BELLEVUE UNIVERISTY

Wine Tasting Predictions

Christine Hathaway
5/24/2020

Abstract

Wine comes in different colors and flavors. But to truly appreciate it, it's important to understand the characteristics that different grapes offer, and how they are expressed in wines (Laube & Molesworth, 1996). Wine characteristics are expressed in five ways: sweetness, acidity, tannin, alcohol, and body. It is important to understand these basic characteristics in order to learn how to taste wine.

This project will look at the varietal characteristics of wine and how they impact how it tastes. Using two data sets for white and red wines, it will examine the characteristics of each wine, as well as the quality rating each receives from a panel of wine experts. Based on these data sets, this study will attempt to answer if it is possible to predict the quality of a wine based on the attributes that make up the measurable, physical characteristics of wine

Intro/Background of the Problem

It seems every day there is a new wine on the market. Traveling across the United States, one can see that more non-traditional places are producing wine. Most people are familiar with European wines and champagnes, while in the U.S., California is usually the first place to come to mind when talking of vineyards. But even states like Iowa are now producing wine, with vineyards springing up in the Loess Hills, the Amanas, and even in the northeast corner in Marquette. And while some people may prefer these sweeter wines over the dryer California varieties, it often raises the question of whether or not wine tasting is a scientific process, or just a matter of taste.

Most wines in Europe are known primarily by geographic appellation (Laube & Molesworth, 1996). In other countries, most wines are labeled by their varietal names, or by grape combinations. But region is just one part of the equation. To understand wine, it is essential to understand the characteristics of the different grapes, and how they should be expressed in wine. There are five fundamental traits used to classify wines: sweetness, acidity, tannin, alcohol, and body (Puckette, How Basic Wine Characteristics Help You Find Favorites, 2020). The first impression of wine is its level of sweetness, which makes taste buds tingle. Acidity refers to how tart the wine is, and wines with higher acidity come across as “spritzy”, while wine that is more rich and round has less acidity. Tannin refers to how astringent or bitter the wine is, and is often confused with the level of dryness, as tannin dries out the mouth. Alcohol content can be tasted by how much the wine warms the throat. Finally wine comes in different body types, such as light, medium, and full. Body is a snapshot of the overall impression of the wine (Puckette, How Basic Wine Characteristics Help You Find Favorites, 2020).

There are hundreds of varieties of red and white wines. Common types of white wine include Chardonnay, Chenin Blanc, Gewurztraminer, Gruner Veltliner, Marsanne, Muscat, Pinot Blanc, Pinot Gris/Grigio, Riesling, Roussanne, Sauvignon/Fume Blanc, Semmillon, and Viognier (Gregutt, White Wine Basics, 2011). Common types of red wine include Cabernet Franc,

Cabernet Sauvignon, Gamay, Grenache/Garnacha, Malbec, Merlot, Mourvedre/Mataro, Nebbiolo, Pinot Noir, Sangiovese, Syrah/Shiraz, and Zinfandel (Gregutt, Red Wine Information & Basics, 2019). With so many red and white wines available, this project will attempt to answer the question of whether the measurable characteristics of a wine can predict whether or not a wine expert will classify the wine's quality as "good" or "bad".

Data Understanding

This project utilizes two datasets, one using red wine samples, and the other using white wine samples. The 11 input variables consist of objective tests that measured acidity, sugar, chlorides, sulfur dioxide, density, pH, sulphates, and alcohol. The output variable is the quality of the wine, based on a scale ranging from 0 (very bad) to 10 (very excellent). The quality was graded based on the evaluation of wine experts, with a median of at least 3 evaluations per instance. The two datasets are related to red and white variants of the Portuguese "Vinho Verde" wine (Cortez, Cerdeira, Almeida, Matos, & Reis, 2009). There are 1,599 instances of red wine, and 4,898 instances of white wine.

Methods

The data sets were examined using Python and R software programs. Histograms of all of the input and output variables were performed on both datasets. The red wine histograms revealed that the density and pH are normally distributed, while the remaining input variables are somewhat right skewed. The white wine histograms revealed that the pH is normally distributed, while the remaining input variables are also somewhat right skewed. The quality variable for both data sets has a semi-normal distribution, meaning there are more wines with a "normal" quality than a good or bad quality. With these datasets, the outliers will be the wines of interest, as they will be of either very good quality, or very bad quality.

A correlation matrix was created for each dataset. For the red wine data set, the correlation matrix revealed that **fixed.acidity** is highly positively correlated with **density** and **citric.acid**. **total.sulfur.dioxide** is highly positively correlated with **freesulfur.dioxide. pH** is highly

negatively correlated with **fixed.acitiy**. And **citric.acid** is correlated negatively with **volatile.acidity** and **pH**. **quality** was negatively correlated with **alcoho**l, and positively correlated with **volatile.acidity**. The white wine data set revealed that **residual.sugar** is highly positively correlated with **density, free.sulfur.dioxide**, and **total.sulfur.dioxide. alcohol** is highly negatively correlated with **density, residual.sugar,** and **total.sulfur.dioxide**. **quality** was negatively correlated with **alcohol** and positively correlated with **density**.

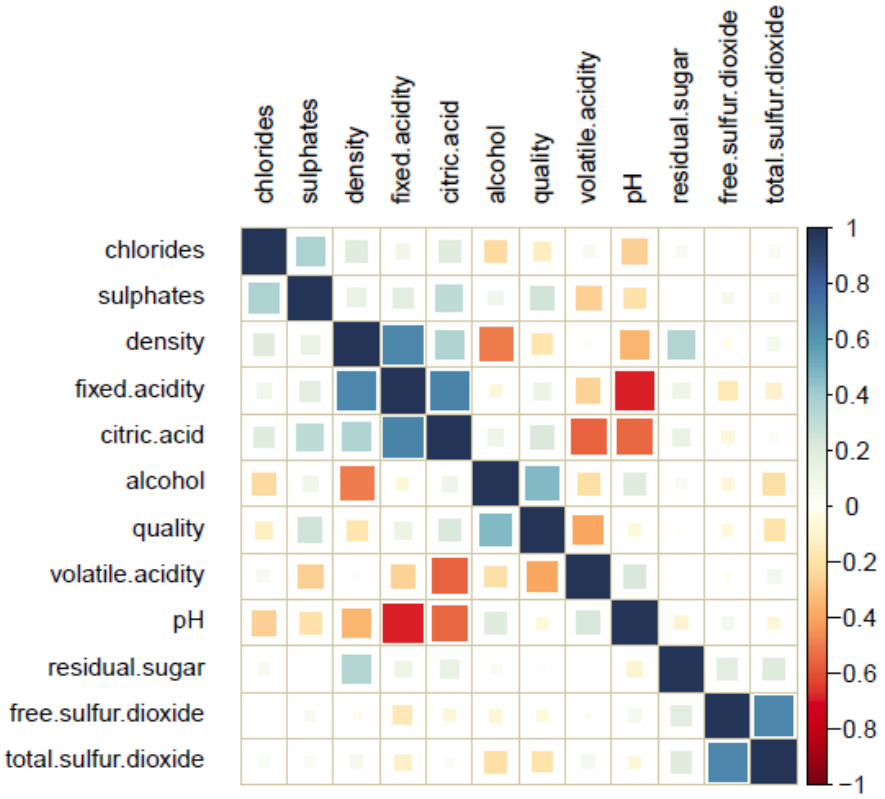

Figure 1 Red Wine Correlation Matrix

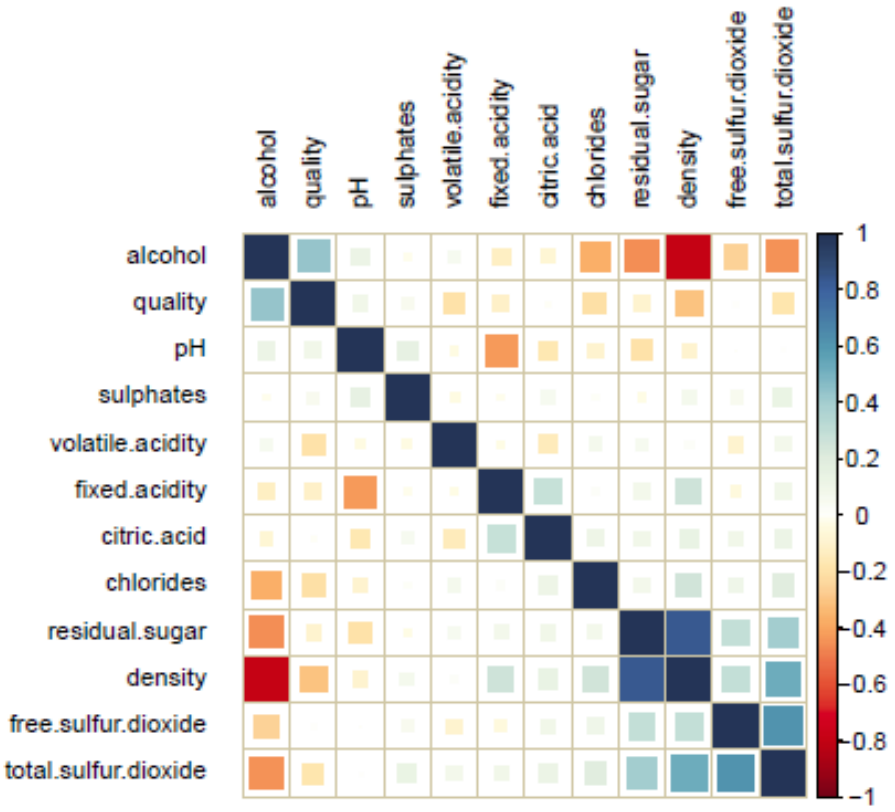

Figure 2 White Wine Correlation Matrix

Modeling and Results

The data sets can be viewed as classification tasks, since the classes are ordered and not balanced. Therefore, regression models were created using both Python and R software. For each data set, the data was split into training and testing sets. Binomial logistic regression models were created in R using the step method, while a linear regression model was created using Python. Models in each software gave impressive results.

Looking at the models in R, the accuracy of the training set was 77% for the red wine and 77% for the white wine. The accuracy of the testing set was 75% for the red wine and 78% for the white wine. Using the stepped method, the final regression model used the variables for **free.sulfur.dioxide, density, sulphates, residual.sugar, alcohol,** and **volatile.acidity**.

Looking at the models in Python, the root-mean-square error(RMSE) was calculated to measure the difference between values predicted by the model and values actually observed. If the model is a good fit, the RMSE for the training and tests sets should be very similar. For the red wine data, the RMSE was .652 for the training set and .627 for the test set. For the white wine data, the RMSE was .749 for the training set and .757 for the testing set. The Python models also indicate that for red wine, an increase in sulphates leads to an increase in the quality of the wine, while increases in volatile acidity and density decrease the quality of the wine. For white wine, increases in pH and sulphates lead to an increase in the quality of the wine, while increases in volatile acidity and density decrease the quality of the wine.

Discussion/Conclusion

The results of the models in both Python and R gave impressive results. Using training and testing data, models in both led to predictions with over 75% accuracy in R and over 62% in Python, with better results for the white wine data sets. The modeling and predictions indicate that there are strong correlations between the measurable characteristics of wine, and the quality of the wine. Given these results, it would be an accurate test to measure the attributes of the wine, and from them determine if the wine will be considered good or bad quality by wine experts. This could help potential wine makers understand if their wines will do well in the market or not, before they are even released.

Acknowledgments

I would like to acknowledge Corez et al for their large data sets. Much work is necessary to compile data, tests, and results and allow them to be freely used by students.

References

Cortez, P., Cerdeira, A., Almeida, F., Matos, T., & Reis, J. (2009, November). Modeling wine preferences by data mining from physicochemical properties. *Decision Support Systems, 47*(4), 547-553.

Gregutt, P. (2011, March 16). *White Wine Basics*. Retrieved from Wine Enthusiast: https://www.winemag.com/2011/03/16/white-wine-basics/

Gregutt, P. (2019, March 19). *Red Wine Information & Basics*. Retrieved from Wine Enthusiast: https://www.winemag.com/2015/10/27/red-wine-basics/

Laube, J., & Molesworth, J. (1996, April 13). *Varietal Characteristics*. Retrieved from Wine Spectator: https://www.winespectator.com/articles/varietal-characteristics-1001

Puckette, M. (2020, April 8). *How Basic Wine Characteristics Help You Find Favorites*. Retrieved from Wine Folly: https://winefolly.com/deep-dive/wine-characteristics/