

Machine Learning for Wine Quality | Analytics Writeup

SMU Data Analytics Bootcamp Capstone | March 2023

Group Members:

Matthew Bailey, Garrett Kidd, Shaunhnessy Robertson, and Brenna Wallace

Introduction:

This project uses machine learning techniques to build a predictive classification model that can classify wines into two quality categories while also classifying whether the wine is Red or White. The model is based on two publicly available Kaggle datasets, [Wine Quality Data Set \(Red & White Wine\)](#) and [Wine Rating & Price](#). The datasets contain data for red and white wines, one for wine composition features and the second for highly rated wines worldwide. Our group was inspired by understanding the features that make up the quality of a good vino. According to data analysis, understanding what goes into the quality can also be economically beneficial because there might be wines that are cheaper than more expensive ones but at a higher quality score. We expected the model to be accurate enough for wine enthusiasts and vintners to improve upon their understanding of quality red and white wine.

Research Questions:

We aimed to build a machine learning model that can accurately classify wine quality based on its makeup of features. The output will be a predicted quality score ranging from 0 to 10, which can be used to identify wines of different qualities. Furthermore, we want to be able to classify and identify wine as either red or white based on each wine's attribute measures.

Data Cleaning/Processing:

Our data from both sources were clean to begin with. The only data engineering we did was combining the red and white wine CSVs into one for our Tableau visualizations. In preprocessing the data for the machine learning models, we dropped the type column and changed our quality column to be labeled as binary quality with the values set to true or false. We elected to engineer the data in this way to simplify the output, and qualify wine as either a binary “good” or “bad”. If the wine's quality value was at six or above, it was designated as good quality and anything below as poor quality.

Data Visualizations in Tableau:

In Tableau, we used the [Wine Rating & Price](#) dataset to create map visualizations showing the wineries by country, to see where the best quality of wine originates from. We originally wanted to create a visualization showing the winery locations to find the region where the most wine is produced. To do this, we plugged our dataset into a Google API that provided us with the latitude and longitude coordinates for each winery. However, our dataset originally contained over 12,000 wines, so we narrowed it down to 4-star and above ratings. This gave us a much more manageable dataset of just over 4000 wines. After using this data with the API, we discovered that the coordinates were inaccurate, with wineries for France populating in Texas, USA. We moved to group wineries at the country level, and created a choropleth map of the world symbolized from a light pink to dark red color scale, to show the average rating of wineries within each country. We also include pop-up information for each country when a user hovers

their mouse over an area of interest, to show metrics such as average price, average rating, and number of wines included in the analysis.

Our first Tableau dashboard consists of an exploration of our data. We created three interactive visualizations as a bar chart, lollipop chart, and a bivariate bar chart. These visualizations show average rating by type, top wine producing countries, and average price by type. A user can manipulate the visualizations using custom built filters for wine type and country of origin.

Our second Tableau dashboard is titled: The Most Influential Features. This dashboard displays five box-and-whisker plots for quality ranges for total sulfur dioxide, alcohol, volatile acidity, chlorides, and density, all components within our wine quality dataset. Users of this dashboard can interact with the visualizations by using a custom build filter for red or white wine types. We elected to use the box-and-whisker plots to discern what numerical qualities of these wine components, on average, make up a good wine.

Machine Learning - Features:

We found several models in our initial research that all used the same features. One of the [wine quality](#) examples listed the features used and what they mean for wine quality:

- Fixed acidity is due to the presence of non-volatile acids in wine. For example, tartaric, citric or malic acid. This type of acid combines the balance of the taste of wine, brings freshness to the taste.
- Volatile acidity is the part of the acid in wine that can be picked up by the nose. Unlike those acids that are palpable to the taste (as we talked about above). Volatile acidity, or in other words, souring of wine, is one of the most common defects.

- Citric acid is allowed to be offered in winemaking by the Resolution of the OIV No. 23/2000. It can be used in three cases: for acid treatment of wine (increasing acidity), for collecting wine, for cleaning filters from possible fungal and mold infections.
- Residual sugar is that grape sugar that has not been fermented in alcohol
- Chlorides: the structure of the wine also depends on the content of minerals in the wine, which determine the taste sensation such as salinity (sapidità). Anions of inorganic acids (chlorides, sulfates, sulfites...), anions of transferred acids, metal cations (potassium, sodium, magnesium...) are found in wine. Their content depends mainly on the climatic zone (cold or warm region, salty soils depending on the observation of the sea), oenological practices, storage and aging of wine.
- Free sulfur dioxide and total sulfur dioxide: sulfur dioxide (sulfur oxide, sulfur dioxide, readiness E220, SO₂) is used as a preservative due to its antioxidant and antimicrobial properties. Molecular SO₂ is an extremely important antibiotic, affecting significant consumption (including wild yeast) that can manifest itself in wine spoilage.
- Density: the density of wine can be either less or more than water. Its value is determined primarily by the concentration of alcohol and sugar. White, rosé and red wines are generally light - their density at 20°C is below 998.3 kg/m³.
- pH is a measure of the acidity of wine. All wines ideally have a pH level between 2.9 and 4.2. The lower the pH, the more acidic the wine; the lower the pH, the less acidic the wine.
- Sulfates are a natural result of yeast fermenting the sugar in wine into alcohol. That is, the presence of sulfites in wine is excluded.
- Alcohol: the alcohol content in wines depends on many tastes: the grape variety and the amount of sugar in the berries, production technology and growing conditions. Wines vary greatly in degree: this Parameter varies from 4.5 to 22 depending on the category.
(<https://www.kaggle.com/code/georgyzubkov/wine-quality-exploratory-data-analysis-ml>)

Machine Learning - Quality:

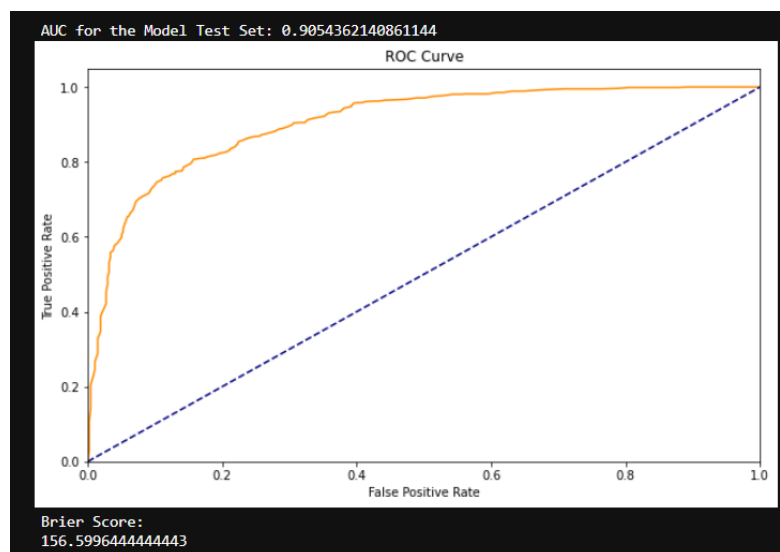
We decided to have two models in our machine learning. The first machine learning model will be able to predict the quality of the wine based on the features listed above. For predicting the quality of wine, our group used a RandomForest model with an accuracy score of 82%, which was lower than our goal of trying to have our models at 90% or more accuracy. However, this was the most accurate model we found after trying various things including using oversampling smoothing: SMOTE(). We attempted binning the original quality scores which ranged from 0 to 10, but the models did not increase in accuracy. Fixing oversampling was the most beneficial fix for our quality model. The model gave us a hierarchical list of feature importance with individual scores. The most valued feature for quality was alcohol, followed by volatile acidity.

```
[
(0.0638691113053404, 'fixed acidity'),
(0.06525115046938254, 'pH'),
(0.06921418831390205, 'citric acid'),
(0.0701710707397306, 'residual sugar'),
(0.07559276008177804, 'sulphates'),
(0.07697401663052739, 'total sulfur dioxide'),
(0.0808579534156564, 'free sulfur dioxide'),
(0.08917249596524728, 'chlorides'),
(0.10537903552080241, 'density'),
(0.12331548042623221, 'volatile acidity'),
(0.18020273713140075, 'alcohol')]

```

METRICS FOR THE TESTING SET:				

[[371 106]				
[126 697]]				
	precision	recall	f1-score	support
False	0.75	0.78	0.76	477
True	0.87	0.85	0.86	823
accuracy			0.82	1300
macro avg	0.81	0.81	0.81	1300
weighted avg	0.82	0.82	0.82	1300



Machine Learning - Type:

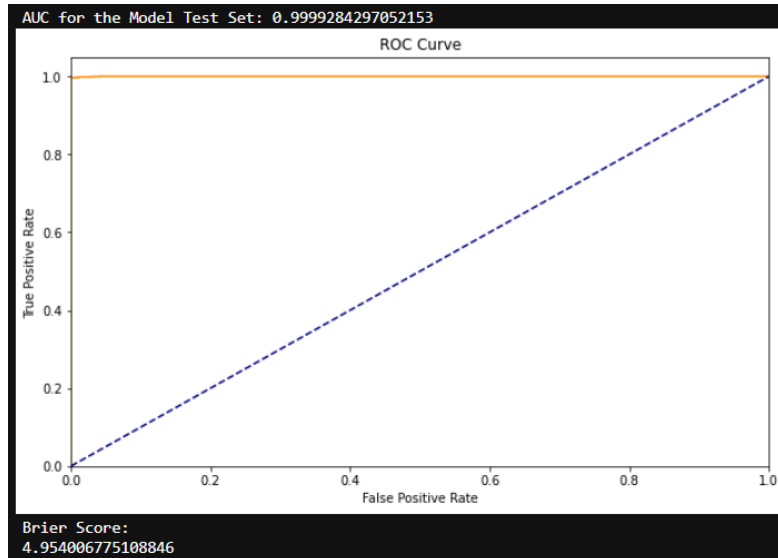
Our second model will be able to predict if the wine is red or white. Our classification model of predicting if the wine is red or white was much more accurate as we used a XGBoost model with a 100% accuracy rating. The feature importance ratings valued total sulfur dioxide as the most appreciated feature, followed by chlorides.

```
[
(0.0051493854, 'free sulfur dioxide'),
(0.008744543, 'citric acid'),
(0.010765948, 'alcohol'),
(0.012624031, 'residual sugar'),
(0.013245545, 'fixed acidity'),
(0.022094695, 'pH'),
(0.028804269, 'sulphates'),
(0.030964078, 'density'),
(0.08135282, 'volatile acidity'),
(0.20809929, 'chlorides'),
(0.5781554, 'total sulfur dioxide')]

```

METRICS FOR THE TESTING SET:				

[[478 2]				
[5 1465]]				
	precision	recall	f1-score	support
0	0.99	1.00	0.99	480
1	1.00	1.00	1.00	1470
accuracy			1.00	1950
macro avg	0.99	1.00	1.00	1950
weighted avg	1.00	1.00	1.00	1950



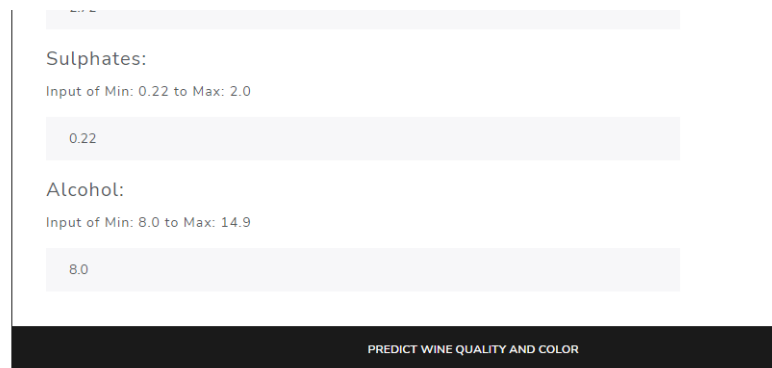
We were surprised by the accuracy of the model, but with limited time and two models to work on, we were unable to go back and try to test if the model is overfitting.

Once both models were created in Jupyter Notebook, we saved each to a .sav file. We then used pickle in our website to locate and render the model for our wine prediction, where users could input features and get an output of wine type and quality.

Web Application:

Our web application contains a landing page, two pages for Tableau dashboards, a page for a Tableau map, an interactive Machine Learning (ML) wine predictor page, one page for this paper, two pages for our main datasets, a works cited page, and an about us page. The landing page contains a brief high-level description of our data analysis project and motivations, and users can use the top “navbar” to navigate throughout the site. On our ML predictor page, we integrated our two models and created a wine prediction functionality with a flask web app. Our goal was to have users input wine component metrics and click a single button to predict both the quality and the type of

wine. To achieve this, we created a function using a combination of Python, Javascript, and HTML coding languages . As long as users of the app input numeric measures within the described minimum and maximum parameters, they will be greeted with a pop-up at the bottom of the page telling them about their wine of interest. For our about us page, you will find a flattering photo of each of us, and what our favorite wines are. Lastly, our works cited page contains the titles and links for all of the sources we referenced.



Sulphates:
Input of Min: 0.22 to Max: 2.0
0.22

Alcohol:
Input of Min: 8.0 to Max: 14.9
8.0

PREDICT WINE QUALITY AND COLOR

IT'S A RED WINE.
SAVE IT FOR THOSE ANNOYING PEOPLE YOU DON'T
LIKE. :(

Future Work and Limitations:

Our datasets and models had several limitations. We wanted to pull the exact latitude and longitude for the winery locations to produce a more detailed map, but our [Wine Rating & Price](#) dataset was extremely large, so we narrowed it down to just 4-star and above ratings, and even after attempting the Google API with that data we were receiving inaccurate coordinates. In a future iteration of this project, we would like to find more accurate geospatial data to show wine quality around the world, specific to regions

of a country. Another limitation we dealt with for the [Wine Quality Data Set \(Red & White Wine\)](#) was that it contained wine data mainly in the 5 or 6 level rating quality. Creating a model that predicted multiple classifications didn't seem logical, as the model would have expected one classification most of the time. Our machine learning model for the quality of wine did not reach our goal and expectation of having a 90% and above accuracy score, so our quality prediction model might give us unreliable results. Our type model may be overfitting so in future work it would be ideal if we could get a larger and more diversified dataset to run through our model to double check for overfitting. Another idea for future work would be to use the longitude and latitudes to create another machine learning model that could predict the quality of the wine based on location or a model predicting the likelihood of the region of origin. To do this, we must find a way to efficiently get the latitudes and longitudes for the wineries or the origin of the grapes used at the wineries.

Conclusion:

Our project was primarily successful in that we could answer the research questions we initially set out to answer. The only drawback was that our quality prediction model was less accurate than what we ideally liked it to be. However, we believe we met our goal of producing a ML model that users could use to successfully predict the quality and type of wine based on the input of the given features in the dataset. We also believe that users will then be able to understand the region where the best rating of wine is produced. Cheers!