Wine Study

Team 2: M. Borunda, K. Childers, and F. Valdez

Southern Methodist University

Data Science Bootcamp (09/2023)

**Introduction**

This study explores wine variants of “Vinho Verde” wine from Portugal. Team 2 was interested to find any correlation between these wine variants and their features which included fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, alcohol, and quality score. A database of these features was developed by combining two datasets for machine learning. The machine learning would help with developing a predictive model on wine type (red or white) based on the features and educate the drinker on what aspects make a high-quality wine. A third data set was used to bring in cost, region of origin, and rated points to enhance findings from the first and second datasets and determine price points and regions of high-quality wine.

# Data Cleaning

The first two datasets were transformed for machine learning by first cleaning and combining them. The first data set brought in red wines and the second one consisted of white. To combine them, team two had to reformat the second set to match the first. Then a column had to be added to each data set to label the wines as red or white.

A screenshot of a data

Description automatically generated

**Figure 1: Original read in of dataset 1 & 2**

**A screenshot of a computer

Description automatically generated  
Figure 2: Added wine type columns to dataset 1 & 2**

**A screenshot of a computer

Description automatically generated**

**Figure 3: Combined datasets 1 & 2**

The range of the data was not extreme, and scaling was not needed. It did look like there was a large skew due to the number of red wines. After analysis, about 93% of the data fell between the quality rating 5-7. The features “fixed\_acidity” and “alcohol” seemed to better determents of wine type. When looking for correlation between the wine features and “quality” the most correlation that could be seen (it was from low to moderate) was “alcohol”, “density”, and “volatile\_acidity.”  
A colorful squares with white text

Description automatically generated

A computer screen shot of a number of objects

Description automatically generated

**Figure 4: Correlation between wine features**

**Findings/Analysis**

Machine learning was performed on the dataset to create a predictive model. Team 2 first performed various analysis to determine the best model to predict quality of wine. The first analysis utilized linear regression. The values that came out showed that it was not the best model to use. The R2 value was low (below 0.4), meaning there was low correlation. The MSE value was slightly higher than team 2 would prefer because the closer to 0, the better the model. In this case it was around 0.54. In addition, the RMSE was kind of high, the model is better the lower it is. Finally, the MAE value should be lower for a better fit model, but it was kind of high as well.

A number on a white background

Description automatically generated

A graph with blue dots

Description automatically generated

A diagram of a plot

Description automatically generated  
**Figure 5: Linear regression analysis to predict wine quality**

Random Forest was utilized next to determine if it would be a better fit model to predict wine quality. It had an accuracy score of 68%. It did verify that alcohol, density, and volatile were the best determinants of wine quality as seen in figure 7 below.

**A screenshot of a graph

Description automatically generated**

**Figure 6: Random Forest analysis**

**A screenshot of a computer program

Description automatically generated**

**A graph with text on it

Description automatically generated**

**Figure 7: Feature importance**

KNN analysis was the next analysis and was only a 54% accuracy. PCA was the next technique and came out okay as a predictive model.

**A number on a white background

Description automatically generated**

**Figure 8: PCA analysis**

Decision Tree Analysis was performed, and accuracy was about 61% as seen in the figure 9 below.

A screenshot of a graph

Description automatically generated

**Figure 9: Decision tree analysis**

In conclusion, team two decided that quality did not have strong enough correlations to the other wine features to use in a predictive model. Therefore, attention was switched to focus on predicting wine type as red or white based on wine features (jupyter notebook title “machine\_learning\_classification”). One-hot encoding was implemented to id the wine type as either 1 for red or 0 for white. Predicting wine types had much stronger correlations as seen in figure 11.A screenshot of a computer

Description automatically generated

**Figure 10: Converting wine types to 1 for red and 0 for white**

A colorful squares with white text

Description automatically generated with medium confidence  
**Figure 11: Correlation with wine\_type**

Team 2 implemented SMOTE and NearMiss techniques to balance out the data due to the higher count of white wines (white: 4,898; red: 1,599). After performing SMOTE and re-sampling   
the data with logistic regression analysis the accuracy skyrocketed to 98%.

A number of numbers in a row

Description automatically generated with medium confidence

**Figure 12: Analysis after SMOTE**

Team two explored some neural network analysis for fun and found it came out with an accuracy level of 96% for one node, 97% for two, and 99% for three.

Further analysis showed that specifically, volatile\_acidity was the best indicator of wine type.   
A number of numbers on a white background

Description automatically generated

**Figure 13: Stats on volatile acidity and wine type**

**Dashboard**

The Dashboard that was created to have an approachable way for people to explore wine and its varying features that make it either a great or bad wine. The Tableau dashboard also incorporates wine regions, prices, and ratings to make the average drinker knowledgeable on what aspects make a highly rated wine. Team 2 organized Tableau into four main dashboards.

The first dashboard focuses on wine types, exploring the distinct difference with alcohol, density, and citric acid that can determine if a wine is white or red.

A screenshot of a computer

Description automatically generated

**Figure 14: Wine type dashboard**

The second dashboard provides an overview of wine quality. A user can interface with the dashboard by selecting a quality to see how alcohol, volatile acidity, pH, and sulphates effected it. Although there were not strong correlations, some could be found with these features to determine a wine quality.

A screenshot of a computer

Description automatically generated

**Figure 15: Wine quality dashboard**

The third dashboard covered wine ratings from the third dataset. A user can view the average wine points per Country and which countries had the highest points for wine.

A screenshot of a computer

Description automatically generated

**Figure 16: Wine ratings dashboard**

The final dashboard displays wine price, point ratings, and average cost of wine per country. A user can enter a price range and points to see which wine would best fit their criteria.

A screenshot of a computer

Description automatically generated

**Figure 17: Wine prices dashboard**

**Conclusion**

Naturally this study is intriguing because wine is within the top 10 beverages consumed globally. This study educates the average drinker on the various features of wine and how they can be used to predict wine type, price, and point ratings. Not every wine has to be super expensive to be a great drink and a multitude of aspects come together to create a perfect bottle of vino.

References

Ambient (outdoor) air pollution (December 2022). World Health Organization.

https://www.who.int/news-room/fact-sheets/detail/ambient-(outdoor)-air-quality-and-health

Air Quality Index, Ozone Alerts & PM Alerts, and Health Advisories. Oklahoma Environmental Quality.

https://www.deq.ok.gov/air-quality-division/ambient-monitoring/aqi-alerts-advisories/

Earth Day 2020: A Guide for All Ages (2020). <https://digitalprojects.davidson.edu/earthday2020/air-pollution/>

Ramachandran, A. World Air Quality Index by City and Coordinates (CC BY-NC-SA 4.0).Kaggle. <https://www.kaggle.com/datasets/adityaramachandran27/world-air-quality-index-by-city-and-coordinates/data>

World Economic Forum (Sep 2020). Can We Put a Price on Clean Air? Yes – And Here It Is.

<https://www.weforum.org/agenda/2020/09/we-can-put-a-price-on-clean-air/>

Terminology

**Air Quality Index (AQI)**: Index is used for reporting daily air quality. It tells you how clean or polluted the air is in a region.

**Particulate matter (PM2.5)**: Fine Particulates such as sulfates, nitrates, ammonia, sodium chloride, black carbon, mineral dust and water.

**Carbon monoxide (CO):** Toxic gas produced by the incomplete combustion of carbonaceous fuels.

**Ozone (O3)**: Ozone at ground level – not to be confused with the ozone layer in the upper atmosphere – is one of the major constituents of photochemical smog and it is formed through the reaction with gases in the presence of sunlight.

**Nitrogen dioxide (NO2)**: NO2 is a gas that is commonly released from the combustion of fuels in the transportation and industrial sectors.