The Effects of Wine Attributes on Quality Using Machine Learning

SYSEN 5880

## Data Analytics and Machine Learning (2022SP)

Daphne Duvivier | dld237

Gordon “Max” Schmidt | gms288

Ashley Turke | ajt238

Daniel Yeung | djy27

Jeet Dhoriyani | jd825

**Abstract**

…

**Introduction**

The quality of wine is important to wine producers, marketers, and consumers. Wine producers and marketers strive to supply the highly regarded wines to support their brand and ultimately generate greater revenues. Customers want to ensure they have an optimally enjoyable experience by searching for fine wines. Unfortunately, there is no clear standard or consensus between these suppliers and buyers on what constitutes a high quality wine, according to a dissertation published by S. J. Charters [1]. Charters’ research asserts that one person’s perception of a finely-crafted wine may differ from that of another person. This difference in perception may be related to the innumerable types of wine available. There are now more than 10,000 wine grape varieties throughout the world according to MasterClass [2], and there are around 11,300 wine producers in the United States alone according to The Wine Economist [3]. In the midst of thousands upon thousands of wine types and brands, how can a consumer make an effective choice when picking which wine to purchase? How can one discern a quality wine over a mediocre one?

Wine quality is important to producers in order to maximize profitability. According to the Statista Wine Report 2020, annual wine revenue was estimately to be approximately half a trillion dollars in 2021, and the wine market growth is projected to be approximately 7% annually through 2025. In just the last year, the U.S. generated a revenue of $48 on average per person. Over half of wine transactions are typically made in a bar or restaurant setting. In these settings, wine prices can normally be marked up anywhere from 200% to 300%. Pricing for restaurants that serve especially rare quality wine are typically even higher. Determining the quality of wine is critical for individuals who sell wine because of these prospects for pricing growth and, as a result, increased revenues.

What is in the wine itself that determines its quality? The pH of wine is a well-acknowledged component in any wine’s quality. A typical wine pH would range from around 3 to 4, according to the Consumer Fresh Winemakers. For reference, the most basic pH is 14 and the most acidic pH is 0. The pH of water is around 7. The reason why the pH of wine is an important aspect of wine quality is its effect on how the wine ages. A wine that is relatively more acidic (lower pH value) takes a longer time to mature than a relatively less acidic wine (higher pH value). The age of the wine can be revealed by its color, so the pH of the wine is a discernible factor that has a significant impact on how a wine is regarded.

In addition to pH, alcohol content is a key factor in determining how fine a wine is. Yeast is used in the fermentation process to change sugars into alcohol. Certain types of alcohol particles can be sensed, which contribute to adding layers of complex flavors and smells to the wine. These alcohol particles interact with acids in the wine over time during the aging process to produce an enjoyable complex flavor profile. The specific kind of wine and individual preferences tend to vary widely, so the ideal percentage of alcohol in one glass of wine may be perfect to one person, but not ideal to another. Nevertheless, according to empirical studies generated from taste testing, the target wine alcohol content is approximately 13.6% on average. Other results from these studies revealed that certain types of wine with alcohol content exceeding 14% were rated higher, so there is no hard and fast rule for an ideal alcohol concentration.

Quality of wine also depends on the environmental conditions of the vineyard, known as terroir. A change in the terroir of a wine product may result in a change of taste, which may be undesirable when a wine producer is trying to deliver a signature flavor profile. One of the aspects impacted by terroir is chloride content. The saltiness of the water used to water the grapes directly influences the chloride content. If the water used for irrigation has a higher salinity, the wine itself will have a higher chloride content. Additionally, certain kinds of grapes are more susceptible to chloride absorption, such as Syrah grapes. These chloride concentrations contained in the wine will ultimately change how a wine tastes.

Based on the discussion above, wine quality is complex, multifaceted, and not easily achieved. Historically, the various factors contributing to wine quality have been checked and fine-tuned by conducting numerous tests and experiments. These findings can be validated and expanded upon using a different approach: machine learning algorithms. Machine learning solvers, namely random forest, decision tree, AdaBoost, GBM, and XGBoost, will be used to predict the wine quality scores of several different wines. It is hypothesized that these machine learning predictions will tend to corroborate with previous findings that factors, such as pH, alcohol content, and chloride concentration, significantly affect wine quality scores. Therefore, based on the literature review above, the team hypothesizes that acidity (pH) will have the most significant effect on wine quality.

The random forest, decision tree, AdaBoost, GBM, and XGBoost solvers will be ranked according to their prediction accuracy. Ensemble machine learning algorithms, which are an amalgamation of several different models, are of particular interest since, due to their widely-acknowledged effectiveness and multifaceted nature, it may be difficult to determine which one is the best solver to implement. As less-complex models, classification machine learning algorithms are also included for comparison. Based on preliminary investigation, the team hypothesizes that the random forest solver will produce the most accurate predictions.

Two datasets will be run through machine learning solvers: a red wine dataset and a white wine dataset. The most significant factors used to predict red wine quality and white wine quality will be determined and compared. The team hypothesizes that the most significant factors used to predict wine quality for red wines and white wines will be the same.

# 

# **Background**

In recent years, machine learning has been used to analyze multivariate problems like wine classification, which dates back to ancient times. According to Wine Facts [4], the classification of wine can be traced back to the Roman Empire where Roman wine producers would catalog different methods and uses for wine. The article continues to describe how European governments would impose strict regulations on wine production in order to authenticate traditional regional wines. Classifications of wine were performed based on more readily identifiable characteristics, such as sweetness or place of origin, but with advances in chemistry, the means for identifying and measuring more hidden traits, such as sulfur dioxide, chloride, and sulfate levels. With modern computing power, machine learning has proven itself to be a useful classification tool for complex, multifaceted problems. By combining discoveries in chemistry and machine learning techniques, greater insights can be brought to light to determine what goes into producing and selecting the highest quality wines.

Many have seen this opportunity to gain greater knowledge of wine quality using modern analysis techniques, which have resulted in numerous studies of wine quality using machine learning. A study by D. Pawar et al. used logistic regression, stochastic gradient descent, support vector classifier, and random forest machine learning solvers to compare several wine attributes to its quality [5]. Another study by Hopfer et al. performed a principal component analysis on 27 wine samples and compared sensory attributes, such as the color of the wine to its quality [6]. Limionet et al. compared the results of additive logistic regression, weighted linear regression, and k-Nearest-Neighbors methods using wine quality data [7]. A study by N. Sharma on the quality of red wines using a variety of machine learning techniques revealed that certain machine learning solvers produced superior results than others [8]. Similarly, B. AG produced a similar conclusion that certain machine learning algorithms produce superior results than that of others [9].

From research conducted during the literature review, no studies could be identified that focused specifically on comparing and contrasting the quality of white and red wines using machine learning techniques. The wine quality data from the University of Irvine Machine Learning Repository will be analyzed, which includes datasets for both white and red wines [10]. The 12 features of both wine datasets along with their corresponding descriptions are as follows:

1. Fixed Acidity - acids in wine that are less prone to evaporation
2. Volatile Acidity - acids in wine that are more susceptible to evaporation
3. Citric Acidity - organic flavor-enhancing preservative
4. Residual Sugar - remaining post-fermentation grape sugar
5. Chloride Concentration - salinity of the wine
6. Free Sulfur Dioxide - antimicrobial and anti-oxidizing preservative
7. Total Sulfur Dioxide - how much sulfur dioxide (SO2) is in the wine
8. Density - amount of concentration of dissolved solids, which is directly related to wine sweetness
9. pH Level - how acidic the wine is
10. Sulfate Content - wine preservative that helps generate sulfur dioxide (SO2)
11. Alcohol Content - percentage of alcohol in the wine
12. Wine Quality Score - a rating from 0 (worst) to 10 (best) denoting the quality of the wine

The white wine dataset has 4,898 rows of data points whereas the red wine dataset has 1,599 rows of data points. The team will implement multiple ensemble and classification machine learning algorithms to determine how to choose and to ultimately predict the highest quality wines based on these features and determine how decisions are changed depending on if one is choosing a white wine or red wine.

# **Methods**

Since one of the outcomes of the evaluation is to predict the wine quality rating, supervised learning algorithms will be used for this project. The project does not involve discovering traits that are not already defined, so unsupervised algorithms will not be used. Furthermore, classification algorithms will be used since this project is looking at classifying these wines into 10 different rating groups. Ensemble learning algorithms will be used over traditional classification algorithms, since ensemble learning algorithms combine several weaker classification solvers into more powerful solvers. Several different ensemble learning algorithms will be investigated for this project.

It has been observed through investigation that bagging algorithms tend to provide the most accurate predictions for the wine quality data being used; therefore, the Random Forest Classifier and Decision Tree Regressor have been shown to provide the best models from initial investigations. Further investigation with this data will be conducted to decide which specific variables have the biggest impact on the final predictions. Initial boosting algorithm investigations have shown that Gradient Boost has performed the best with the data provided as compared to other boost algorithms evaluated.

Random Forest Classifier uses solely classification while Decision Tree Regression employs a combination of classification and regression for its solving; therefore, both options are being explored along with additional boosting algorithms. As the dataset contains continuous values for most of the data, that is what will be used as opposed to abstracting all the data to discrete variables as this may lead to less precise predictions.

Initially all included variables will be considered in order to find initial predictions. Subsequently, specific sets of variables will be explored to find out which combinations of these are the most influential in arriving at the most accurate predictions using the established model. This information will lead to a better understanding of what factors of wine quality are the most important to consider when predicting the quality of the resulting wine.

The combinations of variables will be created using prior knowledge of the attributes important to wine creation and selection by consumers, and this combination will serve as the hypotheses to this proposal. These combinations will include the following groups:

* Fermentation: Residual Sugar, Density, Alcohol Content, Chloride
* Acidity: Fixed Acidity, Volatile Acidity, Citric Acidity, pH
* Sulfur Dioxide: Sulfates, Free Sulfur Dioxide, Total Sulfur Dioxide

Evaluation using these variable combinations will allow the team to determine which attributes of wine are the most influential in deciding the quality of the wine; thus, helping to inform consumers what to specifically look for in a wine. In order to do this, machine learning solvers will be used with the different groups of variables included as features and the accuracy of the predictions will be assessed to judge which are the best at reaching the correct wine quality for the data being input.

Overall, the solvers used during this project will utilize multiple machine learning algorithms to predict the wine quality score for ten data points with a mix of both white and red, and to evaluate these predicted quality values to the actual wine quality values. Another objective of this project would be to determine which group of attributes is the most influential in deciding the quality of a wine in the dataset. The final high level objective of this project will be to rank five different machine learning algorithms based on the error observed when using each algorithm. The algorithms being evaluated include Random Forest, Decision Tree, Gradient Boost, AdaBoost, and XGBoost. The algorithm with the least error will be chosen as the best method for the dataset.

# Results and Discussion

The data was split up into a training dataset and a testing dataset. The white wine dataset has 4,898 data points and the red wine dataset has 1,599 data points. The team used 200 data points for validating the machine learning models and the rest (4,698 white wine data points and 1,399 red wine data points) for training the models.

The mean squared error (MSE) was used to compare the performance of the different models that were generated since the MSE is a proven statistical method for measuring the error between predicted values and actual values in a dataset. The MSE was calculated using the following formula:

where…

MSE = mean squared error of the model

N = Number of data points in our test dataset (200 data points in this case)

= Actual wine quality scores

= Predicted wine quality scores

The team ran both the red wine and white wine data through five machine learning solvers. The solvers predicted the wine quality score for each data point. The predictions were then compared to the actual wine quality score by calculating the mean squared error (MSE). The solvers were then ordered by MSE values from least to greatest. The solver with the lowest MSE is deemed the best of the five solvers. See the table below.

*Table 1: Solvers sorted by MSE*

| **Solver** | **MSE for red wine data** | **Rank from Low to High** | **MSE for white wine data** | **Rank from Low to High** |
| --- | --- | --- | --- | --- |
| Random Forest Classifier | 1.075 | 2 | 0.95 | 1 |
| Decision Tree Regressor | 1.385 | 4 | 1.11 | 5 |
| AdaBoost Classifier | 1.14 | 3 | 1.04 | 4 |
| GBM Classifier | 1.41 | 5 | 0.98 | 3 |
| XGBoost Classifier | 1.07 | 1 | 0.965 | 2 |

As shown in Table 1 above, the XGBoost Classifier exhibited the best performance of all the solvers for the red wine dataset with the lowest overall MSE, but was closely followed by the Random Forest Classifier. For the white wine dataset, the Random Forest Classifier performed the best with the lowest MSE and was closely followed by the XGBoost Classifier. Hence, these two algorithms would be recommended for this particular dataset. The worst (least accurate) solvers were the Gradient Boosting Machine Classifier for the red wine dataset and the decision tree classifier for the white wine dataset as these two solvers had the highest MSE. It is noted that the white wine dataset exhibited lower MSEs across all five different solvers than that of red wine. This result is expected since the white wine dataset had more than 3000 more data points than the red wine dataset, so the white wine models were trained on significantly more data.

The team initially hypothesized that the Random Forest Classifier solver would produce the most accurate predictions. Based on Table 1 above, the random forest solver was the best machine learning algorithm for white wine and was a close second for red wine. The XGBoost algorithm outperformed the Random Forest Classifier only slightly for red wine (difference in MSE was only 0.005). For white wine, the XGBoost Classifier was only slightly worse than the Random Forest Classifier (difference in MSE was only 0.015) Hence, the Random Forest Classifier and XGBoost Classifier are both excellent choices when deciding which solver to run for either dataset. Hence, the results were very close to the team’s initial hypothesis.

Each attribute was then assigned to three overall groupings: fermentation, acidity, and sulfur dioxide. A model was built based on each grouping. The predictions of each model were then compared to the actual wine quality score by calculating the MSE. The solvers were then ordered by MSE values from least to greatest. The model with the lowest MSE is deemed the most significant group of attributes for predicting the wine quality score. See the table below.

*Table 2: Attribute groupings sorted by MSE*

| **Grouping** | **MSE for red wine data** | **Rank from Low to High** | **MSE for white wine data** | **Rank from Low to High** |
| --- | --- | --- | --- | --- |
| Fermentation | 0.98 | 3 | 1.12 | 3 |
| Acidity | 0.88 | 2 | 1.02 | 1 |
| Sulfur Dioxide | 0.75 | 1 | 1.03 | 2 |

As shown in Table 2 above, the fermentation group of attributes resulted in the highest MSE values for both datasets; hence, machine learning models based on the fermentation group of attributes performed the worst. The sulfur dioxide group outperformed the fermentation and acidity groups for red wine. For white wine, the acidity group outperformed the fermentation and sulfur dioxide groups.

The team initially hypothesized that acidity (pH) will have the most significant effect on wine quality. Based on Table 2 above, this hypothesis was confirmed for the white wine data, but was rejected for the red wine data. Although the acidity performed second overall for the red wine dataset, sulfur dioxide was determined to be the most important grouping of attributes impacting wine quality for red wine. Therefore, the team’s initial hypothesis was partially confirmed.

Further analysis of the different wine attributes led to the analysis of the individual attributes. Models were run based on single attributes for the red and white wine datasets. The MSEs were calculated, ranked, and compared to the correlation coefficient. The correlation coefficient is a ubiquitous measure of the relationship between two variables. The correlation coefficients were calculated using the following formula:

where…

CC = Correlation coefficient

N = Number of data points in our test dataset (200 data points in this case)

= Actual wine attribute values

= Average (mean) of the wine attribute values

= Actual wine quality scores

= Average (mean) of the wine quality scores

The MSEs and correlation coefficients were compared to validate the findings. In Table 3 below, one can see the MSE and correlation coefficients of each attribute side-by-side.

*Table 3: Attributes’ MSE and Correlation Coefficients*

| **Grouping** | **MSE for Red Wine Data** | **Rank from Low to High** | **MSE for White Wine Data** | **Rank from Low to High** | **Correlation Coefficient for Red Wine** | **Rank from High to Low** | **Correlation Coefficient for White Wine** | **Rank from High to Low** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Fixed Acidity | 1.065 | 8 | 0.66 | 1 | 0.12 | 8 | 0.11 | 6 |
| Volatile Acidity | 1.03 | 6 | 0.78 | 4 | 0.39 | 2 | 0.19 | 4 |
| Citric Acid | 0.845 | 2 | 0.7 | 3 | 0.23 | 4 | 0.0092 | 10 |
| Residual Sugar | 1.005 | 4 | 0.815 | 6 | 0.014 | 11 | 0.098 | 8 |
| Chlorides | 1.145 | 10 | 0.815 | 6 | 0.13 | 7 | 0.21 | 3 |
| Free Sulfur Dioxide | 0.905 | 3 | 0.78 | 4 | 0.051 | 10 | 0.0082 | 11 |
| Total Sulfur Dioxide | 0.835 | 1 | 0.9 | 8 | 0.19 | 5 | 0.17 | 5 |
| Density | 1.325 | 11 | 1.465 | 9 | 0.17 | 6 | 0.31 | 2 |
| pH | 1.1 | 9 | 0.675 | 2 | 0.058 | 9 | 0.099 | 7 |
| Sulphates | 1.045 | 7 | 0.79 | 5 | 0.25 | 3 | 0.054 | 9 |
| Alcohol | 1.02 | 5 | 0.875 | 7 | 0.48 | 1 | 0.44 | 1 |

The team initially hypothesized that the most significant factors used to predict wine quality for red wines and white wines will be the same. Based on table 3 above, The two wines varied significantly with regards to what factors played a factor in their rating. Both results show that the density attribute resulted in one of the highest MSEs of all the attributes; however, red wine appears to be most affected by sulfur dioxide whereas white wine appears to be most affected by acidity. In comparing the MSEs and correlation coefficients for red wine and white wine, it can be seen that the statistical measure changes the rankings. The MSE is taken to be the preferred statistical measure since it directly compares the predicted outcomes to the actual outcomes. The correlation coefficient was computed by comparing the data points to their averages, so it is not as precise a measure for determining an attribute’s relationship with the wine quality score.

# Conclusion and Recommendations

In conclusion, the team’s initial hypothesis on which solver would perform the best was mostly correct. Random forest was hypothesized to be the best solver. Based on the results, random forest and XGBoost appear to provide the best predictions on wine quality as opposed to other algorithms. Random forest was used for the individual attribute investigations since it performed just slightly better than XGBoost. Nevertheless, bagging algorithms were not determined to be dominant over boosting algorithms, since decision tree regression seemed to perform as one of the bottom three regressors in both categories. The other boosting methods varied between the two wine types. Additional research is needed to discover why these algorithms varied.

White wine quality was found to respond most to acidity changes. Understanding the flavor and composition of a white wine, it can be assumed that this is because white wine is known to be more acidic in flavor than red wine. White wines are known for “brighter” flavors, so acidity plays a role in getting this. More research is needed to understand whether higher or lower acidity affects white wine quality positively or negatively. Red wine, on the other hand, responds more to the sulfur dioxide content. Natural wines are known to have lower sulfur dioxide content, and are sometimes preferred to other wines with a higher content. If the wines used in this project were to have data on whether they are considered “natural” or not, more research can be done to include this information in the model.

A problem that came up during this project was with the boosting algorithms producing just a “6” as the answer for each test value, especially with the white wine dataset. Initially, only the last 100 data points were going to be used, but that was increased to 200 points when it was found that the increase in data size helps ameliorate these effects. Furthermore, defining stricter parameters also led to just output “6” each time, so the solvers were left on their default settings. In the end, the SVR algorithm was unable to be fixed, and thus the results can be mostly ignored there. Additional work can be done to figure out why this kept happening.

# Copy of Python Code

<insert Python Code here>

# 

# References

[1] S. J. Charters, “Perceptions of wine quality,” Edith Cowan University. [Online]. Available:<https://ro.ecu.edu.au/cgi/viewcontent.cgi?article=1115&context=theses>. [Accessed: Mar. 28, 2022].

[2] MasterClass staff, “What Are the Different Types of Wine Grapes? A Guide to the Various Types of Red and White Wine Grapes in the World,” MasterClass. [Online]. Available: <https://www.masterclass.com/articles/what-are-the-different-types-of-wine-grapes-a-guide-to-the-various-types-of-red-and-white-wine-grapes-in-the-world#:~:text=There%20are%20now%20more%20than,on%20where%20they%20are%20grown>. [Accessed: Mar. 28, 2022].

[3] M. Veseth, “Category Archives: wine brands, Has U.S. Wine Industry Consolidation Gone Too Far?,” The Wine Economist. [Online]. Available: <https://wineeconomist.com/category/wine-brands/>. [Accessed: Mar. 28, 2022].

[4] “Wine Report 2020 Statista Consumer Market Outlook - Segment Report”, Statista, 2020. [Online] Available from: <https://www.statista.com/study/48818/wine-report/>. [Accessed: Mar. 30, 2022]

[5] R. Casertano, “pH & Wine Quality”, Consumer Fresh Winemakers, 2017. [Online] Available from: <https://www.cfpwinemakers.com/blog/post/?pH-Wine-Quality-30>. [Accessed: Mar. 30, 2022]

[6] M. S. Coli, A. G. P. Rangel, E. S. Souza, M. F. Oliveira, A. C. N. Chiaradia, “Chloride concentration in red wines: influence of terroir and grape type”, Food Science and Technology, Campinas, 2015, vol. 35, n. 1 , pp. 95-99, ISSN 1678-457X. [Online] Available from: <https://doi.org/10.1590/1678-457X.6493>. [Accessed: Mar. 30, 2022]

[7] “Classification of Wine,” Wine Facts. [Online]. Available: <http://www.wine-facts.net/facts-about-wine/classification-of-wine/>. [Accessed: Mar. 28, 2022].

[8] D. Pawar, A. Mahajan, and S. Bhoithe, “Wine Quality Prediction using Machine Learning Algorithms,” International Journal of Computer Applications Technology and Research, vol. 8, issue 9. [Online]. Available:<https://ijcat.com/archieve/volume8/issue9/ijcatr08091010.pdf>. [Accessed: Mar. 28, 2022].

[9] H. Hopfer, J. Nelson, S. E. Ebeler, H. Heymann, “Correlating Wine Quality Indicators to Chemical and Sensory Measurements,” Multidisciplinary Digital Publishing Institute. [Online]. Available:<https://www.mdpi.com/1420-3049/20/5/8453/pdf>. [Accessed: Mar. 28, 2022].

[10] A. Lemionet, Y. Liu, and Z. Zhou, “Predicting quality of wine based on chemical attributes,” Stanford University. [Online]. Available:<https://cs229.stanford.edu/proj2015/245_report.pdf>. [Accessed: Mar. 28, 2022].

[11] N. Sharma, “Quality Prediction of Red Wine based on Different Feature Sets Using Machine Learning Techniques,” International Journal of Science and Research (IJSR), vol. 9, issue 7. [Online]. Available:<https://www.ijsr.net/archive/v9i7/SR20718002904.pdf>. [Accessed: Mar. 28, 2022].

[12] B. AG, “Wine Quality Prediction Using Different Machine Learning Techniques,” International Journal of Science, Engineering and Technology. [Online]. Available:<https://www.ijset.in/wp-content/uploads/IJSET_V8_issue4_231.pdf>. [Accessed: Mar. 28, 2022].

[13] “Wine Quality Data Set,” University of California Irvine Machine Learning Repository. [Online]. Available:<https://archive.ics.uci.edu/ml/datasets/Wine+Quality>. [Accessed: Mar. 28, 2022].