Predicting Wine Quality

December 18, 2024

Abstract:

Due to the complexities of wine quality ratings, an analysis was conducted to determine what physicochemical elements are significant in predicting wine quality. The analysis was conducted on a Vinho Verde wine dataset. Out of the twelve variables considered, eight were found to be significant. Further analysis of these predictors revealed relationships that are helpful for winemakers to improve their wine quality ratings.

Introduction:

The quality of wine is a multifaceted attribute that involves a range of sensory, chemical, and environmental factors, making it a topic of subjective evaluation. While it has traditionally been assessed by sensory characteristics such as color, aroma, taste, and mouthfeel, recently more scientific measures have been used to assess the underlying factors that contribute to these sensory attributes. In order to understand the odor contribution of wine components, for example, researchers look at the biosynthetic pathways and interactions between aromatic compounds present in wine. This involves complex techniques such as Gas Chromatography Olfactometry (GC–O) and sensory evaluation (Ebeler, 2001). The GC–O technique involves the combination of traditional gas chromatography with human olfactory assessment, enabling the accurate identification of compounds that have an impactful sensory component.

It has also been found that the wine-making process, specifically the brewing environment, largely impacts the flavor of wine. For example, brewing in a low-temperature environment can change the process of microbial evolution to achieve greater aroma (Shi et al., 2022) and storing wine in glass bodies for a certain period can reduce bitterness to improve flavor (Echave et al., 2021). There are also measures that combine both aroma and flavor – and other factors to give wine a certain "grade" or ranking, as a way to standardize wine tasting and consumption. Such measures include the 100-point scale, where wines in the upper echelons of quality are assigned a grade from 96 to 100, versus bottles in the 80-84 range, which indicates a suitable bottle of wine for casual consumption, but do not carry the allure of higher-ranked types (Puckette, 2015). This scale, however, can be subjective; one critic may assign a high grade to a wine that another critic did not like. Wine producers need an objective method for classifying their wine, as wine classification can be used to preserve the integrity of their production and justify prices. Through analysis of chemical factors such as pH, acidity, alcohol levels, and density, among others, this report will help producers combine the scientific research with the tried and tested point-ranking method to develop an objective wine classification method.

Methods:

To answer this research question, data was obtained from Kaggle (Yasser, 2021). In particular, the data to be used was uploaded by M. Yasser, H, a machine-learning researcher and cloud engineer, likely for the purpose of assisting winery owners and wine connoisseurs in predicting the quality of different wines they create, further helping them pinpoint pricing for selling purposes. He uploaded it three years ago, which likely has little to no significance on the quality of the data, as the wine industry has adopted new technologies like optical berry sorting and micro-oxygenation in recent years to refine traditional methods, but these changes have largely focused on enhancing efficiency rather than overhauling established practices (Sullivan, 2023) . While the uploader is originally from India, the data comes from the Minho region in Portugal. Because the quality and taste of wine are largely impacted by the origins of the grapes that produce it, the geographical region of the entries is a major factor to consider. Specifically, each entry in the dataset is related to the red variants of the Portuguese "Vinho Verde" wine, which translates literally to "green wine" (Costa, 2022). However, the name is not due to the color of the wine but rather the fact that the wine is typically consumed three to six months after harvesting, making it young in comparison to other, similar wines.

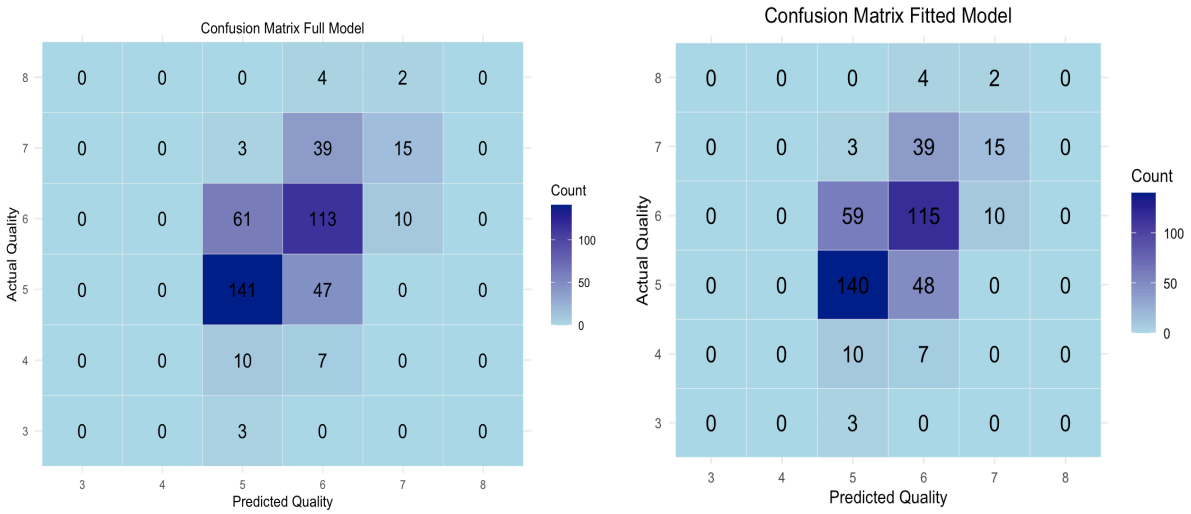
The Wine Quality dataset contains twelve variables to predict the quality of Vinho Verde wines: Fixed Acidity, Volatile Acidity, Citric Acid, Residual Sugar, Chlorides, Free Sulfur Dioxide, Total Sulfur Dioxide, Density, pH, Sulphates, Alcohol, and Quality. The response variable in this dataset was a rating of the wine quality from levels three to eight. The data was then partitioned into a training and test data set to measure model accuracy.

Result:

A proportional odds model was implemented using a vector-generalized linear model with a parallel assumption for this data. This model was used to account for the ordinal nature of the response

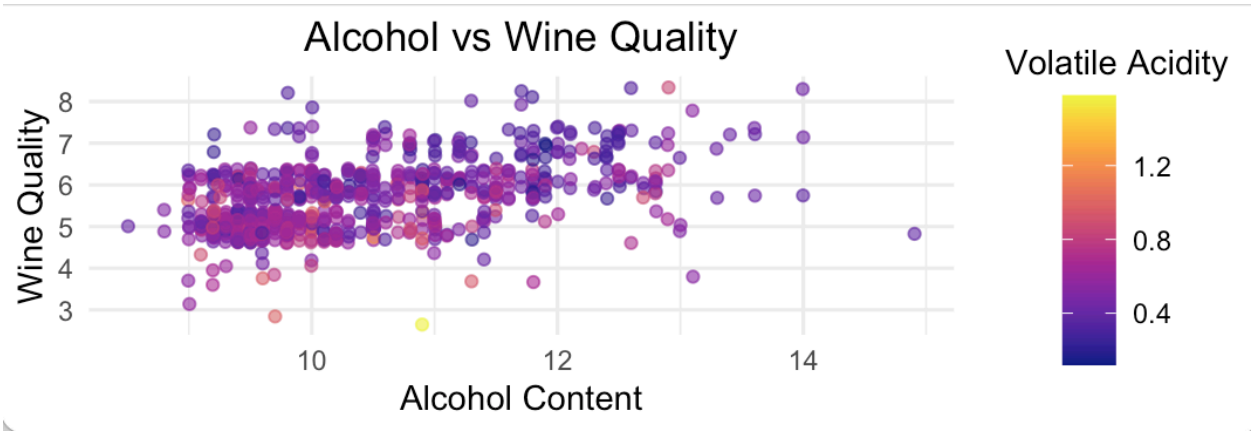
variable. The purpose of using this model was to determine which variables were significant in predicting wine quality. The original model was a full model that was fitted with all available predictor variables such as fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, and alcohol.

In order to perform variable selection and find the best model fit, the AIC method was applied to the full model. The outcome resulted in a fitted model which included fixed acidity, volatile acidity, chlorides, free sulfur dioxide, total sulfur dioxide, density, sulphates, and alcohol. Both the saturated model and the reduced model were trained on the training data and used to predict wine quality for the test data. The models had very similar accuracies with the full model resulting in a 59.34% accuracy and the fitted model resulting in a 59.12% accuracy. The similarities in the model accuracies are further shown in the confusion matrices for both models. The confusion matrices for both matrices are identical, showing that the fitted model successfully preserves predictive power while reducing model complexity.

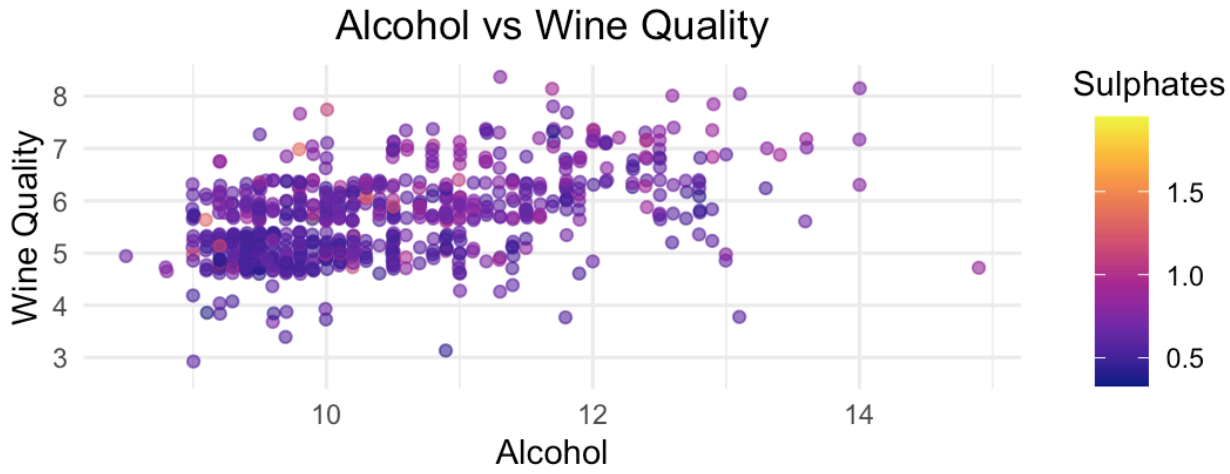


While the model accuracies and confusion matrices indicate that the reduced model maintains comparable performance to the full model, it is also important to understand which predictors contributed most to the predictions. Out of the eight predictor variables in the fitted model, the three variables that were most significant for predicting wine quality were volatile acidity, sulphates, and alcohol. Exploring further how these predictors affect wine quality, it was found that high levels of volatile acidity are generally associated with lower wine quality levels, whereas high alcohol and sulphate levels are associated with high wine quality.

According to the Australian Wine Research Institute, volatile acidity is used to measure the poorness of wine quality (Australian Wine Research Institute, 2023). High levels of volatile acidity can produce a vinegar-like taste and odor, disrupting the flavor profile of the wine. Additionally, volatile acidity levels are used to measure wine spoilage. Elevated volatile acidity levels indicate spoilage or bacterial contamination. When bacterial contamination is present, acetic acid is produced, resulting in higher levels of volatile acidity. Wine experts recommend a volatile acidity level between 0.3 and 0.8 gallons per liter. This relationship was confirmed by a scatter plot, which showed that lower levels of acidity are associated with higher wine quality ratings.



While high levels of volatile acidity are found to reduce wine quality ratings, sulphate levels and alcohol content were found to have the opposite effect. A scatter plot comparing alcohol, wine quality, and sulphate levels was produced to show this interaction. Observations with higher wine quality ratings had both higher alcohol and sulphate levels. High alcohol levels are known to produce intense wines with full body flavors, often resulting in higher wine quality ratings (Cellaraiders, 2020). Similarly, sulphates produce sulfur dioxide which acts as an antioxidant and preservative (OIV, 2021). The antioxidant properties of sulphates help preserve wine aroma compounds and its preservative properties help reduce the risk of bacteria development.



Conclusion:

The results of this analysis suggest that fixed acidity, volatile acidity, chlorides, free sulfur dioxide, total sulfur dioxide, density, sulphates, and alcohol are physicochemical elements of wine that are helpful in predicting wine quality. However, other elements such as citric acid, residual sugar, and pH levels did not have a significant impact on predicting wine quality and were removed from the model. This is likely due to the presence of correlated variables sharing explanatory power. Even without these variables, the model maintained similar predictive accuracy.

The most significant physicochemical wine elements included volatile acidity, alcohol, and sulphates. Based on the analysis of these variables, it is recommended that wine professionals monitor the

volatile acidity, alcohol, and sulphate levels of their wines. Specifically, they should try to maintain lower volatile acidity levels, and higher alcohol and sulphate levels to improve the chances of a high quality rating for their wines. However, it should also be noted that the next most significant predictors of wine quality included fixed acidity, chlorides, and total sulfur dioxide, so these elements should not be ignored.

While this analysis led to helpful insights, further research on wine quality predictions would be helpful. The M Yasser H wine quality dataset lacked observations for low wine quality ratings. Although the wine quality rating scale was from one to five, the dataset only included rankings starting at three and did not include many observations within those rankings. Running the model on data that contains more observations for the different quality levels would be helpful in gaining more insights for predicting wine quality. Additionally, it is important to consider that wine quality may have a different meaning depending on wine type. For further analysis, it may be helpful to look at wine quality datasets by wine type, since they are likely to require different physicochemical element levels to achieve a certain ranking compared to other wine types. Ultimately, while this study highlights critical elements influencing wine quality, addressing data limitations and wine type variability in future research will strengthen predictive models and enhance winemaking strategies tailored to diverse consumer preferences.

References

Dataset:

M Yasser, H. (2021). Wine Quality Dataset. Kaggle. Retrieved September 16, 2024 from https://www.kaggle.com/datasets/yasserh/wine-quality-dataset/data.

Sources:

Ebeler, S. E. (2001). Analytical chemistry: Unlocking the secrets of wine flavor. Food Reviews International, 17(1), 45–64. https://doi.org/10.1081/fri-100000517

Echave, J., Barral, M., Fraga-Corral, M., Prieto, M. A., & Simal-Gandara, J (2021). Bottle aging and storage of wines: A review. Molecules, 26(3), 713. https://doi.org/10.3390/molecules26030713

How Alcohol Content Affects Wine. Cellaraiders. (2020, March 30). https://cellaraiders.com/blogs/news/abv-of-wine?srsltid=AfmBOoogWEQNpp-ioyrz7CzUSarhhR7iO6lp3CFTgdxhmTeuqrGHzoxJ

International Organisation of Vine and Wine. (2021, March). SO2 and Wine: A Review. https://www.oiv.int/public/medias/7840/oiv-collective-expertise-document-so2-and-wine-a-review.pdf

Puckette, M. (2015). *A Pragmatic Approach to Using Wine Ratings*. How Wine Ratings Work. https://winefolly.com/tips/wine-ratings-explained/

Shi, X., Liu, Y., Ma, Q., Wang, J., Luo, J., Suo, R., & Sun, J. (2022). Effects of low temperature on the dynamics of volatile compounds and their correlation with the microbial succession during the fermentation of Longyan wine. LWT, 154, Article 112661. https://doi.org/10.1016/j.lwt.2021.112661

Top Winemaking Innovations. Wine Enthusiast. (2023, May 5). https://www.wineenthusiast.com/culture/wine/science-and-the-future-of-winemaking/

U.S. Department of the Treasury. (2024, May 1). *American Viticultural Area (AVA)*. Alcohol and Tobacco Tax and Trade Bureau. https://www.ttb.gov/wine/american-viticultural-area-ava#:~:text=An%20American%20Viticultural%20Area%2C%20or,regulations%20at%2027%20CFR%204.25%20.

Wine flavours, faults and Taints. The Australian Wine Research Institute. (2023, September 29). https://www.awri.com.au/industry_support/winemaking_resources/sensory_assessment/recognition-of-wine-faults-and-taints/wine_faults/