Coursera Capstone Project

Red Wine Quality Predictor

Jordan Chester Sze Chong

September 1, 2020

Table of Contents

1.0 Introduction

Alcohol is enjoyed by many to destress after a long day of working. According to the National Survey on Drug Use and Health, "86.3 percent of people ages 18 or older reported that they drank alcohol at some point in their lifetime; 70.0 percent reported that they drank in the past year; 55.3 percent reported that they drank in the past month. (National Institute of Alcohol Abuse and Alcoholism, 2020)" This tells us that most people would like to drink alcohol during their breaks, but of all the alcohol present in our daily lives, what is the most prominent alcohol people usually drink? As stated in The Wine Industry Network, there are about 77 million wine drinkers around America (Press Release, 2020). Of all the alcohols out there, what is wine? Also, what are different kinds of wine if there are any? Wine is a kind of alcoholic beverage made by fermented juices of grapes. Wine can be of different colors, textures, and fermentations, and can vary in results. There are different kinds (or as they like to call, styles) of wine as stated by Wine Tasting Ljubjana. There are sparkling wines, dessert and fortified wines, aromatic white wines, light-bodied white wines, full-bodied white wines, rose wines, light-bodied red wines, medium-bodied red wines, and full-bodied red wines. In this project, we will be talking about the quality of wine and how to predict the quality of these red wines.

1.1 Problem

Wine has different kinds of processing. As stated by the Wine Month Club, there are five steps in making wine. These are "harvesting", "crushing and pressing", "fermentation", "clarification", and "aging and bottling". In this project, we will be looking at what kind of processes can increase the quality of wine. If so, based on the info, what process can hypothetically make the best quality of wine.

1.2 Interest

Of course, with this information, people who are part of the wine making business may be interested so that they would know what type of processes can improve the quality of the wine itself. Sommeliers (wine stewards), wine testers, or even people who have a hobby of drinking wine may be interested with the information since they will want to know what kinds of processes did they wine makers do, or maybe just knowing what content does each wine have.

2.0 Data

The data that we will be using is based on the csv file by Kaggle located at the reference table below highlighted in yellow (UCI Machine Learning, 2017). Please take note that with this data we will only be comparing different kinds of red wines only. There are no white wine comparisons since the fermentation process between the two are different. Also, please take note that we can only compare the "quality" of the wines, not the future prices of red wine. There will also be no comparison of brands of red and white wine.

2.1 Data Cleaning

The data downloaded from Kaggle has datasets already adjusted for machine learning. There were no empty cells inside the table as stated in Figure 1.

```
No of empty cells:
 fixed acidity           0
volatile acidity         0
citric acid              0
residual sugar           0
chlorides                0
free sulfur dioxide      0
total sulfur dioxide     0
density                  0
pH                       0
sulphates                0
alcohol                  0
quality                  0
dtype: int64
```

Figure 1. No. of Empty Cells in the Table

There was also not much redundant information as everything will be needed to calculate the quality of the wines.

With this, we can explore the dataset and analyze what we can see at the dataset.

3.0 Data Analysis

3.1 Amount and Quality of Wine

We need to know the quantity of wine and the spread of quality inside the file are to know what the possible outcomes for the machines are. As stated in Figure 2 below, there are more wine qualities in the 5 and 6 markers.

```
3     10
8     18
4     53
7    199
6    638
5    681
Name: quality, dtype: int64
```

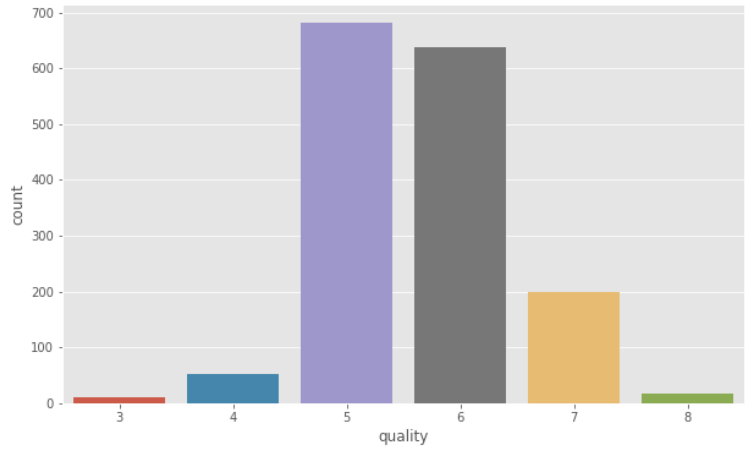

Figure 2. Wine Distribution in terms of Quality

3.2 Relationships Between Quality and Other Chemicals in Wine

With this in mind, we can check what makes the quality of wine better. As stated in Figures 3, volatile acidity and chlorides negatively affect the quality of wine. On the other hand, in Figures 4, citric acid, sulphates, and alcohol increases the quality of wine. It is safe to say that wine should have more citric acid, sulphates, and alcohol in order for the wine's quality to be better.

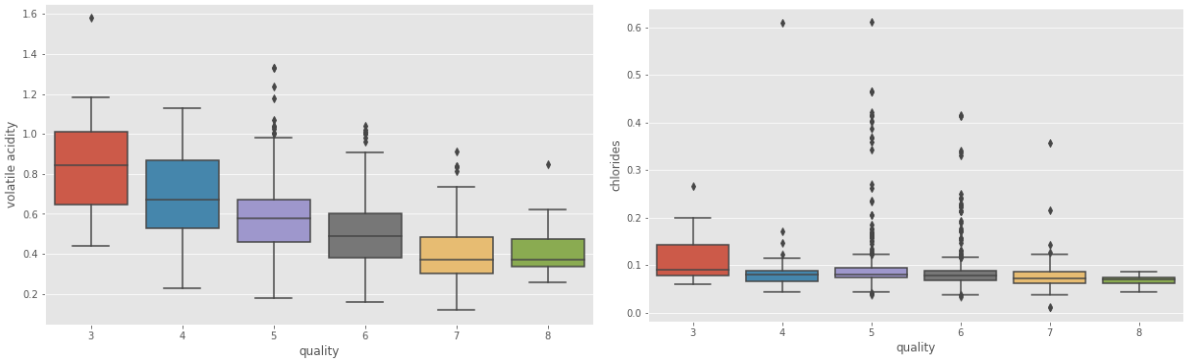

Figure 3. Negative Relation Between Volatile Acidity and Chlorides with Quality

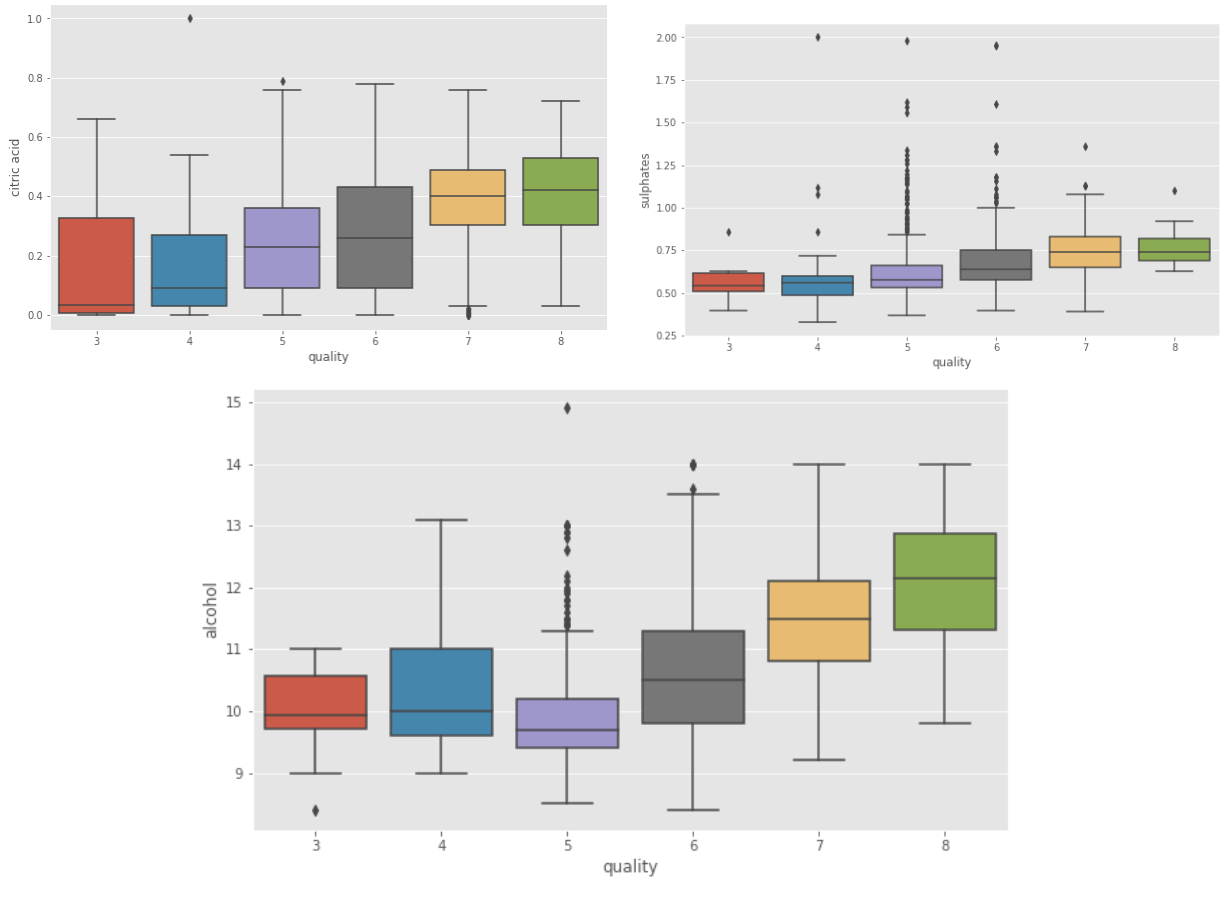

Figure 4. Positive Relation Between Citric Acid, Sulphates, and Alcohol with Quality

3.3 Outliers with Data

As we can see in Figure 4, on the upper right side, there is a bar plot full of outliers. With that, we want to know why. We can look at the data and check why is it. In Figure 5, down at the sulphates row, there is a huge discrepancy with regards to sulphates, making it a little bit complicated for the machine to learn.

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol
count	1599.000000	1599.000000	1599.000000	1599.000000	1599.000000	1599.000000	1599.000000	1599.000000	1599.000000	1599.000000	1599.000000
mean	8.319637	0.527821	0.270976	2.538806	0.087467	15.874922	46.467792	0.996747	3.311113	0.658149	10.422983
std	1.741096	0.179060	0.194801	1.409928	0.047065	10.460157	32.895324	0.001887	0.154386	0.169507	1.065668
min	4.600000	0.120000	0.000000	0.900000	0.012000	1.000000	6.000000	0.990070	2.740000	0.330000	8.400000
25%	7.100000	0.390000	0.090000	1.900000	0.070000	7.000000	22.000000	0.995600	3.210000	0.550000	9.500000
50%	7.900000	0.520000	0.260000	2.200000	0.079000	14.000000	38.000000	0.996750	3.310000	0.620000	10.200000
75%	9.200000	0.640000	0.420000	2.600000	0.090000	21.000000	62.000000	0.997835	3.400000	0.730000	11.100000
max	15.900000	1.580000	1.000000	15.500000	0.611000	72.000000	289.000000	1.003690	4.010000	2.000000	14.900000

Figure 5. General Description of the Data

4.0 Predictive Modeling

There are different kinds of predictive modeling, but the closest modeling that can match with the data is classification since we need to classify the quality of wine once we put down data. Because of this, most of the models are mostly kinds of classification models.

4.1 Applying Classification Models

Before applying classification models to the data, I separated the data into two binary parts. From 3 to 6, I labeled them as “bad” quality wine while from 7 and above, labeled them as “good” quality wine. With that, I began applying these data to the model. These are the results. As SVC garnered the best positives and negatives, we tried improving the SVC model, and Figure 10 was the result once we applied weight to the model.

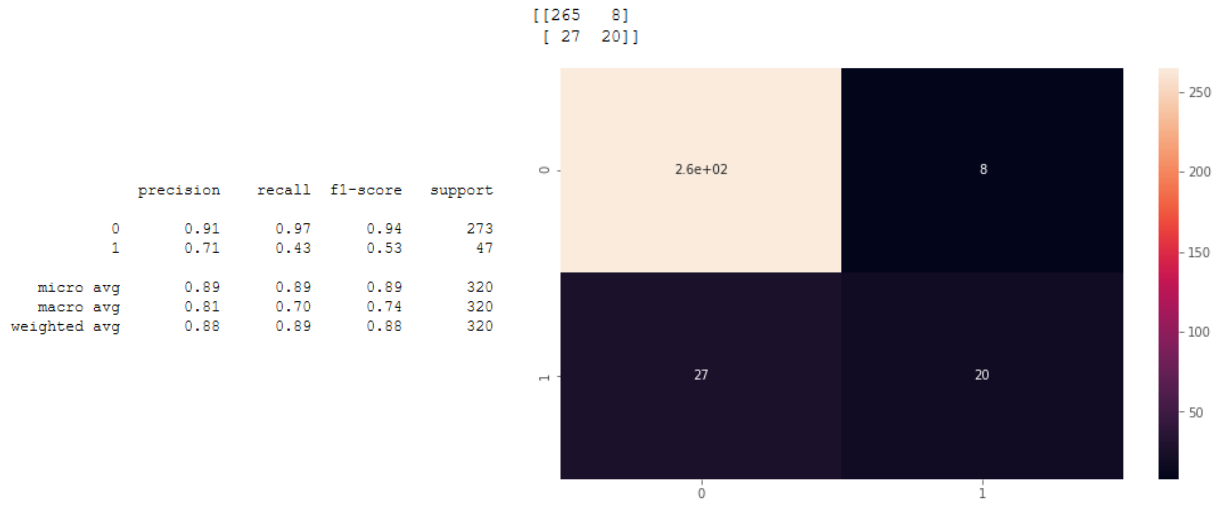

Figure 6. Using Random Forest Classifier

	precision	recall	f1-score	support
0	0.88	0.86	0.87	273
1	0.26	0.30	0.28	47
micro avg	0.78	0.78	0.78	320
macro avg	0.57	0.58	0.57	320
weighted avg	0.79	0.78	0.78	320

```
[[234  39]
 [ 33  14]]
```

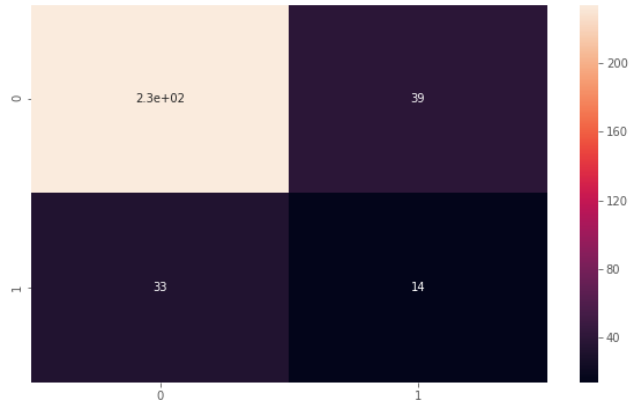

Figure 7. Using Stochastic Gradient Decent Classifier

	precision	recall	f1-score	support
0	0.88	0.98	0.93	273
1	0.71	0.26	0.37	47
micro avg	0.88	0.88	0.88	320
macro avg	0.80	0.62	0.65	320
weighted avg	0.86	0.88	0.85	320

```
[[268   5]
 [ 35  12]]
```

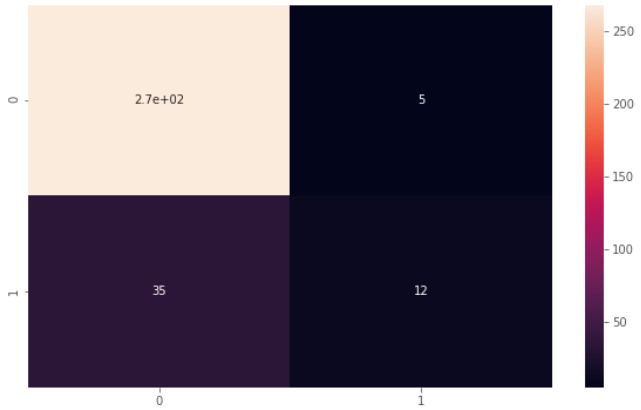

Figure 8. Using Support Vector Classifier

	precision	recall	f1-score	support
0	0.90	0.99	0.94	273
1	0.89	0.34	0.49	47
micro avg	0.90	0.90	0.90	320
macro avg	0.89	0.67	0.72	320
weighted avg	0.90	0.90	0.88	320

```
[[271   2]
 [ 31  16]]
```

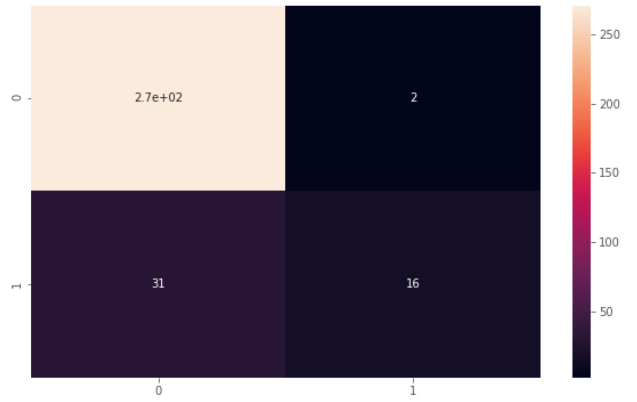

Figure 9. Using Weighted Support Vector Classifier

References

National Institute of Alcohol Abuse and Alcoholism. (2020, February). *Alcohol Facts and Statistics*. Retrieved from National Institute of Alcohol Abuse and Alcoholism: https://www.niaaa.nih.gov/publications/brochures-and-fact-sheets/alcohol-facts-and-statistics

Press Release. (2020, May 5). *US Drinkers Have Increased Wine Consumption During Lockdown, Led by More Involved Drinkers, as Interest in Locally Produced Wine Surges*. Retrieved from Wine Industry Advisor: https://wineindustryadvisor.com/2020/05/05/us-drinkers-increased-wine-consumption-lockdown

UCI Machine Learning. (2017, November 28). *Red Wine Quality*. Retrieved from Kaggle: https://www.kaggle.com/uciml/red-wine-quality-cortez-et-al-2009

Wine Month Club. (n.d.). *Learn The 5 Steps of The Wine Making Process*. Retrieved from Wine Month Club: https://www.winemonthclub.com/the-wine-making-process

Wine Tasting Ljubljana. (2017, June 14). *9 Main Styles of Wine and How They Are Made*. Retrieved from Wine Tasting Ljubljana: https://winetastingljubljana.com/9-main-styles-of-wine-and-how-they-are-made/