Wine Quality Classification from Physicochemical Properties

Introduction

In recent decades, wine has become a much more affordable commodity. To support the growth of the wine industry, new quality assessments techniques must be developed. Currently, evaluations of wine quality are determined heavily by expert evaluations; however, these evaluations are prone to many subjective factors. The development of objective analytic methods to assess wine quality is extremely challenging, because the relationship between physicochemical and sensory analyses are complex and not yet fully understood. The use of machine learning techniques on a large dataset could shed light on the nature of the relationship between physicochemical and sensory analyses. The goal of this analysis is to predict wine quality based on physicochemical data without heavy reliance on the volatility of wine tasters.

Within the wine industry, certification and quality assessment are conducted via physicochemical and sensory tests. Meanwhile, some of these sensory tests such as taste preference are performed by wine experts. Several studies have found that physicochemical properties are also useful in predicting human wine preferences, suggesting that the long-standing reliance on wine experts to perform these sensory tests is not entirely necessary. In a November 2009 study, Paulo Cortez and colleagues used support vector machine (SVM), multiple regression, and neural networks analyses to predict human wine taste preferences. After comparing the accuracies of these three methods, the authors conclude that the SVM method outperformed both the multiple regression and neural network analyses (Cortez *et al.,* 2009).

Two studies published in the years following the 2009 study by Cortez and colleagues utilized the same dataset to investigate similar topics related to wine quality classification. The first, a 2012 study by P. Appalasamy and colleagues, compared the results of two different algorithms: the decision tree-based ID3 and Naive Bayesian. Based on the accuracy and processing times for the two models, the authors concluded that the tree-based ID3 was the better of the two classification models. However, the authors also concluded that because of the relatively high misclassification rates in both of these models, neither were successful in modeling something as complex as the human taste (Appalasamy *et al.,* 2012).

The second, a 2013 study by A. Nachev and M. Hogan, used four data mining techniques for their wine quality prediction assessment: multilayer perceptrons, cascade-correlation neural

networks, general regression neural networks, and support vector machines (SVM). The authors of this study compared the predictive abilities of their models by looking not only at prediction accuracy, but at mean absolute deviation and area over the the regression error characteristic curve. The authors found that SVM with polynomial kernel outperformed all other models, and that the classic neural network method had the highest misclassification rate (Nachev *et al.,* 2013).

In order to re-examine and build upon previous analyses of wine quality classification, this study will compare the predictive power of six different techniques: Logistic Regression (GLM), Support Vector Machines (SVM), Support Vector Machines with Radial Basis Function (SVM-RBF), Neural Networks (NNet), Neural Networks Average (AvgNN), and Random Forest (RF). These different classification methods will aid in the prediction of wine quality based solely on physicochemical data and therefore conquer the ability to understand human taste.

Data and Preprocessing

Data: In their 2009 study on wine quality classification, Paulo Cortez and his colleagues created two datasets related to red and white variants of Portuguese "Vinho Verde" wine that will be utilized in this analysis. The red wine dataset contains 1,599 instances and the white wine dataset contains 4,898 instances. There are 11 input attributes and one output attribute. The input variables include objective tests: fixed acidity (1), volatile acidity (2), citric acid (3), residual sugar (4), chlorides (5), free sulfur dioxide (6), total sulfur dioxide (7), density (8), pH (9), sulphates (10), and alcohol (11). The output is based on sensory data: quality (12). The output is represented by a score between 0 (very bad) and 10 (very excellent), and is determined by taking the median of at least three evaluations by wine experts. Due to privacy and logistical issues, only physicochemical (inputs) and sensory (output) variables are included in this data.

Table 1: Physicochemical Data Statistics for Red and White Wine

Attributes	Red Wine			White Wine		
	Min	Max	Mean	Min	Max	Mean
Fixed Acidity	4.600	15.900	8.320	3.800	14.200	6.855
Volatile Acidity	0.120	1.580	0.538	0.080	1.100	0.278
Citric Acid	0.000	1.000	0.271	0.000	1.660	0.334
Residual Sugar	0.900	15.500	2.539	0.600	65.800	6.391
Chlorides	0.0120	0.611	0.085	0.009	0.346	0.046
Free Sulfur Dioxide	1.000	72.000	15.870	2.000	289.000	35.310

Total Sulfur Dioxide	6.000	289.000	46.470	9.000	440.000	138.400
Density	0.990	1.004	0.997	0.987	1.039	0.994
pH	2.740	4.010	3.311	2.720	3.820	3.188
Sulphates	0.330	2.000	0.658	0.220	1.080	0.490
Alcohol	8.400	14.900	10.420	8.000	14.200	10.510
Quality	3.000	8.000	5.636	3.000	9.000	5.878

Preprocessing: As noted in Cortez's study, red and white wine tastes differ significantly; therefore, this study will analyze the two distinct datasets separately (Cortez *et al.,* 2009). For the purpose of our analysis, the response variable, quality, was redefined as a factor in order to create a binary classification problem. Rather than a quantitative rating from 1 to 10, scores greater than or equal to 6 are classified as "good" and scores less than 6 are classified as "bad." We split our dataset into two sets: training and testing. 70% accounts for the training data and the remaining 30% accounts for the testing data. It is important to note that not using a validating set puts our analysis at risk of being overly optimistic. However, we will only use training and testing as the dataset is not particularly large.

The variables in this dataset have amplitudes of significant size. In addition, these physicochemical properties are recorded using several different units of measure, making them hard to compare. It is therefore inefficient to use the data in its original form. For the purposes of this analysis, the data is scaled to ensure that no one variables is more influential than the others. One potential scaling method is a linear transformation, in which all input variables are divided by the dataset maximum; however, using this method would mean that a large portion of the data would fall very close to zero. As a result, our models would surely perform poorly. Rather, the data is scaled individually for each input variable. In this way, the mean value for each column will be 0 and the mean standard deviation will be 1.

Recursive feature eliminations were performed on both the white wine and red wine datasets in order to identify how many input variables should be examined in order for our classification models to achieve the highest possible testing accuracy. After analyzing the backwards elimination plots for the two datasets in Figure 1, it is clear that both red and white wine classification models perform better when considering all of the potential input variables. These results corroborate Cortez's finding that most of the physicochemical inputs are relevant to the classification of wine quality. As a result, all 11 input variables will be utilized in this analysis.

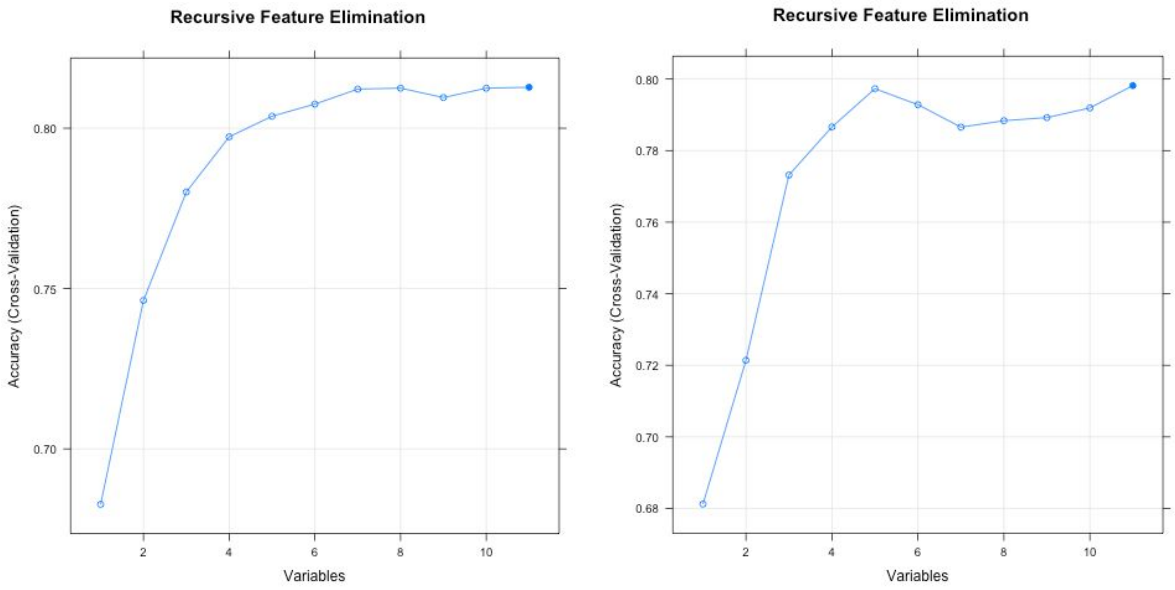

Figure 1: Backward Elimination Plots for White Wine (Left) and Red Wine (Right)

Methods

Logistic Regression: A logistic regression is a binary classification technique. We use this technique when our response variable is binary (i.e. 'bad' or 'good') and a collection of real valued explanatory variables.

SVM: SVM differs radically from approaches such as Neural Networks (NN) because SVM training always find a global minimum in contrast to NN. SVMs are supervised learning methods used for classification. The aim of a SVM is to find the best linear separating hyperplane, tolerating a small error when fitting the data, in the feature space. During training, a linear SVM constructs a dimensional hyperplane that separates the points into two classes. When the classes are not linearly separable, meaning there is no hyperplane that can split the two classes, a variant of SVM chooses a hyperplane that splits the points as clearly as possible. This split maximizes the distance to the nearest cleanly split examples. Maximizing the margin allows one to minimize bounds on generalization error. We chose to test the linear and radial basis function (RBF) kernels for our analysis. The best choice of kernel for a given problem is still a research issue. Because the size of the margin does not depend on the data dimension, SVM are robust with respect to data with high input dimension.

Artificial Neural Networks (ANNs): ANNs learn by training on past experience using an algorithm, which modifies the interconnection weight links as directed by a learning objective for a particular application. A neuron is a single processing unit, which then computes the weighted sum of its inputs. The output of the network relies on cooperation of the individual

neurons. The learnt knowledge is then distributed over the trained network weights. Neural networks are characterized into feedforward and recurrent neural networks. In order to improve NN generalization and avoid overfitting, Average Neural Networks (AvgNN) are also tested in our analysis. Overfitting occurs during NN training. The network has memorized the training examples, but it has not learned to generalize to new situations. Instead, we train multiple neural networks and average their outputs.

Decision trees: Random Forests are one way to improve the performance of decision trees. The algorithm begins by building out trees similar to the way a normal decision tree algorithm works. However, every time a split is made, it uses only a small random subset of features to make the split instead of the full set of features. It builds multiple trees using the same process, and takes the average of all the trees to arrive at the final model. By doing this, the correlation between trees is reduced and the variance of the final tree is reduced as well.

Results and Discussion

Table 2: Mean Accuracy Rates

	Red	White
GLM	71.43%	71.28%
SVM	70.85%	73.62%
SVM-RBF	71.30%	75.47%
NN	69.20%	75.51%
AvgNN	73.78%	76.53%
RF	**77.68%**	**80.90%**

Before fitting the classification models, the input variables were transformed individually so that every input had a mean of 0 and a standard deviation of 1. Recursive feature eliminations were performed on both the red and white datasets in order to determine whether or not the removal of input variables would improve the accuracy of our classification models. The red and white datasets were then split into training and testing sets, each of which constituted 70% and 30% of the data respectively. A 5-fold cross-validation was utilized in order to evaluate the fitted classification models. Results of this cross-validation are reported at a 95% confidence level.

The classification methods performed in this study are evaluated using the testing accuracy, or the number of wine quality attributes that were correctly classified as either ‘good’ (6-10) or ‘bad’ (1-5). The resulting testing accuracies are reported in Table 2 above. The mean testing accuracy for SVM with linear kernel and regularization parameter C equal to 0.001, 0.01, 0.1, 1,

10, 100, and 1000 is 70.85% for red wine and 73.62% for white wine. The mean testing accuracy for SVM with RBF kernel, regularization parameter C equal to 0.001, 0.01, 0.1, 1, 10, 100, and 1000, and sigma equal to 0.001, 0.01, and 0.1 is 71.30% for red wine and 75.47% for white wine. The mean testing accuracy for ANN with 1, 5, and 10 hidden layers and weight decay equal to 0, 0.001, and 0.1 is 69.20% for red wine and 75.51% for white wine. The mean testing accuracy for AvgNN with 1, 5, and 10 hidden layers and weight decay equal to 0, 0.001, and 0.1 is 73.78% for red wine and 76.53% for white wine. The mean testing accuracy for random forest with number of variables randomly sampled as candidates at each split equal to 2, 3, 4, 5, and 6 is 77.68% for red wine and 80.90% for white wine.

The results of this classification analysis point to random forest as the best classifier for both red and white wine. This method significantly outperformed all other methods considered in this analysis by a minimum of around 4% accuracy. It is important to note, however, that all of our classification models performed very well. The findings of this analysis support the conclusion in Cortez's 2009 study that SVM methods outperformed ANN and regression methods (Cortez *et al.,* 2009). However, the AvgNN method outperformed SVM with linear kernel and with RBF kernel for both red and white wine varieties. It is also interesting to note that the random forest decision tree method performed much better than the decision tree-based ID3 method performed in Appalasamy's 2012 study, which resulted in testing accuracies of 60.0% and 52.3% for red and white wine respectively (Appalasamy *et al.,* 2012).

Table 3: Mean Kappa Statistics

	Red	White
GLM	47.17%	39.47%
SVM	47.07%	39.48%
SVM-RBF	51.01%	47.28%
NN	51.42%	44.85%
AvgNN	52.55%	49.49%
RF	**59.66%**	**58.15%**

Another method used to evaluate the classifiers is the Kappa statistic. The Kappa statistic is a measure of agreement between two rates, where agreement due to chance is factored out. It compares an observed accuracy with an expected accuracy. Kappa is always less than or equal to 1. A value of 1 implies perfect agreement and values less than 1 imply less than perfect agreement. Different people have different interpretations as to what is a good level of agreement. Landis and Koch have proposed the following as standards for strength of agreement

for the kappa coefficient: ≤ = poor, 0.01-0.20 = slight, 0.21-0.40 = fair, 0.41-0.60 = moderate, 0.61-0.80 = substantial, and 0.81-1 = almost perfect (Sim *et al.,* 2005).

According to Table 3, white wine's GLM model had the lowest Kappa statistic (39.47%). The Random Forest method displayed the highest strength of agreement for the kappa statistic for both red and white wines. The Kappa statistics are 59.66% and 58.15% respectively. All the classifiers had a fair or moderate strength of agreement according to Landis and Koch's proposed standards. Regardless, the Kappa statistic and accuracy rates both show that the Random Forest classification method outperforms the other three methods: SVM, Logistic Regression, and Neural Networks.

Conclusions

The results of this work are important for the wine industry. Ours is the first wine classification analysis to point to the decision tree-based random forest as the most accurate in identifying wine quality. However, it is important to note that no researchers have compared the same methods. Quality ratings provided by human wine experts are prone to subjectivity, all while being extremely costly and time consuming. With machine learning techniques, however, we are able to determine factors that can objectively classify wine quality in a significantly more timely and cost-effective manner.

Further investigation is needed to determine whether or not red and white wine together can produce classification rates as high or higher than those considering red and white wine separately. Another future consideration entails going beyond a binary classification system to one more multi-faceted and complex. Lastly, the dataset used in this analysis is the largest compilation of wine data to date but is still fairly small for a full cross-validation with training, testing and validating sets. The creation of larger wine datasets would aid in the production of more reliable findings in the area of wine quality classification.

Given more time, we would want to investigate if different physicochemical properties affect the wine quality. For example, does higher alcohol content mean higher wine quality? Does the level of acidity mean lower wine quality? In addition, it is appropriate to use other datasets of wine and its physicochemical properties. There may not be a general model for all wine, as the ratings for wine quality may differ by the country. Our current analysis uses Portuguese wine products but it is unclear as to who are the judges. They could be individuals from Portugal, or the United States, or any country. This can change one's judgment in the quality of wine: Americans may find sweeter wine as 'good' wine, meaning there's less acidity, whereas French people may consider a bitter, more acidic wine as 'good.' Therefore, our analysis may have low external validity, and our results may not be able to transcend international boundaries.

References

A. Nachev, and M. Hogan. Using Data Mining Techniques to Predict Product Quality from Physicochemical Data. *Proceedings on the International Conference on Artificial Intelligence (ICAI)*. The Steering Committee of The World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp), 2013.

J. Sim, and C. Wright. The Kappa Statistic in Reliability Studies: Use, Interpretation, and Sample Size Requirements. *Journal of the American Physical Therapy Association,* 85(3): 257-268, 2005.

P. Appalasamy, A. Mustapha, N.D. Rizal, F. Johari, and A.F. Mansor. Classification-based Data Mining Approach for Quality Control in Wine Production. In Journal of Applied Sciences, 12(6): 598-601, 2012.

P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis. Modeling wine preferences by data mining from physicochemical properties. In Decision Support Systems, Elsevier, 47(4):547-553, 2009.

Appendix

Tables and Figures

Table 1: Physicochemical Data Statistics for Red and White Wine

Attributes	Red Wine			White Wine		
	Min	**Max**	**Mean**	**Min**	**Max**	**Mean**
Fixed Acidity	4.600	15.900	8.320	3.800	14.200	6.855
Volatile Acidity	0.120	1.580	0.538	0.080	1.100	0.278
Citric Acid	0.000	1.000	0.271	0.000	1.660	0.334
Residual Sugar	0.900	15.500	2.539	0.600	65.800	6.391
Chlorides	0.0120	0.611	0.085	0.009	0.346	0.046
Free Sulfur Dioxide	1.000	72.000	15.870	2.000	289.000	35.310
Total Sulfur Dioxide	6.000	289.000	46.470	9.000	440.000	138.400
Density	0.990	1.004	0.997	0.987	1.039	0.994
pH	2.740	4.010	3.311	2.720	3.820	3.188
Sulphates	0.330	2.000	0.658	0.220	1.080	0.490
Alcohol	8.400	14.900	10.420	8.000	14.200	10.510
Quality	3.000	8.000	5.636	3.000	9.000	5.878

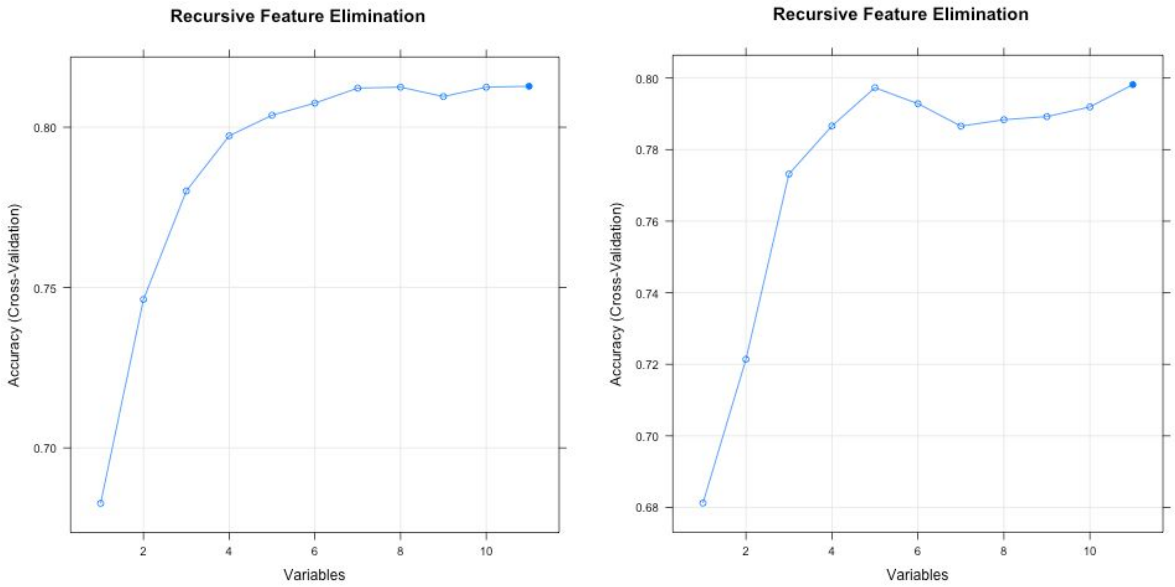

Figure 1: Backward Elimination Plots for White Wine (Left) and Red Wine (Right)

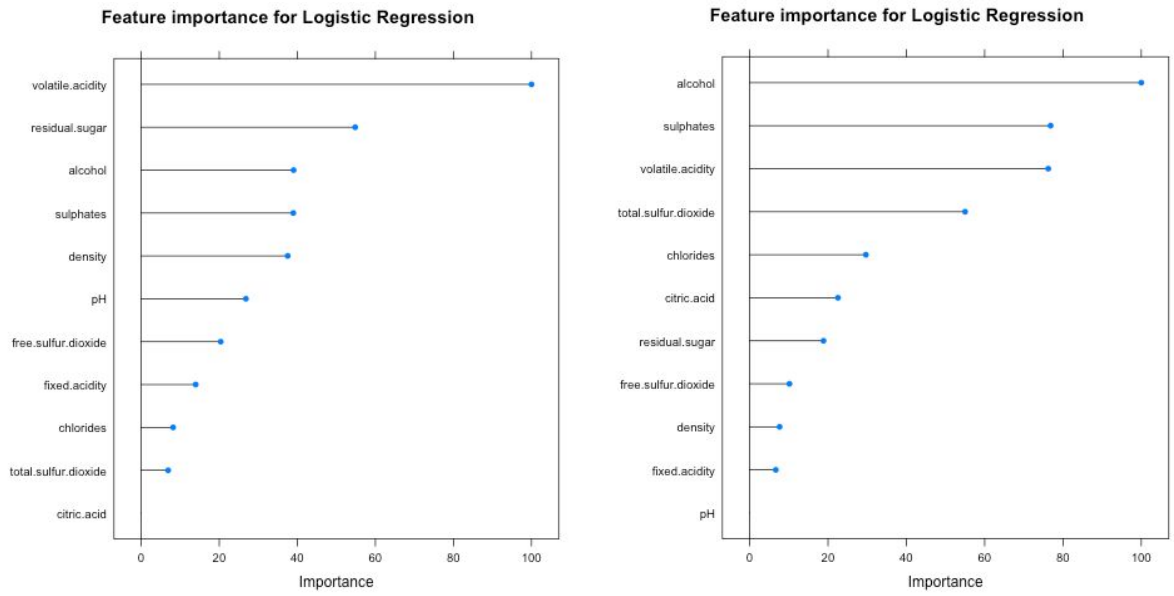

Figure 2. GLM Feature Importance for White Wine (Left) and Red Wine (Right)

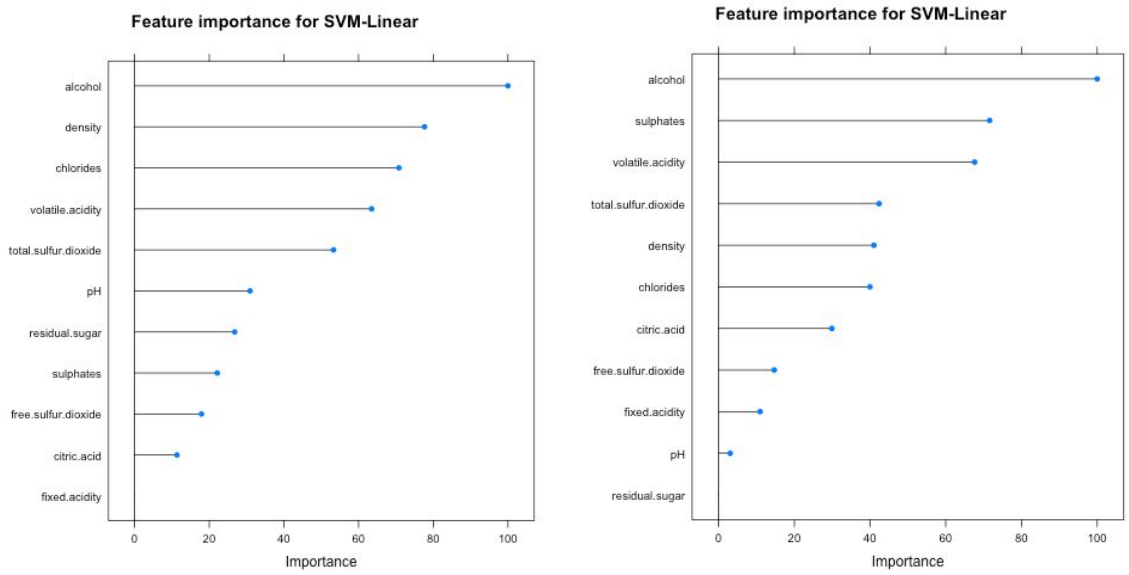

Figure 3. SVM (Linear) Feature Importance for White Wine (Left) and Red Wine (Right)

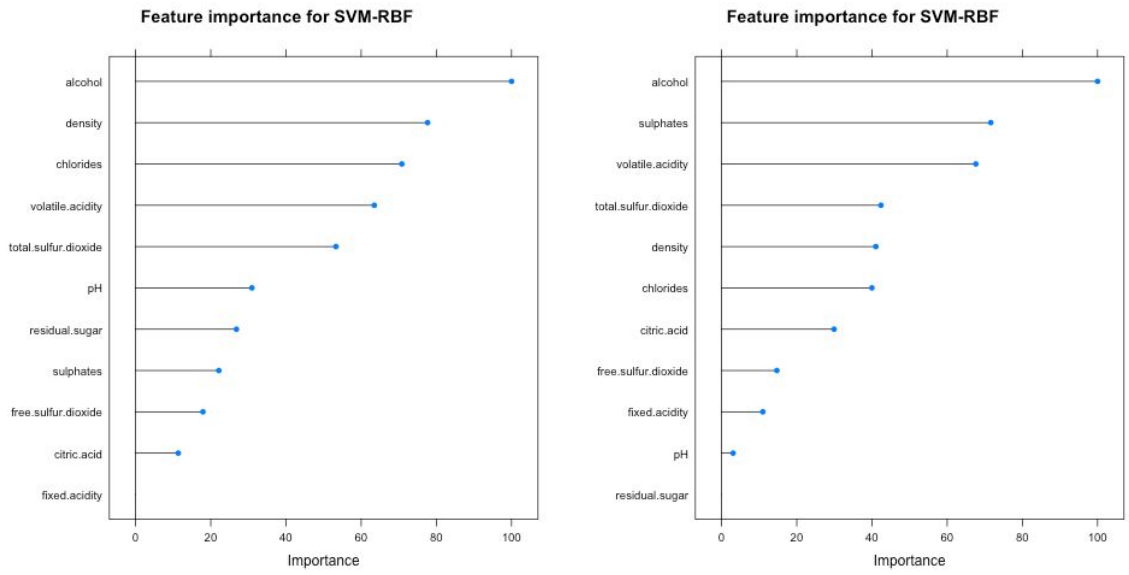

Figure 4. SVM (RBF) Feature Importance for White Wine (Left) and Red Wine (Right)

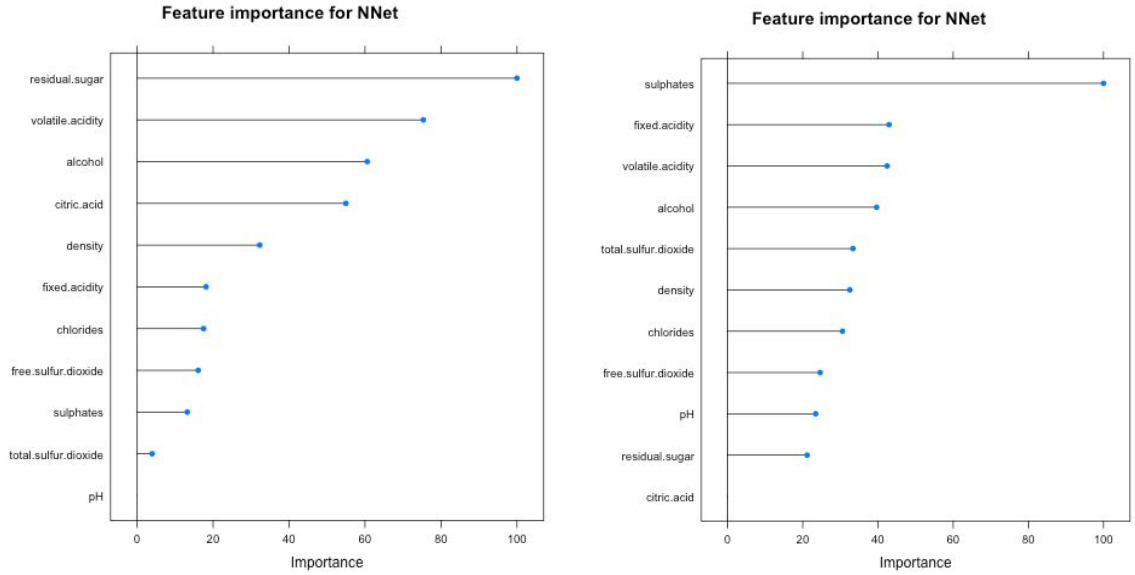

Figure 5. NN Feature Importance for White Wine (Left) and Red Wine (Right)

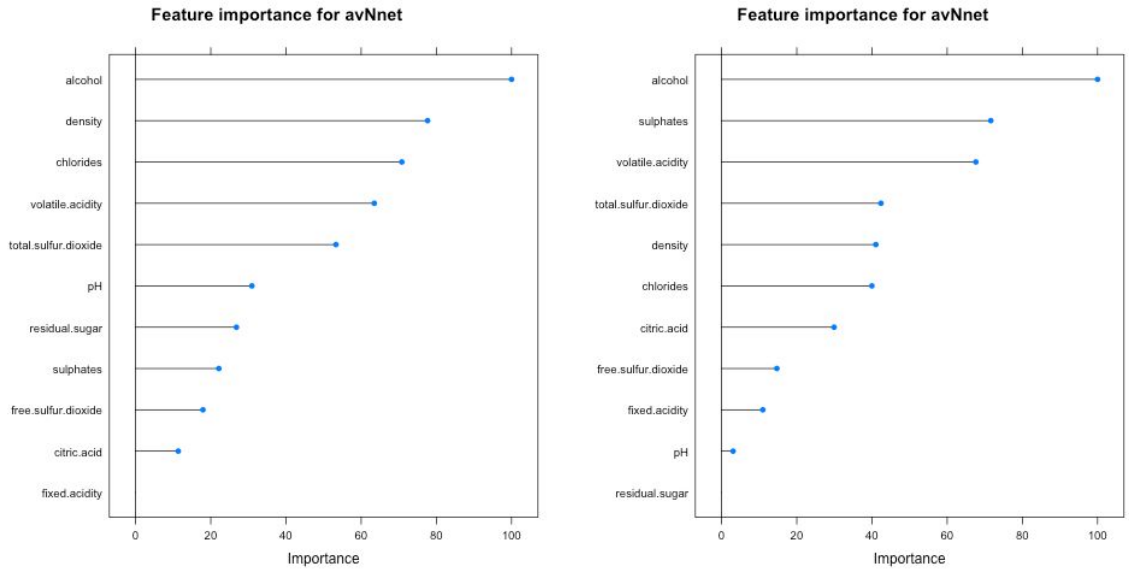

Figure 6. AvgNN Feature Importance for White Wine (Left) and Red Wine (Right)

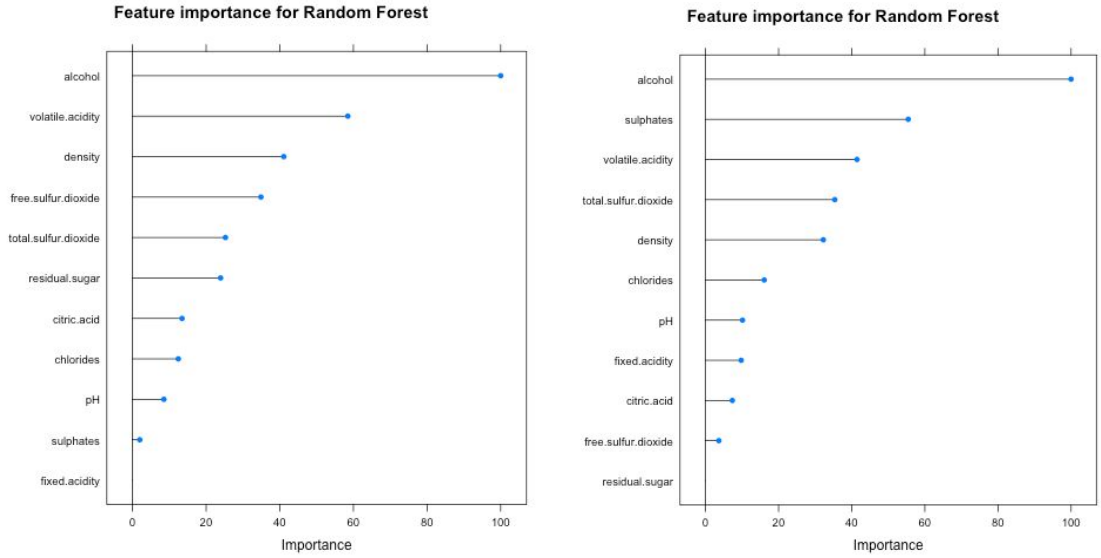

Figure 7. RF Feature Importance for White Wine (Left) and Red Wine (Right)

Table 2: Mean Accuracy Rates for GLM, SVM, SVM-RBF, NN, AvgNN, and RF

	Red	White
GLM	71.43%	71.28%
SVM	70.85%	73.62%
SVM-RBF	71.30%	75.47%
NN	69.20%	75.51%

AvgNN	73.78%	76.53%
RF	**77.68%**	**80.90%**

Table 3: Mean Kappa Statistics

	Red	**White**
GLM	47.17%	39.47%
SVM	47.07%	39.48%
SVM-RBF	51.01%	47.28%
NN	51.42%	44.85%
AvgNN	52.55%	49.49%
RF	**59.66%**	**58.15%**