|  |  |
| --- | --- |
| Outline of wine glass and bottle in dark background  Final Project:IST 707  **Effect of Different White Wine Attribute on the Wine Quality** | *Elizabeth Westbrook, Omar Hanif* |

Contents

[Introduction 3](#_Toc99728288)

[Analysis and Models 4](#_Toc99728289)

[About the Data: 4](#_Toc99728290)

[Exploratory Data Analysis 4](#_Toc99728291)

[Models/Methods: 5](#_Toc99728292)

[Decision Trees: 6](#_Toc99728293)

[K Nearest Neighbor: 6](#_Toc99728294)

[Multinomial Naïve Bayes: 7](#_Toc99728295)

[Random Forest: 7](#_Toc99728296)

[Gradient Boosting Model Results: 7](#_Toc99728297)

[SVM Linear Model: 8](#_Toc99728298)

[SVM with Radial Basis Function Kernel: 8](#_Toc99728299)

[SVM with Polynomial Kernel: 9](#_Toc99728300)

[Comparison of Models' Performance: 9](#_Toc99728301)

[Conclusions 10](#_Toc99728302)

[Appendix 12](#_Toc99728303)

[Visualizations for EDA 12](#_Toc99728304)

[Correlation Matrix 12](#_Toc99728305)

[Discussion: Correlation among variables 12](#_Toc99728306)

[Individual Attribute Exploration 12](#_Toc99728307)

[Boxplots of variable correlation with quality 13](#_Toc99728308)

[Acidity: Fixed Acidity, Volatile Acidity, Density, Residual Sugar Plus Quality 13](#_Toc99728309)

[Citric Acid, Chlorides, Free Sulphur Dioxide, Total Sulphur Dioxide Plus Quality 14](#_Toc99728310)

[pH, Sulphates, Alcohol Content 14](#_Toc99728311)

[Methods and Models 15](#_Toc99728312)

[set train and test sets 15](#_Toc99728313)

[Visualizations of Each Model 15](#_Toc99728314)

[Decision Tree 15](#_Toc99728315)

[Decision Tree #1 16](#_Toc99728316)

[Decision Tree #2 17](#_Toc99728317)

[Decision Tree #3 (best performance) 18](#_Toc99728318)

[Naïve Bayes 20](#_Toc99728319)

[KNN 22](#_Toc99728320)

[Gradient Boosting Machine 23](#_Toc99728321)

[Random Forest 23](#_Toc99728322)

[SVM Linear 25](#_Toc99728323)

[SVM Radial Basis Function 26](#_Toc99728324)

[SVM Polynomial Kernel 27](#_Toc99728325)

[Model Comparison and Variable Importance 28](#_Toc99728326)

# Introduction

Wine quality certification includes physiochemical tests like determination of density, pH, alcohol quantity, fixed and volatile acidity etc. There exist numerous datasets having the physiochemical tests results and quality on the scale of 1 to 10 of wines of the "Vinho Verde" variety. Such a model can be used not only by the certification bodies but also by the wine producers to improve quality based on the physicochemical properties.

Because wine quality ratings are subjective, enologists would be well-served by having objective measures that can help determine a wine's quality, and with the quality determined, they can take steps to improve the quality. Wine quality is assessed based on much more than flavor and opinion. Many attributes are evaluated to help score a wine’s quality.

The project has two-fold objectives which can be summarized in the simplest and easiest terms as:

1. To experiment with different classification methods to see which yields the highest accuracy
2. To determine which features (variables) are the most indicative of a good quality wine

In this study, the aim is to investigate this dataset on physicochemical properties and quality ratings of white wine samples. The project looks at the 12 attributes of white wines in order to determine if the wine quality score can be predicted with any combination of those attributes. Essentially, the project aims to identify the relationship (if there exists one) of the available variables with the target variable: wine quality.

The algorithms used in this analysis are Multinomial Naïve Bayes, Support Vector Machine (SVM), Decision Trees, Random Forest, Gradient Boosting and KNN, all being highly popular algorithms in data analytics and proven to be effective at regression and classification tasks. Parameters for each algorithm are tuned, and the models will be compared to find the one with the highest accuracy.

# Analysis and Models

## About the Data:

The dataset relates to white variant of the Portuguese "Vinho Verde" wine. More details can be found by consulting Cortez et al., 2009. Due to privacy and logistic issues, only physicochemical (inputs) and sensory (the output) variables are available (e.g. there is no data about grape types, wine brand, wine selling price, etc.).

Dataset modeling can be viewed as classification or regression tasks; here, however, modeling is performed as a classification problem. The classes are ordered and not balanced (e.g. there are many more normal (not good) wines than the good ones). Data has been obtained from the following link:

<https://archive.Ics.Uci.Edu/ml/machine-learning-databases/wine-quality/>

## Exploratory Data Analysis

The data consists of 12 columns, with 4,898 rows. For data analytics, 75% of the observations were placed in the training set and the remaining 25% in the test set. 11 of the features are chemical variables (independent variables), and the other feature is wine quality (dependent variable), a subjective measure that is the median of the opinions of three wine experts. Specifically, the features are:

1. fixed acidity: most acids involved with wine or fixed or nonvolatile (do not evaporate readily)
2. volatile acidity: the amount of acetic acid in wine, which at too high of levels can lead to an unpleasant, vinegar taste
3. citric acid: found in small quantities, citric acid can add ‘freshness’ and flavor to wines
4. residual sugar: the amount of sugar remaining after fermentation stops, it’s rare to find wines with less than 1 gram/liter and wines with greater than 45 grams/liter are considered sweet
5. chlorides: the amount of salt in the wine
6. free sulfur dioxide: the free form of SO2 exists in equilibrium between molecular SO2 (as a dissolved gas) and bisulfite ion; it prevents microbial growth and the oxidation of wine
7. total sulfur dioxide: amount of free and bound forms of S02; in low concentrations, SO2 is mostly undetectable in wine, but at free SO2 concentrations over 50 ppm, SO2 becomes evident in the nose and taste of wine
8. density: the density of water is close to that of water depending on the percent alcohol and sugar content
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: a wine additive which can contribute to sulfur dioxide gas (S02) levels, which acts as an antimicrobial and antioxidant
11. alcohol: the percent alcohol content of the wine
12. quality: discrete score between 0 (worst) and 10 (best). Quality scores are subjective.

## Models/Methods:

First, to perform some exploratory data analysis, a few visualizations were created to understand the distribution of wine quality. The histogram on the left below displays the range of wine quality in the data set. Here, 6.5-7 is typically set as the minimum value for “Good” quality. It can, however, be observed that the majority of the wines (80%) have scored a quality value below 7. This is why, to facilitate the analysis, the quality varibale was discretized at the value of 7 and above resulting in a distribution shown on the right below. The next step was to check for missing values and data preparation so that the data can be transformed into a state that can be analyzed by our chosen models and algorithms and help maximize model performance.

Chart, bar chart

Description automatically generatedChart, histogram

Description automatically generatedIn order to achieve reproducible results, seed value was set at 424. The data set was split into training and test sets by utilizing 75% of the observations for training set. For each of the models used in the project, numerous parameters were repeatedly tuned. Eventually, the parameters' fine tuning was stopped upon reaching the maximum accuracy for each model.

Decision Trees: The algorithm has several arguments and parameters that can be used to improve the model performance. After sufficient trial and error, it was determined that with a ***maxdepth*** value of 9 and a ***cp*** of 0.01, the model yielded the best results for this dataset. The model also identified 'alcohol' as the most relevant variable here, which was not surprising in light of the earlier findings gained through correlation matrix.

**Results:** Upon running different DT models, different accuracies were obtained, none of which are impressive compared to other algorithms.

Model 1: Accuracy = 79.74%

Model 2: Accuracy = 80.39%

Model 3: Accuracy = 81.62%

One of the most interesting results from the Decision Tree model is that the top node–alcohol content–branches not to bad and good, but to bad and bad. This reveals the importance of alcohol content in a wine’s composition: if the alcohol content is less than 11%, 90% of wines will be bad Yet, above 11%, there’s still a 59% chance the wine will still be bad. However, above 13% alcohol, there’s a 62% chance the wine will be good.

K Nearest Neighbor: For KNN algorithm, the data was analyzed using the same resampling and preprocessing Models were run with K values of 1-20.

**Results:** The optimal model exhibited prediction accuracy of 85.54% using k = 1. Like Decision Trees, KNN also evaluated 'alcohol' as the most significant variable impacting the wine quality followed by 'density' and 'chlorides'.

Multinomial Naïve Bayes: Laplace smoothing parameters of 0-4 were tested, as well as bandwidth adjustments of 0-7. Gaussian kernel models consistently returned accuracies 5% lower than nonparametric kernels.

**Results:** In the final model, nonparametric kernel was used, the bandwidth adjustment was set to 7 (in order to allow a more flexible density estimate), and Laplace smoothing was set to 0 (though final results were identical with Laplace smoothing of 0 and 1). Performance accuracy improved from about 76% to 80.4% in the final.

Random Forest:For Random Forest Method, the resampling and preprocessing techniques remain the same : 10-fold Cross-Validation repeated 3 times with “center” and “scale.” The final value used for the model was mtry = 2. Very little tuning is required.

**Results:** This model performed better than previous models for this dataset resulting in a very high accuracy of 88.73%. The kappa value was also observed to be significantly higher at a statistic of 0.629.

Gradient Boosting Model Results: Prediction performance improved from the untrained gbm model (Accuracy : 0.8333) with the following tuning parameters: .n.trees=500,.interaction.depth=seq(5,9,by=1),.shrinkage=seq(.13,.21,by=.04), .n.minobsinnode=seq(3,9,by=3)). Final prediction accuracy came in at an impressive 88.64%.

Models were run with few trees (iterations) and with many trees (first, 50-150, then 400-700). After multiple experiments tuning with sequences of trees, n.trees (number of iterations) was set to 500, as performance was negligibly improved with n.trees above 500. The remaining parameters were narrowed to the best range of options, and best model parameters fluctuated within these ranges, yielding different "winning" parameters each time the model was run.

Interaction.depth (highest level of variable interactions) was tested from 1-12. The most accurate model used 9. Shrinkage (learning rate) was tested with parameters from 0.01 to 0.25. Sucessful models performed best between .13 and .17, with the final model using a learning rate of 0.13. Minimum observations per tree (n.minobsinnode) was tested from 2 to 12. The final model utilized n.minobsinnode = 6.

SVM Linear Model:For SVM linear, the resampling parameters were kept at 10-fold Cross-Validation repeated 3 times to ensure consistency throughout the analysis. Preprocessing was also accomplished using the 'center' and 'scale' arguments. Models were run with Cost values of 0 – 20.

**Results:** For the Linear model used, the final cost value used for the model was C = 2. This model performed the worst of all, with an accuracy of 78.35% and a kappa value of 0 for all Costs above 1. The model predicted that all wines were bad.

## SVM with Radial Basis Function Kernel:

Models were run with sequences of sigma values between 0 and 1, sequencing at intervals of 0.1. Later, cost was sequenced from 0.8-1, with intervals of 0.1. Costs of 0 – 40 were explored. Early on, costs in the mid-20’s appeared.

**Results:** In the final model, sigma is 0.9 and cost is 2; this returned a prediction accuracy of 87.34% and a kappa value of 0.584.

SVM with Polynomial Kernel:For SVM Polynomial Kernel, parameters include scale (set initially to 0.001, 0.01, 0.1, then honed to 0.02 - 0.1). Costs of 0-10 were explored, as well as setting the degree to 1 – 5.

**Results:** Results from this model were unstable and varied significantly with each trial. The final values used for the model were degree = 3, scale = 0.05, and C = 3, resulting in an accuracy value of 81.86% and a kappa value of 0.334.

# Comparison of Models' Performance:

A picture containing text, newspaper

Description automatically generatedSeveral different Machine Learning models were applied on the wines data set with an aim to generate insights and reveal any hidden trends or relationship of different variables with the target variable of 'Quality'.

A summary of the accuracies and kappa values obtained during the analysis is reproduced here which shows a comparison of the models' performance relative to each other.Chart, histogram

Description automatically generatedIt was observed that, overall, the Random Forest model performed better than all the other models employed in the project. It correctly identified the greatest number of positive class observations, thus resulting in the highest accuracy and kappa values, as is evident from the following visualization.

Through this project, the most relevant variables were ranked, the ones that affect the target variable the most. All of the algorithms recognized the alcohol content to be the most significant variable affecting wine quality. A comparison of the variables identified by the models is exhibited in the following images.

Graphical user interface, table, Word

Description automatically generatedGraphical user interface, Word

Description automatically generated

# Word Description automatically generated with medium confidenceConclusions

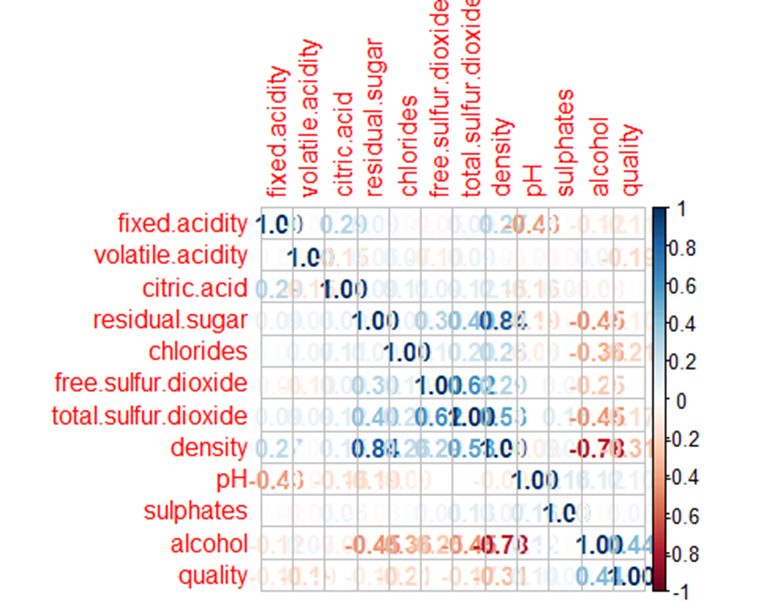
Admittedly, no single approach to machine learning and data analytics works perfectly in all scenarios. Nonetheless, through this project, we have been able to demonstrate that for this particular dataset, Random Forest Model proved to be the most accurate model. This exercise has been instrumental in reinforcing our understanding that machine learning is an incredibly useful tool that can be effectively and successfully used in the decision-making process of any business or other organization.

Although our models’ results and accuracy proved reasonably good, we could not improve the accuracy as much as we would have liked to do. Among other things, this is partially attributable to the fact that the dataset was not sufficiently balanced. Six models were explored in this project, but it seems highly appropriate also to try other machine learning algorithms like neural networks. Yet another way to enhance the models’ performance would be to have more extensive training data sets or increase the number of independent (but relevant) variables.

# Appendix

## Visualizations for EDA

### Correlation Matrix



## Discussion: Correlation among variables

The two strongest correlations with quality are moderate positive correlation to alcohol and moderate negative correlation with density.  
Additionally, Alcohol is negatively correlated with density of wine. Density is strongly positively correlated with residual sugar quantity and moderately correlated with pH. Free sulfur dioxide and total sulfur dioxide are strongly correlated.

Individual Attribute Exploration

Chart, waterfall chart

Description automatically generated

## Boxplots of variable correlation with quality

### Acidity: Fixed Acidity, Volatile Acidity, Density, Residual Sugar Plus Quality

Chart

Description automatically generated

### Citric Acid, Chlorides, Free Sulphur Dioxide, Total Sulphur Dioxide Plus Quality

Chart

Description automatically generated

### pH, Sulphates, Alcohol Content

Chart, box and whisker chart

Description automatically generated

# Methods and Models

All models are run three times with 10-fold cross validation. Preprocessing is done with each train function call. Scale and center are the methods used. The train set contains 75% of the data; the remaining 25% is the test set.

## set train and test sets

set.seed(424)  
  
train\_index <- createDataPartition(wineClass$rating, p= 0.75, list=FALSE)  
  
WC\_train <- wineClass[train\_index, ]  
WC\_test <- wineClass[-train\_index, ]  
  
table(WC\_train$rating)

##   
## bad good   
## 2879 795

table(WC\_test$rating)

##   
## bad good   
## 959 265

# Visualizations of Each Model

# Decision Tree

Decision trees are for classification problems and use training and test data. It recursively splits until a stop condition is met.Stop conditions include: when all data points belong to same class, all records have same attribute values, or model control parameters for pruning have been met.

## Decision Tree #1

Chart, line chart

Description automatically generated

Chart

Description automatically generated

A picture containing text, clock

Description automatically generated

Decision Tree #2

Chart, line chart

Description automatically generated

A picture containing text, clock

Description automatically generated

Decision Tree #3 (best performance)

Chart, line chart

Description automatically generated

Diagram

Description automatically generated

# Naïve Bayes

Chart, line chart

Description automatically generated

Chart, line chart

Description automatically generated

Chart, line chart, histogram

Description automatically generated

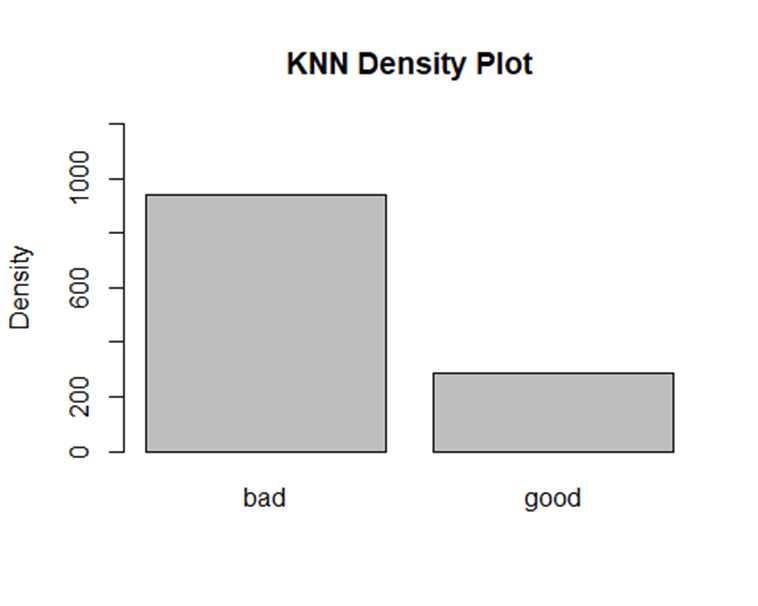
Chart

Description automatically generated

# KNN

Chart, line chart

Description automatically generated



# Gradient Boosting Machine

Chart

Description automatically generated

# Random Forest

Chart, scatter chart

Description automatically generated

Chart, line chart

Description automatically generated

Chart

Description automatically generated

# SVM Linear

Chart

Description automatically generated

# SVM Radial Basis Function

Chart, line chart

Description automatically generated

Chart

Description automatically generated

# SVM Polynomial Kernel

Chart

Description automatically generated

Chart

Description automatically generated

# Model Comparison and Variable Importance

Chart, histogram

Description automatically generated

Graphical user interface, table, Word

Description automatically generated

Graphical user interface, Word

Description automatically generated

Word

Description automatically generated with medium confidence