Wine Quality Prediction

STAT 154

Rowan Cassius, Rachel Henry, Bharvee Patel

I. Introduction

Wine is an alcoholic beverage in which alcohol is generated by the natural process of fermentation. The fermentation process is primarily carried out by the bacteria that are present on the skin of grapes. Six processes are involved in the production of wine, namely destemming and crushing, alcoholic fermentation, drawing the wine off the lees, malolactic fermentation, stabilization and aging and refinement in bottle. It is common knowledge that the quality of wine increases with the time of aging. However, of course, the ingredients of the wine largely determine the fermentation process and, ultimately, the taste of the wine.

As the global wine market trends suggests, this industry will reach 423.59 billion USD by 2023[1], and it will become critical for wine brewers and distributors to take into account how to perform better than their competition and produce wine that consumers will prefer.

Therefore, a question of interest for producers and distributors of different wines would be to predict the quality of wine from a consumer perspective to gauge the predicted popularity of particular products. Furthermore, it may be of particular concern for a critic or consumer of wine to be able to distinguish between exceptional and ordinary wines.

The data in this report are focused, specifically, on red wine varieties. The goals of this project are to:

1. Deduce variable importance in the prediction of red wine quality.
2. Assess the need for a model that predicts red wine quality.
3. Build a predictive model accordingly.

II. Data

i. Description

The dataset, named "Red Wine Quality", contains data related to the red variants of the Portugese "Vinho Verde" wine. This multivariate dataset contains 12 attributes over 1599 varieties of red wines. This data, collected on Kaggle[2] are collected through physiochemical tests. These variables are described as follows:

1. `fixed acidity`: overall concentration of fixed acids found in wine (tartaric, malic, citric, and succinic acids) (in g/L)
2. `volatile acidity`: acetic acid concentration found in wine (in g/L); at high levels, volatile acidity can lead to an unpleasant, vinegar taste
3. `citric acid`: citric acid concentration found in wine (in g/L); in small quantities, citric acid can add freshness and flavor to wines (in g/L)
4. `residual sugar`: the amount of sugar remaining after fermentation process ends (in g/L)
5. `chlorides`: the amount of salt found in the wine
6. `free sulfur dioxide`: the free form of SO_2 existing in equilibrium between molecular SO2 (as a dissolved gas) and bisulfite ion (in ppm)
7. `total sulfur dioxide`: amount of free and bound forms of SO_2 found in wine (in ppm)
8. `density`: the concentration of alcohol, sugar, glycerol, and other dissolved solids in wine (in g/mL)

[1] https://globenewswire.com/news-release/2018/04/09/1467083/0/en/Global-Wine-Market-Will-Reach-USD-423-59-Billion-by-2023-Zion-Market
html

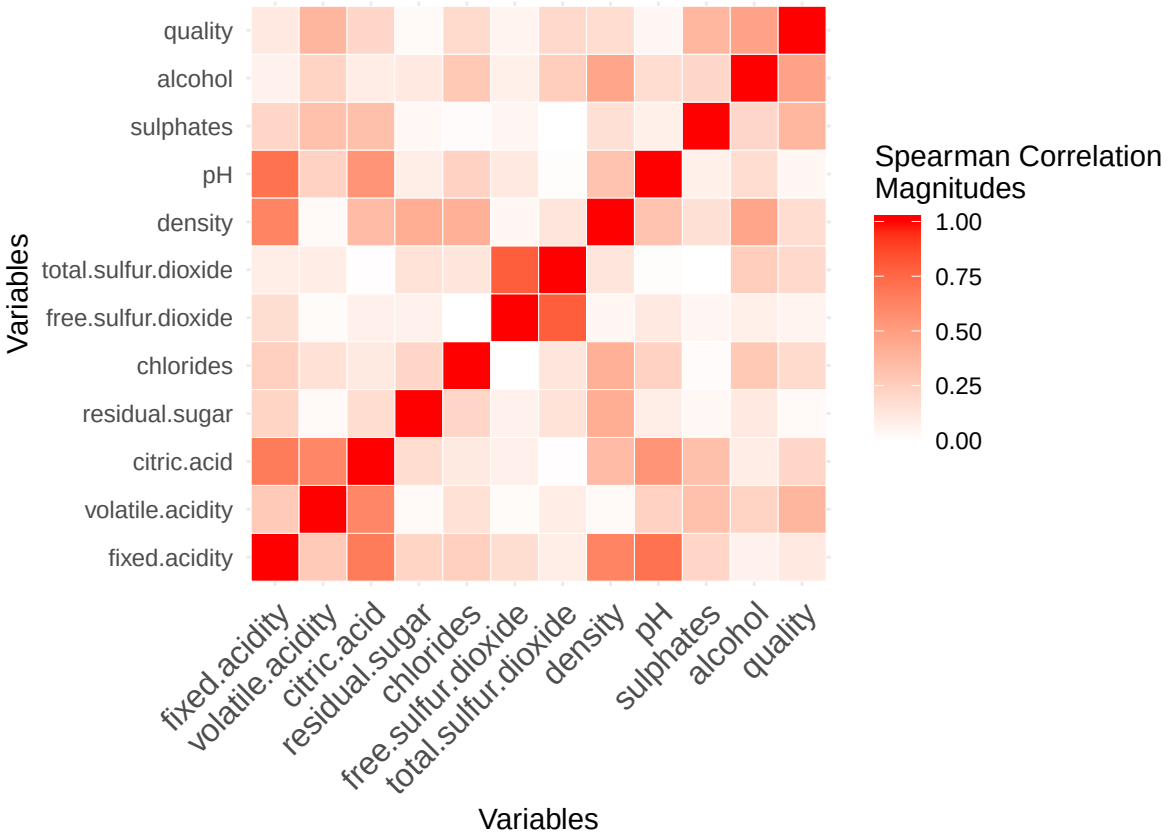
[2] The "Red Wine Quality" dataset can be obtained at: https://www.kaggle.com/uciml/red-wine-quality-cortez-et-al-2009.

9. `pH`: describes how acidic or basic a wine is on a scale from 0 (very acidic) to 14 (very basic); most wines are between 3-4 on the pH scale
10. `sulphates`: level of wine additive which contributes to sulfur dioxide gas levels (in g/L)
11. `alcohol`: the percent alcohol content of the wine
12. `quality`: quality of wine, based on sensory data; score between 0 and 10

ii. Exploration

Berfore attempting to construct a model that can predict wine quality from all predictors variables listed, we will do some exploratory analyses to gain a broad understanding of the data as a whole. Our first step in the exploratory data analysis is to seek an overview of how the data are related to one another.

In order to understand how different variables are correlated, we abstain from looking at their linear correlations to avoid assuming linear relationships between all of the variables. Instead, we use the Spearman Correlation, which only assumes that each pair of variables has a relationship that can be described by a monotonic function. Unconcerned with directionality of correlations, we first visualize the magnitudes of the variables' pairwise Spearman Correlations in a matrix.



It is easy to see in the figure above that all but a few pairs of variables have mild correlations. The variable clusters with the greatest correlations are unsurprisingly `total sulfur dioxide` and `free sulfur dioxide`, the acid-measuring variables `citric acid` and `volatile acidity`, and `pH` and `fixed acidity`. When constructing a linear predictive model that is a generalized linear predictive model, we may have to be cautious about including these pairs together to avoid dangers of colinearlity.

Additionally, it is important to note that `quality` is not strongly correlated with any of its predictors; thus, making preditions about quality is unlikely to depend solely on one, or a couple variables. Of all predictors, `alcohol`, `volatile acidity` and `sulphates` show the strongest, although still moderate, correlations with `quality`. Thus, we expect that these predictors are likely to have most important predictive importance.

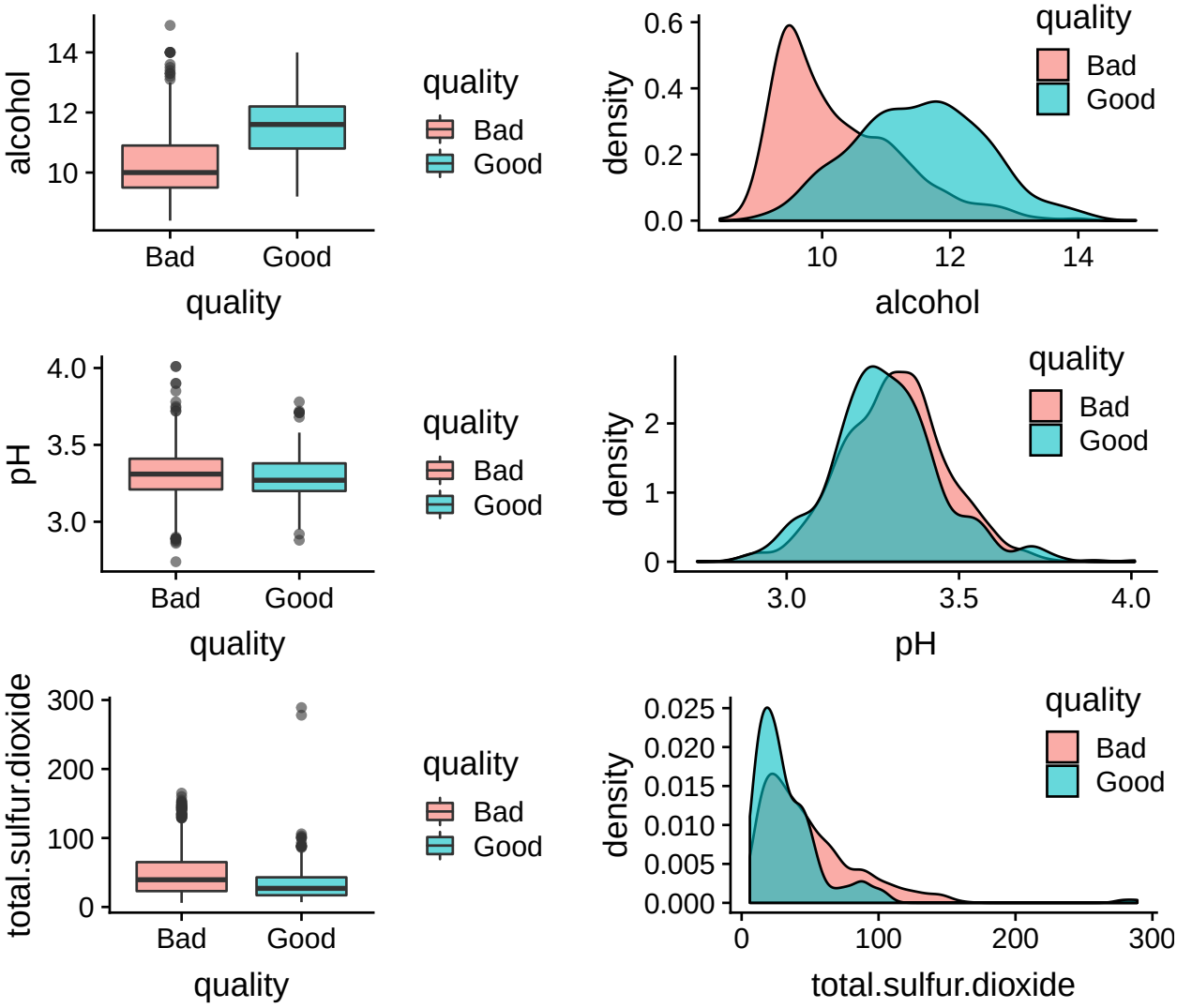
iii. Choosing a Binary Classification Problem

After analyzing our particular data set, we decided to make `quality` a binary classifier, using the classes `Good` and `Bad`. We were motivated to do this after acknowledging the context of our data set and its applications. For example, if we were to present our final model and some test observations to a consumer, it is likely that the consumer would only be interested in knowing which wines are "Good" or "Bad," rather than being given a numeric qualifier. In that same vein, if we were presenting our final model to a wine critic, a critic would likely only be interested in what the "best" wine is. Thus, we set our threshold for "Good" wine at a high level, with a quality greater than or equal to 7 being the criterion to classify as a "Good" wine.

After making our quality variable binary, we have 217 "Good" observations and 1382 "Bad" observations.

iv. Variable Assessment

While we are limited to just 2 dimensions when visualizing the data, we can still gain a comprehensive understanding of how each predictor is related to wine quality by comparing the box plots and kernel densitites of every predictor on the subsets of good and bad wine separately. We begin this variable study by examining 3 predictors: `alcohol`,`total sulfur dioxide`, and `pH`. The results are displayed in the figure below.



Observing the two box plots for `alcohol` for the good (green) and bad (red) subsets of wine, the means

of the two distributions appear drastically different, which provides more evidence to the hypothesis that alcohol will be an important predictive variable when predicting wine quality. In addition, the two kernel densisites of alcohol are rather visually convincing of the notion that the distribution of alcohol is *inherently* different in good and bad wines.

On the other hand, the `pH` box plots show good and bad means that do not appear significantly different, and the two kernel densities for `pH` do not appear to be inherently different distributions of `pH`. In other words, knowledge of a wine's pH level is unlikely to aid the prediction of the wine's quality. Finally, the total `total sulfur dioxide` plots show that whether the distributions of total sulfur dioxide in good and bad wines are different to a useful extent is ambiguous.

Thus, we will quantify the effect size of each variable on wine quality both to avoid having to examine each variable's effect size on wine quality visually, and to have the ability to rank them. To do this, we will calculate the Cliff's Delta statistic for every predictor, and treat the good and bad wine populations separately. We choose Cliff's Delta because it is a standardized and non-parametric measure of effect size that avoids making shaky assumptions about the data. Cliff's Delta is defined as follows:

Cliff's Delta:

$$\text{statistic, } d = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n} I(x_i > x_j) - I(x_i < x_j)}{mn} - 1$$

range: $(-1, 1)$

m, good population size

n, bad population size

We compute Cliffs Delta for each predictor and obtain the following results:

	Predictor	Cliffs.Delta	effect
1	alcohol	-0.6694779	0.6694779
11	volatile.acidity	0.4904284	0.4904284
9	sulphates	-0.4726102	0.4726102
3	citric.acid	-0.3436013	0.3436013
4	density	0.2865457	0.2865457
10	total.sulfur.dioxide	0.2822483	0.2822483
2	chlorides	0.2760959	0.2760959
5	fixed.acidity	-0.2069017	0.2069017
6	free.sulfur.dioxide	0.1325840	0.1325840
8	residual.sugar	-0.1017855	0.1017855
7	pH	0.0838010	0.0838010

The table of predictors and effect sizes shows that `alcohol`, `volatile acidity` and `sulphates` have the greatest effects on wine quality according to Cliff's Delta. This result provides strong evidence that these will likely be the most important predictors of wine quality in all of the predictive models to come.

III. Modeling: A Classification Problem

i. K-Nearest Neighbors

The K-Nearest Neighbors (KNN) algorithm is a non-parametric and lazy learning algorithm, which classifies an observation depending on the majority class of its k neighbors.

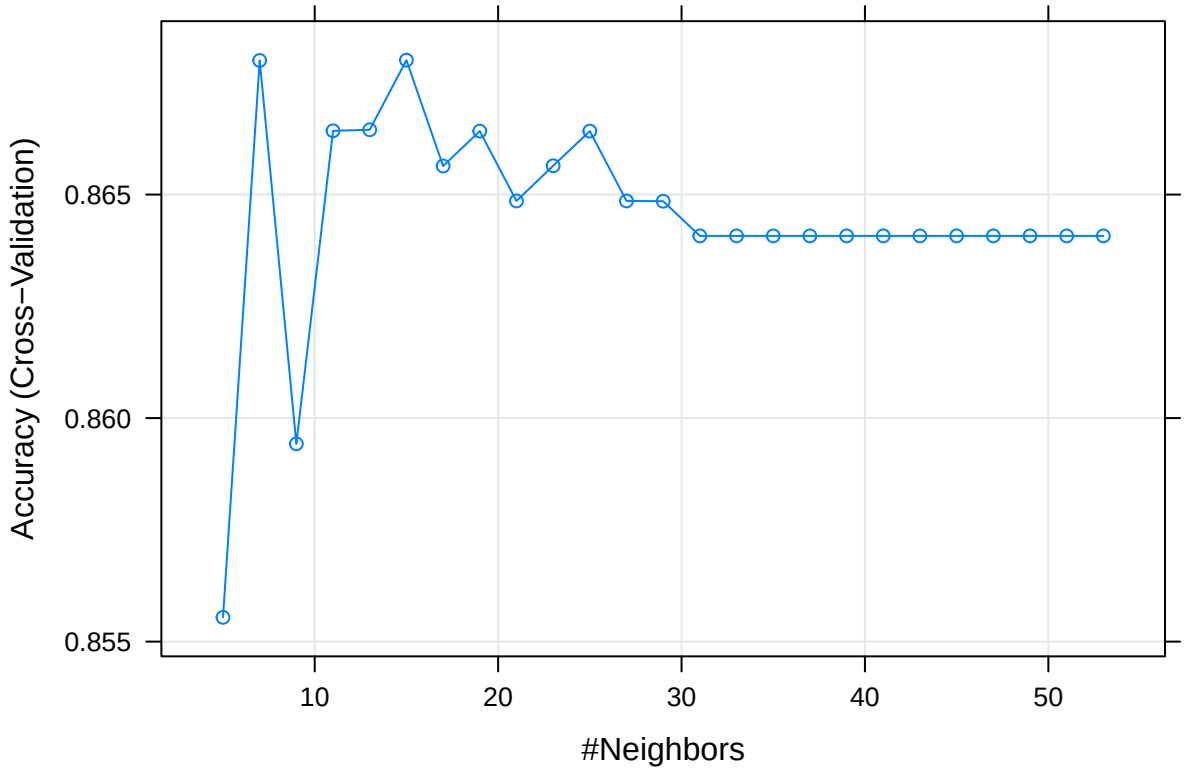
This classification decision is defined according to a distance metric between two data points. The k-nearest-neighbor classifier is commonly based on the Euclidean distance between a test sample and the specified training samples. Let x_i be an input sample with p features $(x_{i1}, x_{i2}, ..., x_{ip})$, n be the total number of input samples $(i = 1, 2, ..., n)$. The Euclidean distance between sample x_i and x_l is is defined as:

$$d(x_i, x_l) = \sqrt{(x_{i1} - x_{l1})^2 + (x_{i2} - x_{l2})^2 + ... + (x_{ip} - x_{lp})^2}$$

Benefits and Disadvantages of KNN on Wine Quality Data:

Because KNN is non-parametric, this method does not make any assumptions on the underlying data distribution. In other words, the model structure is determined from the data. Therefore, this algorithm may be useful to predict wine quality as the underlying distribution of the data is unknown. because in the "real world", most of the data does not obey the typical theoretical assumptions made (as in linear regression models, for example).

One potential concern for implementing KNN on the wine quality dataset is related to the curse of dimensionality. KNN depends greatly on distances between observations; as the number of dimensions in the data increases, the distances between observations will likely be less representative. This is because KNN is sensitive to irrelevant attributes, as features that are not significant predictors will affect the distances between observations and affect the classification decision. Further, as classes are imbalanced in the Wine Quality data, KNN will favor classification of the majority class, which is "Bad" wine.



After cross-validation, we see that the number of neighbors, k, that maximizes accuracy is k=15. The confusion matrix for the K-Nearest Neighbors model is as follows:

	Bad	Good
Bad	270	41
Good	6	2

labels	values
Test Classification Error Rate	0.1473354
Accuracy	0.8526646
Precision	0.2500000
Recall	0.0465116

As the confusion matrix shows, there are very few "Good" wines that are classified corerectly. About 95% of the "Good" wines are classified as "Bad", precisely due to the imbalanced nature of the data. As a result, the KNN approach is unlikely to be the most optimal model to predict wine quality.

ii. Classification Tree

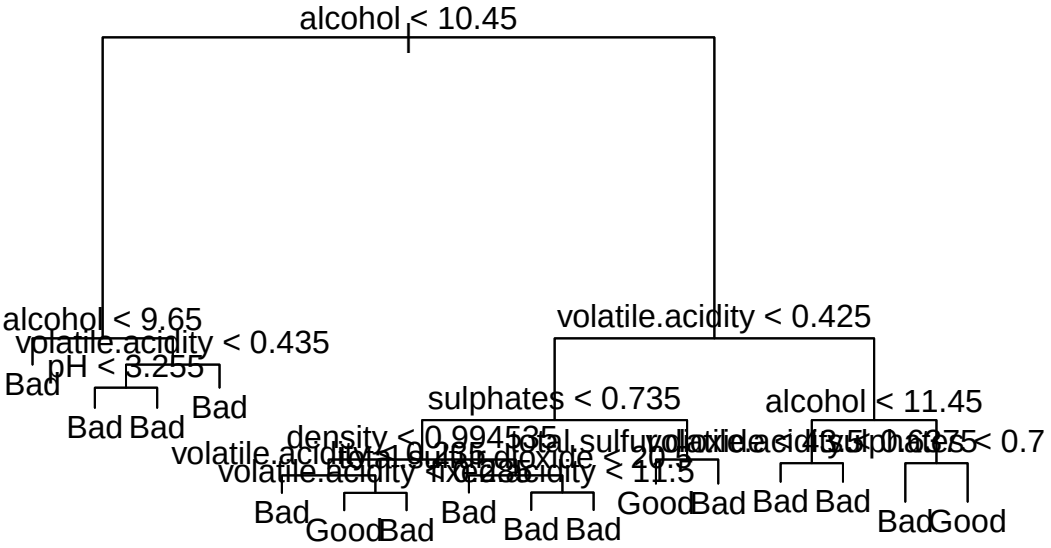
A decision tree is a non-parametric algorithm for regression and classification problems. A decision tree makes sequential, hierarchical decisions about the outcomes variable based on the predictor data.

The model is defined by a series of "rules" that lead to a class label when applied to any observation. Once set up, the model acts as a protocol in a series of "if, then" conditions that produce a specific result from the input data. Decision trees are a non-parametric method, thus there are no underlying assumptions about the distribution of the errors or the data.

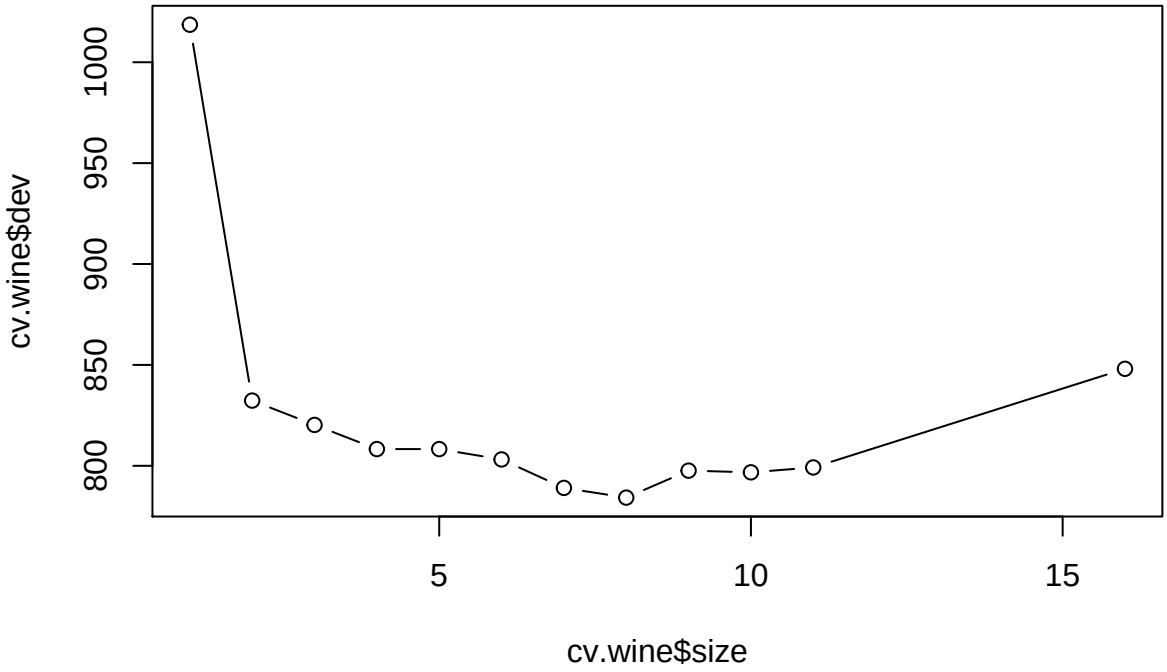
Benefits and Disadvantages of Classification Tree on Wine Quality Data:

This model presents a concern for the wine data, as, at the expense of bias the variance for this model is massive and will likely lead to overfitting. Decision trees can become very complex and may not generalize well from the training data. Further, decision trees are locally optimized, so the greedy algorithm cannot guarantee a return to the globally optimal decision tree. It is an incredibly biased model if a single class takes precendence in the data. The wine quality dataset, however, is not balanced as there is a lower proportion of "Good" wines.

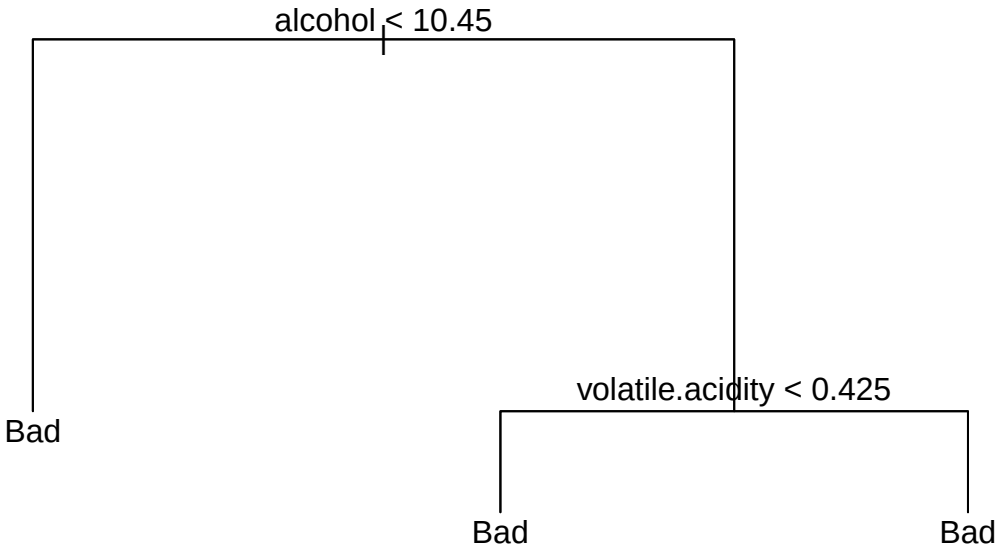
While there are many disadvantages to this model, with regard to the dataset at hand, there are some advantages to decision trees. They are incredibly simple to understand due to their visual representation, can handle large amounts of data and are quite computationally inexpensive. This may be helpful in this context, as producers may be interested in how exactly to manipulate the ingredients of a wine to make future products better preferred by the consumer.



The original tree is very complex, with many rules and nodes. The misclassification error rate resulting from the original tree is 0.205. Due to possible concerns regarding overfitting, cross validation was performed to choose the best number of terminal nodes for the pruned tree.



The cross-validation result shows that the pruned tree should have 3 terminal nodes.



After constructing the pruned true, it is clear that alcohol and volatile acidity are the only variables that contribute to the classification decision of any observation. The The misclassification error rate resulting from the pruned tree is 0.199. This is similar to the original tree. The problem with this model is in its neglection of the "Good" class. As the data is imbalanced, pruning creates a decision that only classifies observations as "Bad." This is extremely detrimental to the problem we are trying to solve, as we are specifically interested in knowing which wines are "Good." The imbalanced nature of the data does not lend itself well to the classification tree model, which suggests that this approach will not likely be the final model in predicting wine quality.

In terms of measuring the misclassification, there are three measures we can use: Entropy, Gini index, and Classification Error. Entropy $= -\sum_j p_j \log_2 p_j$ Gini $= 1 - \sum_j p_j^2$ Classification Error $= 1 - \max p_j$, where p_j is the probability of class j.

The entropy is 0 if all samples of a node belong to the same class, and the entropy is maximal if we have a uniform class distribution. In other words, the entropy of a node (consist of single class) is zero because the probability is 1 and log (1) = 0. Entropy reaches maximum value when all classes in the node have equal probability.

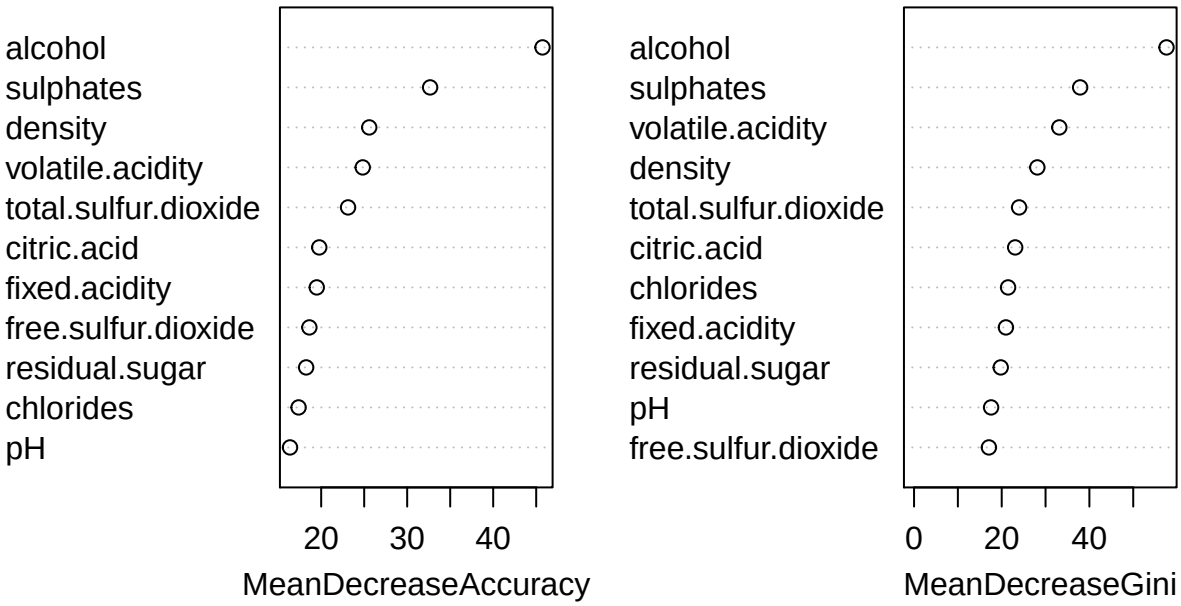
When building a classification tree, either the Gini index or the entropy are typically used to evaluate the quality of a particular split, since these two approaches are more sensitive to node purity than is the classification error rate. Any of these three approaches might be used when pruning the tree, but the classification error rate is preferable if prediction accuracy of the final pruned tree is the goal.Therefore, because the main goal of our project is to improve prediction accuracy, we use the misclassification error rate as the measure to determine the performance of the classification tree model.

iii. Random Forest

Random Forest is a supervised learning algorithm for classification and regression that constructs a number of decision trees and outputs the class that is the mode of classification. Random forests use boostrapped data to correct for decision trees' tendency to overfit to their training sets and can also reduce variance.

Benefits and Disadvantages of Random Forest on Wine Quality Data:

Though random forests can reduce overfitting, random forests are not as easy as decision trees to visually interpret and do not reduce variance if the features are correlated.



The confusion matrix resulting from the Random Forest model is as follows:

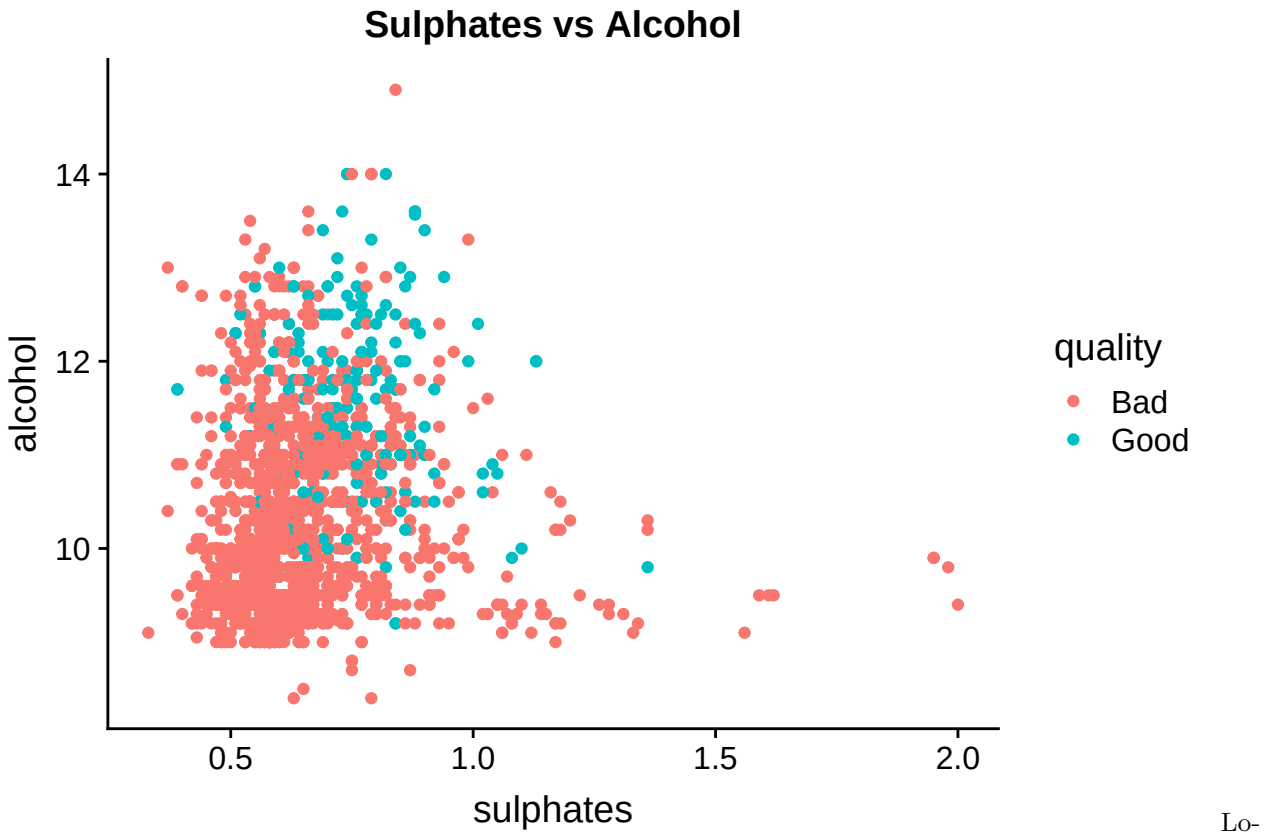
	Bad	Good
Bad	269	17
Good	7	26

labels	values
Test Classification Error Rate	0.0752351
Accuracy	0.9247649
Precision	0.7878788
Recall	0.6046512

iv. Logistic Regression

Benefits and Disadvantages of Logistic Regression on Wine Quality Data:

While tree based methods produced moderate results, Logistic Regression may perform better because it tends to do well in cases where the data are not easily separable. The following view captures the data at one of their most separable.



Logistic regression is a generalized liner model which assumes linearity between the log odds of a wine being "Good" and the predictors; that is,

$$log(\frac{p}{1-p}) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_n, \text{ where } n \text{ is the number of predictors}$$

In order to select the variable to use in the logistic regression model we run both forward and backward stepwise slection processes based on the AIC criterion. The forwards selection converges to a model which uses `fixed acidity`, `volatile acidity`, `residual sugar`, `chlorides total sulfur dioxide`, `density`, `sulphates` and `alcohol` as predictive variables with and $AIC = 654.64$. On the other hand, forwards stepwise selection arrives at a model which uses `alcohol`, `volatile acidity`, `sulphates`, `chlorides`, `total sulfur dioxide` and 'pH as predictive variables with $AIC = 663.61$.

Thus, because forward and backward selection converge to models with different predictors, we test the union of the suggested predictors for significance after training a model based on all of them, $\{forwardpredictors\} \cup \{backwardpredictors\}$ for significance.

	p-value
(Intercept)	0.000
alcohol	0.000
volatile.acidity	0.000
sulphates	0.000
fixed.acidity	0.001
density	0.000
total.sulfur.dioxide	0.001
chlorides	0.008
residual.sugar	0.001
pH	0.910

As displayed above, the predictors `alcohol`, `volatile acidity`, `sulphates`, `total sulfur dioxide`, `chlorides` and `residual sugar` are considered significant at the 5% significance level. Thus the logistic regression model that is used to make predictions will use these significant variables.

% Table created by stargazer v.5.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Fri, Dec 07, 2018 - 18:36:34

Table 7: Logistic Regression Results

	Dependent variable:
	quality
alcohol	1.010^{***}
	(0.095)
volatile.acidity	-3.998^{***}
	(0.671)
sulphates	3.693^{***}
	(0.613)
chlorides	-7.012^{**}
	(3.026)
residual.sugar	0.081
	(0.063)
Constant	-13.177^{***}
	(1.291)
Akaike Inf. Crit.	719.513
Note:	*p<0.1; **p<0.05; ***p<0.01

The confusion matrix for the Logistic Regression model is as follows:

	Bad	Good
Bad	263	32
Good	13	11

labels	values
Test Classification Error Rate	0.1410658
Accuracy	0.8589342
Precision	0.4583333
Recall	0.2558140

v. Support Vector Machine

A support vector machine is a supervised model used for classification and regression. A SVM training algorithm uses the given data, each data point of which belongs to one of two categories, to build a model

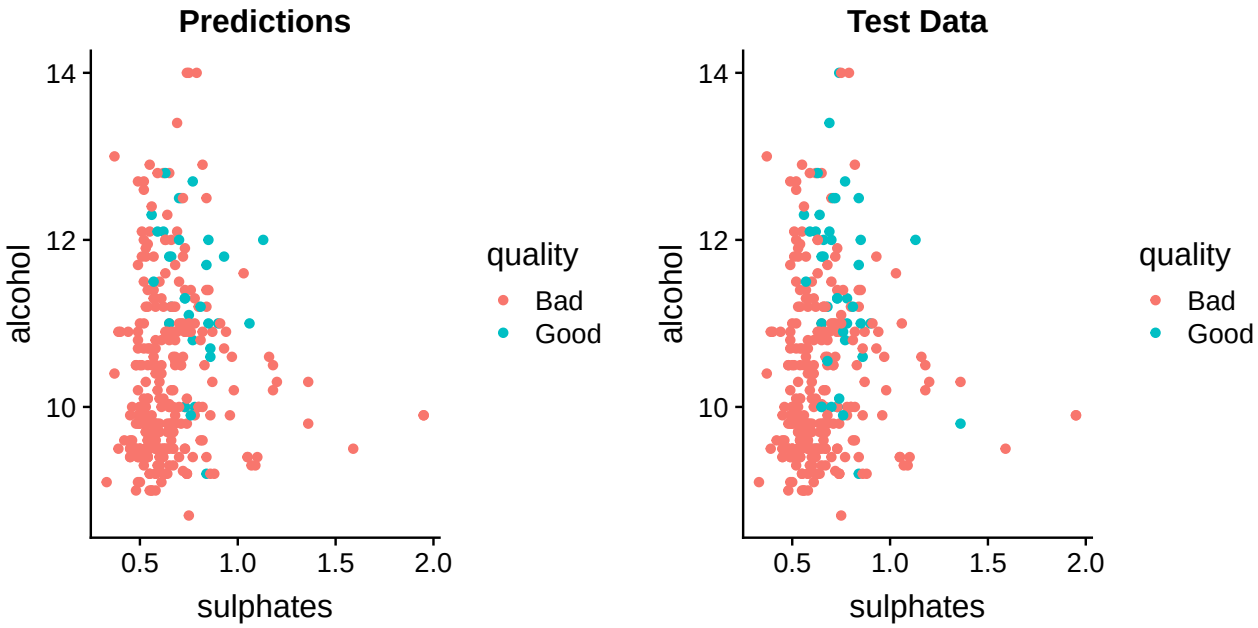
that assigns new examples to one of the two categories. When mapped, the SVM separates the 2 categories with a clear gap and predicts on which side of the margin (or decision boundary) the new data will fall.

Benefits and Disadvantages of SVM on Wine Quality Data:

In addition to performing linear classification, SVM can also be used for non-linear classification, by means of the kernel trick. A kernel trick takes the data and transforms it in a complex way to compute a much more optimal and possibly non-linear hyperplane. Kernels allow SVM a high level of flexibility. SVM also has the ability to deliver a unique solution and handles high dimensional data well. Because the underlying structure of the wine quality dataset is not linear, the SVM will be beneficial in its non-linear modifications.

Nonetheless, SVM is not optimal for non-separable classes, since SVM fundmentally attempts to separate data into classes. In this way, predictions of wine quality will likely suffer if classes "Good" and "Bad" are not easily separable by a margin. The following graphs show the non-separability of the data:

```
##  Bad Good
##    1   10
```



By comparing the graphs for Prediction Data and Test Data, it is clear that most of the misclassifications occur near the overlap, as the classes are not separable.

SVM is also susceptible to overfitting and training isues, depending on the kernel. For this particular data set, we will be using a radial kernel to accommodate unusual shapes in the hyperspace.

For our particular data set, we can also see that our classes are heavily imbalanced, with there being significantly more "Bad" classes than "Good." In order to address this imbalance, which could lead to bias in the SVM analysis, we weighted our classes so as to allow a greater penalty for missing "Good" wine. By doing this, the result should ensure that more of the "Good" wines are actually classified as "Good" by this model.

```
##  Bad Good
##    1   10
```

We get the following confusion matrix for SVM, which shows that we are more likely to get "Bad" classes than "Good" classes in our test data:

	Bad	Good
Bad	268	20
Good	8	23

labels	values
Test Classification Error Rate	0.0877743
Accuracy	0.9122257
Precision	0.7419355
Recall	0.5348837

Results

To summarize the results of each model, it is clear that some models perform very poorly, while others do quite well in predicting wine quality. The worst performing model, the classification tree, was unable to produce a prediction for a "Good" wine in that it only produced prediction of class "Bad." For this reason, the classification tree failed to help answer the main question of the project: Can we predict a "Good" wine?

The other models each have their benefits and disadvantages. A comparison of the accuracy, precision, recall and test classification error rate is as follows:

Table 12: Wine Quality Prediction Results

Measure	KNN	Random Forest	SVM	Logistic Regression
Accuracy	0.8526646	0.9247649	0.9122257	0.8589342
Precision	0.2500000	0.7878788	0.7419355	0.4583333
Recall	0.0465116	0.6046512	0.5348837	0.2558140
Test Classification Error Rate	0.1473354	0.0752351	0.0877743	0.1410658

From this comparison, the Randon Forest model is the best in answering the question of interest. Not only does it have the highest accuracy, but this model also had a relatively high precision and recall, compared to the other models. This is important for our project because we are not just interested in classifying "Good" wines as "Good," but we are also particularly careful in trying to make sure we capture as many "Good" wines as possible in the prediction.

Conclusion

After analyzing the results of all the models we surveyed, we found that the most important predictive variables are alcohol, sulphates, and volatile acidity. For example, these three variables showed the highest mean decrease accuracies and nearly the highest Gini indices according to our Random Forest model. Additionally, all three variables were determined to be highly significant predictors according to a test for significance in logistic regression. Thus, this model-based conclusion regarding the most significant predictors supports the results of our original data exploration, which used visualizations and Cliff's Delta.

In particular, our analyses have shown that higher alcohol content and higher sulphate content tend to be associated with good wine quality, while a higher level of volatile acidity tends to be associated with bad wine quality. In fact the results of logistic regression suggest that an increase of a single standard deviation in alcohol content increases a wine's odds of being good by 194%. Additionally an increase of one standard deviation in volatile acidity decreases a wine's odds of being good by 51%, and an increase in one standard deviation of sulphate levels increases a wine's odds of being good by 81%. These are useful and actionable results to a wine brewery.

A survey of the different model performances shows that Random Forest, compared with all other models, is most effective at predicting whether red wine is excellent or average. In particular, Random Forest offered a desirable balance between precision and recall, with a precision of 0.788 and recall of 0.605, while overall

accuracy stood at 0.925. This strongly suggests that good wine and bad wine represent distinct partitions of the hyperspace spanned by chemical characteristics.