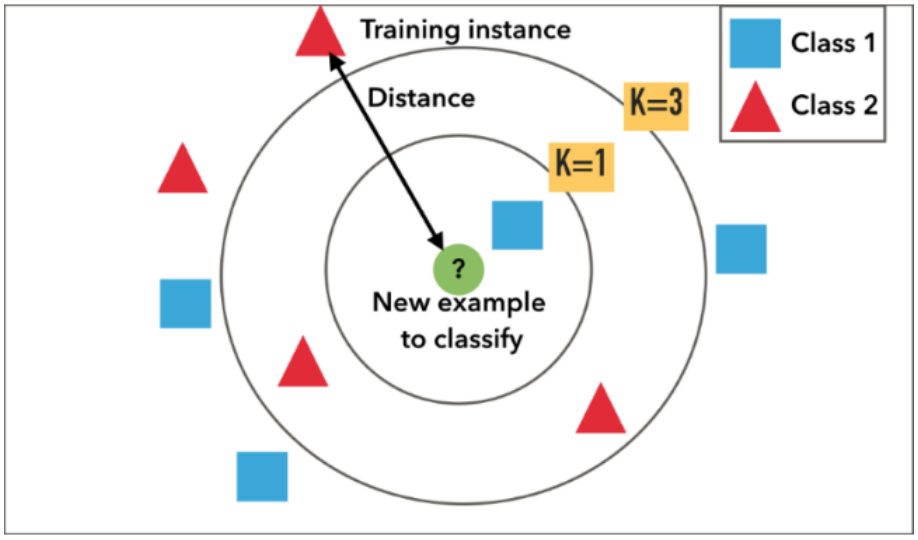

Figure 3: How kNN works [1]

Supplying our kNN classifier with all of the wine attributes, we were able to predict White Wine quality with a 55.92% accuracy and Red Wine quality with a 52.5% accuracy

2.2 Random Forest Classifier

RFC is also somewhat simple, utilizing a very large number of decision trees. This makes it suitable for High Dimensional datasets like this and should be much faster than SVM and kNN.

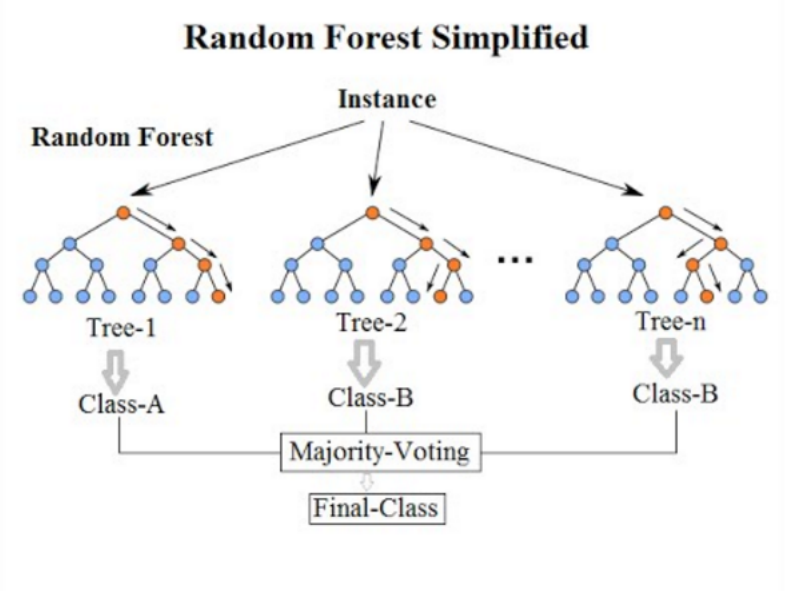

Figure 4: How RFC works [5]

Once again giving our classifier all of the wine attributes, we were able to predict White Wine quality with a 69.32% accuracy and Red Wine quality with a 65.83% accuracy

2.3 Support Vector Machine

An SVM classifies data by drawing Hyperlanes, which separate the data. This can be done both linear and non-linear.

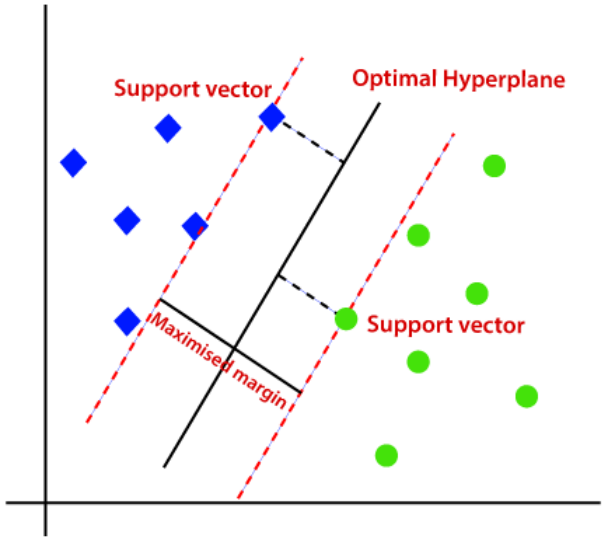

Figure 5: How Linear SVM works [3]

The linear SVM was able to predict White Wine quality with a 52.04% accuracy and Red Wine quality with a 50.42% accuracy. Non-linear SVMs resulted in significantly less accuracy.

3 Looking at the results

The model with the greatest accuracy was the Random Forest Classifier (69.32%/65.83%), followed by k-Nearest Neighbor (55.92%/52.4%), then SVM (52.04%/50.42%). k-NN and RF classification took a similar amount of time to complete, whilst the SVM took significantly longer.

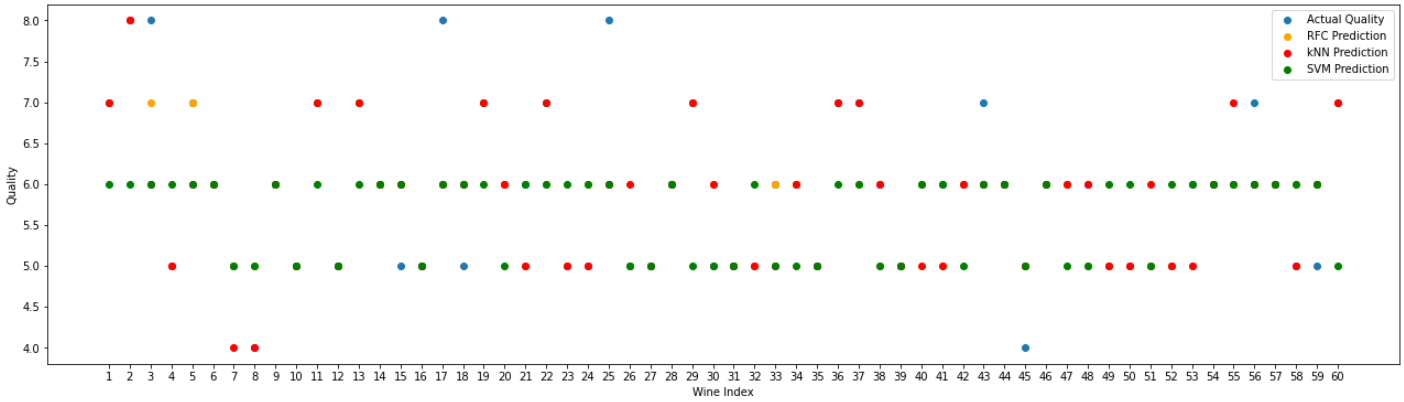

Figure 6: Scatter plot of first 60 White Wines

Figure 6 gives us a comparison between the predictions for each model as well as the data itself. The plot shows why SVM fared so poorly; we were using it to perform Binary classification, but there were much more than just two qualities it could predict. When the model was told to predict whether the quality was less than or equal to Five, or greater than Five (Making it a binary classification), it had an accuracy of 87.5%. This was still less than the RFC in the same scenario, however, which had a 92.1% accuracy.

The Random Forest Classifier performed much better than the k-NN because the data is high-dimensional, and RFC works particularly well with Categorical data like this data set. Interestingly, however, the k-NN model produced an incredibly accurate distribution, even more so than the RFC (See Figures 7 and 8). We can infer from this that the kNN algorithm puts a lot of weight into matching the distribution, whereas the RFC puts less importance on it.

We then trained and tested the Random Forest Classifier with each attribute individually (See Figures 9 and 10). The point was to determine which individual chemical property gave the best accuracy, and which one gave the worst accuracy. Alcohol appeared as the best predicting factor for our data set and sulphates appeared to be the worse. In red wine, chlorides were a significantly worse predictor compared to white wine. In white wine, pH appeared to be a worse predictor compared to red wine.

We will be using the Random Forest Classifier as our final model because it has the highest accuracy. It is a fairly simple solution, but the data is fairly simple. The only difficulty is the high dimensions of the data but that is something the RFC excels at and unlike the other models, it worked very well with no tweaking at all. It was also the fastest method, taking no noticeable time at all, especially in comparison to the SVM which would take upwards of two minute to train. The data was in the form of two csv files, had no null values and was in generally great shape. Overall, this meant little preprocessing was necessary.

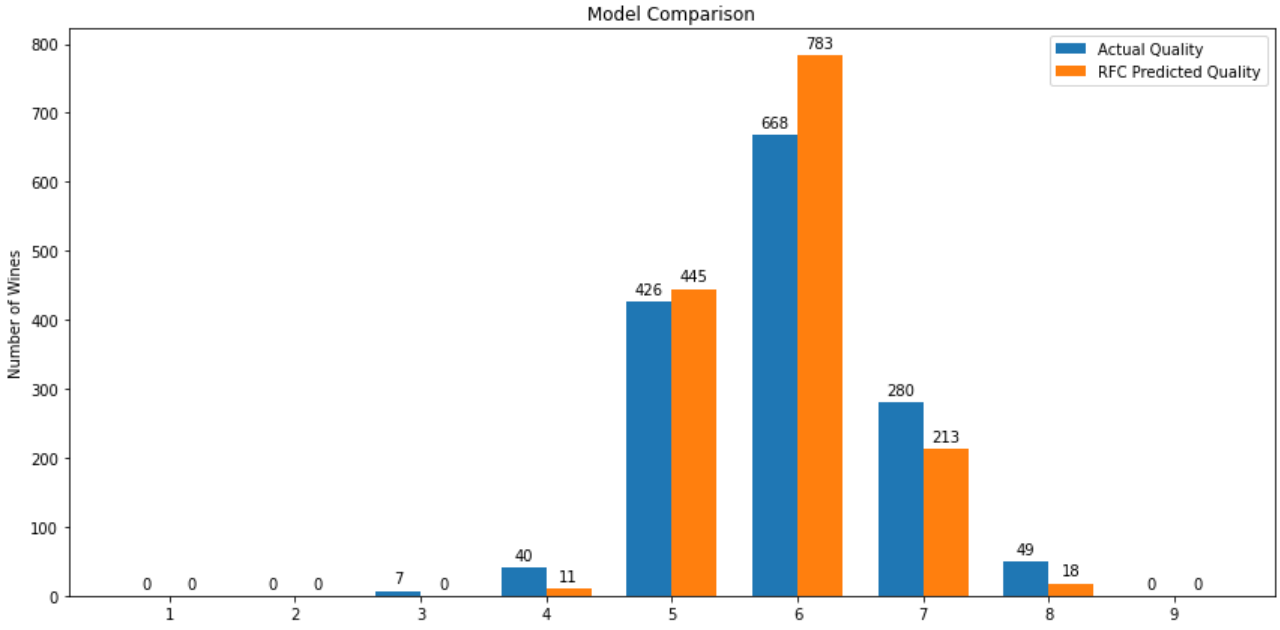

Figure 7: Actual Distribution vs RFC Distribution

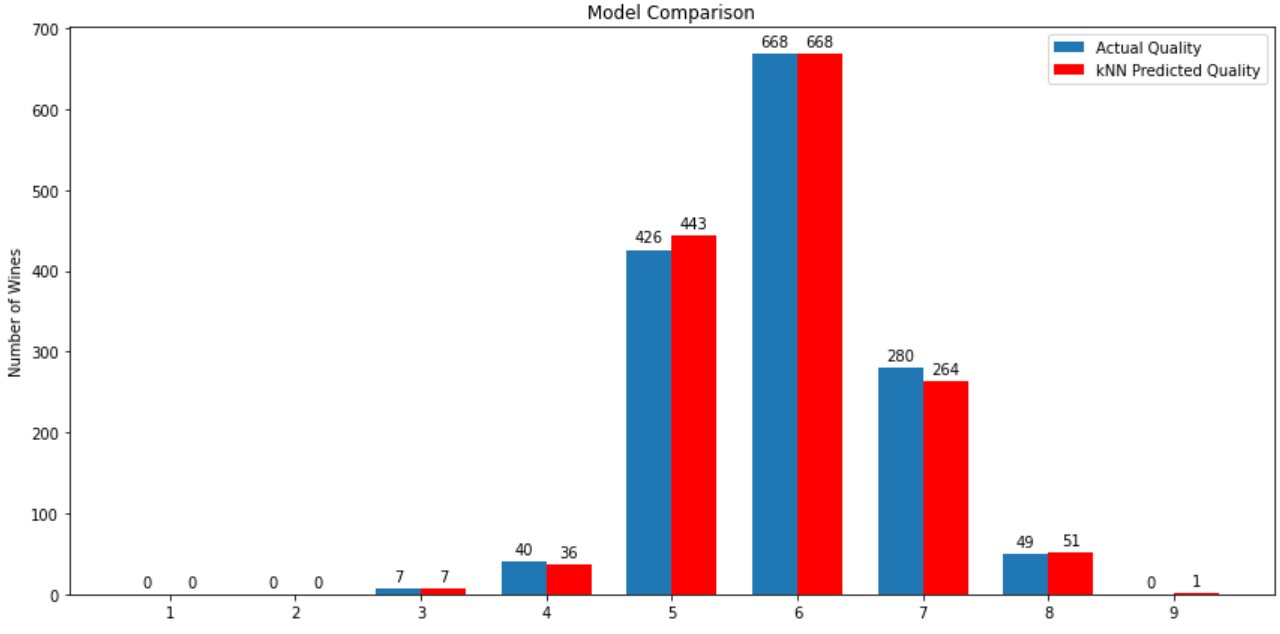

Figure 8: Actual Distribution vs kNN Distribution

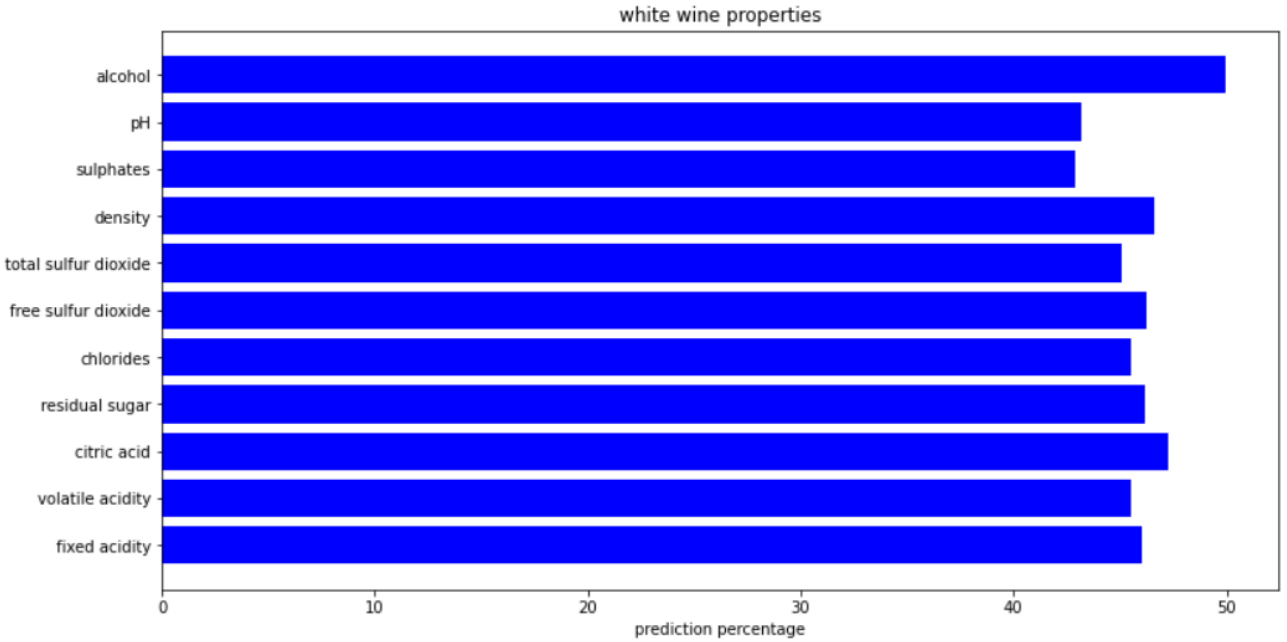

Figure 9: Random Forest Classifier accuracy for White Wine based on individual attributes

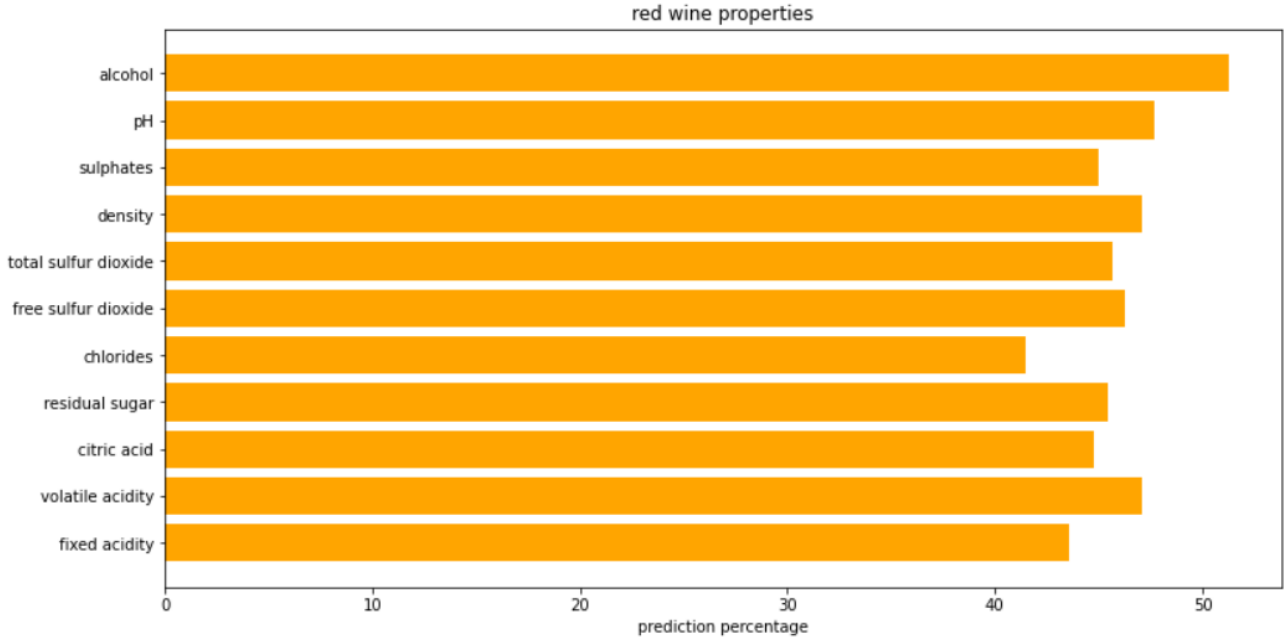

Figure 10: Random Forest Classifier accuracy for Red Wine based on individual attributes

4 Conclusion

4.1 Answering the Question

For there to be no correlation between a wine's perceived quality and its chemical property (ie. entirely placebo effect), we would expect our most accurate model to have around 14.29% accuracy; the probability of picking a quality at random and being correct (1/7).

As our most accurate model was able to predict quality almost 70% of the time, we can conclude that Wine Quality is strongly affected by its chemical properties. While this does not mean there is no Placebo effect with wine, this evidence suggests it could be a relatively minor factor in determining wine quality.

The attribute comparison data also had some interesting results (Figures 9 and 10). Alcohol content was a significantly better predictor of wine quality than any of the other attributes. We can infer from this that of all the chemical properties of wine, having the correct Alcohol content is most important. This is a reasonable assumption, because alcohol has a very strong flavor and would be the first thing a taster would use to categorize a wine.

4.2 How to improve

With more knowledge on the topic, we could better pick chemical properties that could effect quality, which would probably improve our mode. Attempting linear regression with the alcohol content could have some interesting results as well. We could also make better use of SVM by training an SVM for each possible quality.

Whilst the k-NN classifier was less accurate at actually predicting quality than the Random Forest Classifier, it was remarkably good at predicting the Distribution. Perhaps there is a way to feed this to the RFC and potentially improve our final model.

If we acquire more precise wine quality data (e.g 6.4, 7.2) we could see how much better our algorithm predicts the quality and whether the accuracy lies in a more acceptable range compared to what we have at the moment.

4.3 Final Notes

Our Github contains both the csv's and the .ipynb containing our Machine Learning code. Be forewarned, it is not particularly neat or professional.

URL = `https://github.com/sambrege/ECE196ProjectWINE`

References

[1] Apr 2017 Adi Bronshtein. *A Quick Introduction to K-Nearest Neighbors Algorithm*.
URL = https://blog.usejournal.com/a-quick-introduction-to-k-nearest-neighbors-algorithm-62214cea29c7.

[2] 2009 Cortez et al. *Wine Quality Data set*.
URL = https://archive.ics.uci.edu/ml/datasets/wine+quality.

[3] javatpoint.com. *Support Vector Machine Algorithm*.
URL = https://www.javatpoint.com/machine-learning-support-vector-machine-algorithm.

[4] 2017 University of Bonn. *Why expensive wine appears to taste better*.
URL = https://www.sciencedaily.com/releases/2017/08/170814092949.htm.

[5] Dec 2017 Will Koehrsen. *Random Forest Simple Explanation*.
URL = https://medium.com/@williamkoehrsen/random-forest-simple-explanation-377895a60d2d.