DATA MINING IN WINE QUALITY

Group 10:

Jennifer Zappala

Pei-Ying Liang

Shabana Ajamal Hannure

Uyen Le

Instructor
Prof Mohammadreza (Reza) Ebrahimi
University of South Florida

Fall 2021

Abstract

The main purpose of this study is to predict wine quality based on physicochemical data. In this study, The dataset was taken from Kaggle. These data sets contain 11 features of physicochemical data such as alcohol, chlorides, density, total sulfur dioxide, free sulfur dioxide, residual sugar, and pH. First, We successfully classified the quality into Low, Medium, and High where a score of 3, 4, and 5 are low quality, 6 and 7 are medium quality, and 8 and 9 are high quality. Then, the following four different data mining algorithms were used to classify the quality of wine: multiclass decision jungle, multiclass logistic regression, multiclass neural network, Multiclass Decision Forest. We have conducted descriptive statistics on the quality of wines. We also used Permutation Feature Importance to determine what are the most important attributes that decide the quality of the wines.

Introduction and Problem Specification

Today, varied consumers enjoy wine more and more. Wine industry is researching new technologies for both wine making and selling processes in order to back up this growth. Physicochemical and sensory tests are used for evaluating wine certification. The discrimination of wines is not an easy process owing to the complexity and heterogeneity of its headspace. The classification of wines is very important because of different reasons. These reasons are economic value of wine products, to protect and assure the quality of wines, to forbid adulteration of wines, and to control beverage processing. Nowadays wine is increasingly enjoyed by a wider range of consumers. Wine certification and quality assessment are key elements within this context. Certification prevents the illegal adulteration of wines (to safeguard human health) and assures quality for the wine market. Quality evaluation is often part of the certification process and can be used to improve wine making (by identifying the most influential factors) and to stratify wines such as premium brands (useful for setting prices).

Certifying the quality of food products is the major concern of the country. The citizens of the country are recommended to use only quality assured products. The same thing needs to be applied for the wine industry also. The quality of wine needs to be assessed and it should be classified into different categories based on the quality assessment. Data mining is the right approach to achieve this as it extracts the useful information by analyzing the data set. Data mining technologies have been applied to classification of wine quality. The aim of machine learning methods similar to other applications is to create models from data to predict wine quality.

With this Analysis we are going to address the following question:

1. Which classification method yields the highest accuracy when predicting wine quality?
2. How to choose a high-quality wine?

Data Characteristics

Wine is an alcoholic beverage made from fermented grapes (Mgmarques). There are many different factors that can influence the quality of wine. This is why paying attention to every detail that goes into each bottle is very important ("Fresh Vine Wine"). In the particular data set that we chose the attributes are alcohol, volatile acidity, free sulfur dioxide, residual sugar, sulphates, pH, chlorides, density, total sulfur dioxide, citric acid, fixed acidity, and quality. Of these included, alcohol, volatile acidity, free sulfur dioxide, residual sugar, sulphates, pH, chlorides, density, total sulfur dioxide, citric acid, and fixed acidity are all independent variables. Of all the attributes this only leaves quality. This is because quality of the wine is the dependent variable that all of the other variables included in the data set are influencing.

In addition to knowing the attributes included in the data set, it is also important to understand the connection between the attributes with the wine making process. Knowing what the attributes do in addition to just what they are is pivotal to the overall understanding of the data model. Some of these attributes include acidity, sweetness, salty, sulfites, alcohol, and body.

Acidity is a fundamental property of wine because it contributes to the overall taste of the wine. Wines with higher acidity tend to feel lighter-bodied, while a wine with slightly less acidity will taste more rich and round. Significantly reducing the acidity can lead a wine to tasting flat, which is why acidity can influence the overall quality of a wine. In food and drinks, acidity can often lead to a "tart" or "zesty" taste. When tasting for acidity one might feel a tingling sensation like that of pop rocks along the front and sides of the tongue, a gravelly texture along the roof of the mouth, and wetness. In the dataset acidity is measured by the following:

- Fixed Acidity: Fixed acids found in grapes such as tartaric, malic, citric, and one not found in grapes, succinic.
- Volatile Acidity: These acids are distilled out of the wine before the production process is complete. This is because an excess of volatile acids can lead to an unpleasant flavor. Volatile acids are mostly made up of acetic acid, lactic acid, formic acid, and butyric acid.
- Citric Acid: As can be seen above, this is one of the acids included as a fixed acid. It is what gives a wine its freshness.
- pH: Also known as the "potential of hydrogen" operates on a numeric scale. Solutions with a pH less than 7 are acidic, while solutions with a pH higher than 7 are basic. Most wines are acidic with a pH level between 2.9 and 3.9.

In addition to acidity, sweetness is also a factor in the wine making process. Sweetness refers to how sweet or dry a wine is. A person might discern how sweet a wine is by looking at the "legs" caused by the higher wine viscosity and a tingling sensation on the tip of the tongue. In the data set, sweetness is determined by the residual sugar.

- Residual Sugar: Refers to the natural sugar left behind after the grape fermentation process.

Another property that can influence the quality of a wine is saltiness. As we are accessing the quality of wine there are some factors that influence wine quality in a negative way, salt is one of them. Wines with the highest salinity come from countries where the vineyards are irrigated using salty water.

- Chlorides: Chloride concentration caused by salty water is a major contributor to the saltiness of wine.

Like many of the previously mentioned characteristics, sulfites are also the result of grape fermentation during the wine making process. Sulfites, also known as sulphur dioxide, help prevent the growth of undesirable yeasts and microbes, while also combating against oxidation. A few ways wine drinkers can gauge the amount of sulfur in a wine is through acidity, color, sugar content, and temperature. In the data set sulfur is measured by the following attributes:

- Sulphates: Mineral salts containing sulfur that are an essential part of a wine aroma and flavor.
- Free Sulfur Dioxide: A part of the sulfur that continues to be free after the other part binds. Too much causes a pungent “rotten egg” odor.
- Total Sulfur Dioxide: The total of the bound and free sulfur.

Wine is considered an alcoholic drink. The alcohol is formed as a result of yeast converting sugar during the fermentation process. Alcohol can be interpreted by one's taste receptors. Wines with more alcohol have a bolder body while wines with less alcohol taste lighter bodied.

- Alcohol: Measured in % volume.

As previously mentioned, another way to describe wine is by body. Body can be described as a snapshot of the overall impression of a wine. A wine being light, medium, or full bodied is a result of the many factors mentioned earlier and in the data set is measured by density.

- Density: A specific volume of wine compared to an equivalent volume of water (Mgmarques).

All of these attributes thus far have been the independent variables. The dependent variable, quality, is an evaluation by wine experts graded between 0 and 10. In our dataset we categorized these quality evaluations into Low, Medium, and High where a score of 3, 4, and 5 are low quality, 6 and 7 are medium quality, and 8 and 9 are high quality. When combined with the aforementioned variables the score will help us to establish what makes a high quality wine. We also removed the type category classifying a wine as being either red or white. This is because we did not want our results being affected by type when we were looking at quality instead.

Lastly, for our data set we chose a 70/30 Train and Test Split. This is because it is a commonly accepted split and it worked well with our data.

DM Model Construction

Look at our experiment as below, we used 4 models including multiclass decision forest, multiclass decision jungle, multiclass logistic and multiclass neural networks. Classification Multi-Class is in the example we have classified the quality of a wine with the attributes “Good” and “Bad”, in this example we will analyze the quality of a wine with numbers between 0 and 10. To do this we must use other Multi-Class classification algorithms but the principle remains the same. Deep Neural Network is the most simple neural network is the "perceptron", which, in its simplest form, consists of a single neuron. The perceptrons only work with numerical data, so, you should convert any nominal data into a numerical format.

Classification and clustering are techniques used in data mining to analyze collected data. Classification is used to label data, while clustering is used to group similar data instances together. So we are targeting classification Since regression is supervised learning. In supervised learning, the concept of adjustment consists in looking for a prediction function which, by means of the predictive attributes, makes it possible to best adjust the attribute to be predicted. Some model prediction interpretations, for this, we will be leveraging skaters and look at model predictions. We will try to interpret why the model predicted a class label and which features were influential in its decision. We needed interpretability so we could see which attributes have a greater influence on the quality of the wine. This leads us to be interested in the distribution of the output variable for the different combinations of attribute values, that is to say in the representation in the form of a contingency table of the data.

The results on the left correspond to the “Multiclass Decision Forest” classifier, the one on the right to the “Multiclass Decision Jungle” classifier. The trained model is evaluated using the source data, a comparison between the predictions and the real data is therefore carried out to measure the efficiency of the model in predicting new data sets. The “Multiclass Decision Forest” classifier provides better overall predictions with about 84% accuracy, multiclass decision jungle had 83% accuracy, multiclass logistic regression had 76% accuracy and multiclass neural network had 76% accuracy,in fact the higher the data located on the transverse, the less the model is wrong. That is why we choose “Multiclass Decision Forest” as our best model.

Metrics Multiclass Decision Forest

Overall accuracy	0.772704
Average accuracy	0.848469
Micro-averaged precision	0.772704
Macro-averaged precision	0.717503
Micro-averaged recall	0.772704
Macro-averaged recall	0.592104

Confusion Matrix

Metrics Multiclass Decision Jungle

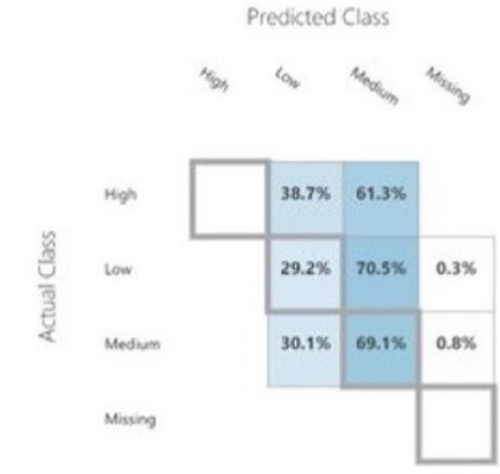
Overall accuracy	0.757311
Average accuracy	0.838208
Micro-averaged precision	0.757311
Macro-averaged precision	0.665678
Micro-averaged recall	0.757311
Macro-averaged recall	0.518271

Confusion Matrix

Metrics Muticlass Logistic Regression

Overall accuracy	0.521108
Average accuracy	0.760554
Micro-averaged precision	0.521108
Macro-averaged precision	NaN
Micro-averaged recall	0.521108
Macro-averaged recall	NaN

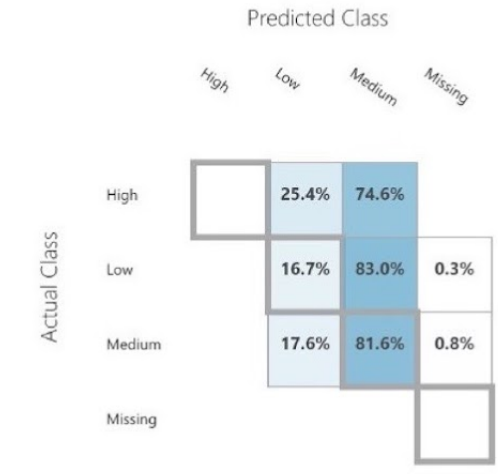
Confusion Matrix



Metrics Multiclass Neural Network

Overall accuracy	0.549033
Average accuracy	0.774516
Micro-averaged precision	0.549033
Macro-averaged precision	NaN
Micro-averaged recall	0.549033
Macro-averaged recall	NaN

Confusion Matrix



Results

label	fixed.acidity	volatile.acidity	citric.acid	residual.sugar	chlorides
Low	7.331956	0.3975464	0.3043676	5.632125	0.06445531
Medium	7.166628	0.3068285	0.3267951	5.332974	0.05170955
High	6.853299	0.2893147	0.3331472	5.372081	0.04083756

free.sulfur.dioxide	total.sulfur.dioxide	density	ph	sulphates	alcohol
29.48250	119.2123	0.9957532	3.214595	0.5240978	9.874001
30.94376	113.4504	0.9941657	3.220280	0.5364176	10.809425
34.53299	117.6954	0.9925036	3.224822	0.5119289	11.685787

Figure 1

Figure 1 above shows the mean of the attributes in our data set. We have conducted descriptive statistics on the quality of wines. The number above is the means of all the independent variables in our data set. From the summary above, there are some highlights that we found from the data set:

- The average alcohol concentration increased by around 1% at each level as the quality improved.
- The chlorides and volatile acidity are less present as well as there is no significant difference between each quality of wines.
- The free sulfur dioxide is higher with higher quality
- Higher quality has less fixed acidity

After comparison between the 4 models which are decision forest, decision jungle, logistics regression, and neural network, the best performance out of the 4 models is decision forest. We conducted a model and used Permutation Feature Importance to determine which are the most important attributes that decide the quality of the wines. We came to the conclusion that these are the key ingredients to make the best wines:

- Alcohol
- Volatile acidity
- Free sulfur dioxide

To answer the proposed questions that we stated at the beginning of the project, these are the relationship between the attributes and the quality of wines:

- The higher the alcohol, the better the quality
- While volatile acidity will decrease the quality of wine
- Higher quality wine will have higher free sulfur acidity.

These visualizations shown below describe the relationship between significant attributes and the quality of wines.

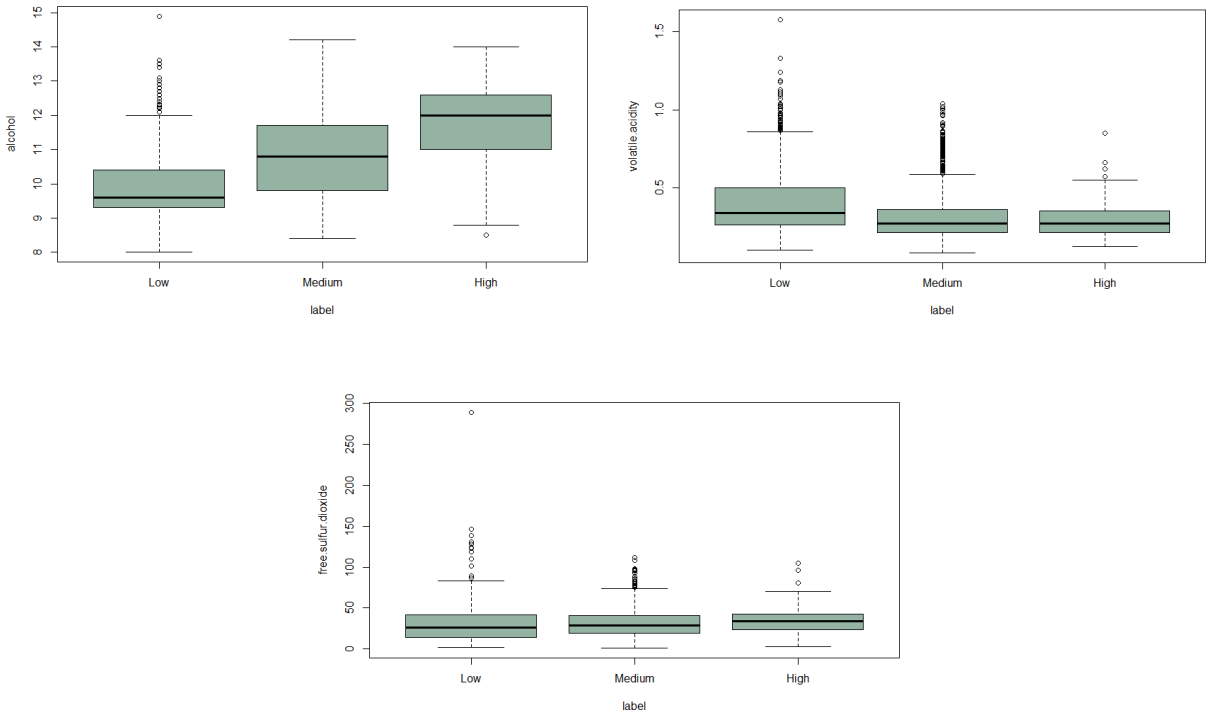


Figure 1

In comparison to wine samples with medium and low ratings, higher quality wine samples have lower levels of volatile acidity and higher levels of alcohol content. In the meanwhile, the free sulfur dioxide will make the quality of wines better
In conclusion, alcohol, volatile acidity, and free sulfur dioxide have a significant impact on the quality of wines. Higher alcohol and volatile acidity will increase the quality of wines while free sulfur acidity will decrease.

Besides 11 attributes that are presented in the data set, the quality of wines also depends on weather, temperature, growing practices, and the process of winemaking. According to Buckley Fine Wines, wines from cooler climates have more acidity but less sugar and alcohol. Hotter conditions promote ripening, resulting in wines with more sugars, alcohol, and body. So that producers that try to grow types that don't thrive in that climate will end up with a lower-quality wine. Temperatures between 60 and 70 degrees Fahrenheit are required for grapevines. However, if the weather is too hot, the grapes will mature too rapidly, reducing the amount of time it takes for flavor, color, and other compounds to fully emerge (*The 4 factors and 4 indicators of Wine Quality*). In addition to what the land and sky give, how a winemaker manipulates the vines has an impact on the quality of the finished product. Leaves and shoots are sometimes removed from the canopy to maximize sunshine exposure, while pruning removes specific branches to control

yields and keep vines healthy (*The 4 factors and 4 indicators of Wine Quality*). Another important component is harvesting, since grapes harvested too early or too late may lack the appropriate balance. Grape quality is also influenced by whether grapes are harvested manually or mechanically (*The 4 factors and 4 indicators of Wine Quality*). The winemaking process has an equal role in determining the wine's final quality. When making wine, wineries follow four key steps: maceration, fermentation, extraction, and aging, and they must maintain consistency to get the most out of their grapes. Sulfur dioxide and processing enzymes, as well as oak barrel aging and oxygen management, all contribute to the quality of wine, which ranges from superb to insipid.

Wines industry is a complicated and expensive field that people who are interested in the field need to know exactly what they need to deal with. Besides a large amount of investment (about $1 million) to get started, farm work all year round, complex paperwork as well as fatigue sale legwork, people cannot get revenue until several years of losses. By being an expert in the process of making wines along with exploring the key ingredients to produce high-quality wines, businesses can make up to 8 to 9 figures of money. Focusing on the main attributes, producers carefully weigh and adjust the amount of each attribute so that all the ingredients in each bottle can be in harmony which can lead to the key to success in the wine industry.
The attributes that make the best wine could be of interest to a CEO or winery owner. In America, many celebrities are creating their own wine labels, so it is important for them to know how to make a competitive wine for the market. Ex) Fresh Vine Wine by Julianne Hough and Nina Dobrev and Spade and Sparrows by Kaitlyn Bristowe.

Works Cited

"Fresh Vine Wine: About Us." *Fresh Vine Wine: About Us*, https://www.freshvinewine.com/about-us/.

Mgmarques. "Wines Type and Quality Classification Exercises." *Kaggle*, Kaggle, 24 Oct. 2018, https://www.kaggle.com/mgmarques/wines-type-and-quality-classification-exercises/notebook.

"Spade & Sparrows Wine by Kaitlyn Bristowe." *Spade & Sparrows*, https://spadeandsparrows.com/.

"The 4 Factors and 4 Indicators of Wine Quality." *JJ Buckley Fine Wines*, https://www.jjbuckley.com/wine-knowledge/blog/the-4-factors-and-4-indicators-of-wine-quality/1009.