Red Wine: Classification Model for Predicting Quality Based on Chemical Data

Michael Bryant

Final project for

DSE 200x UC San Diego

Motivation

I work in the food industry and want to explore how specific chemicals in wine affect quality and to determine if there is any predictive power in the chemical composition by creating a prediction model (decision tree classifier).

This would have implications in product design and manufacturing for red wines. The insights from the analysis could be used to optimize the levels of specific chemicals in manufacturing to make a higher quality red wine.

Data Set

- Red Wine Quality: Simple and clean practice dataset for regression or classification modelling (from Kaggle.com, UC-Irving data set)
    - This model has data regarding chemical concentrations and the quality ranking of red wines. It contains the following variables

fixed acidicity	difference between total acidity (can be measured by pH) and the volatile acidity.
volatile acidity	a gaseous acid that would eventually dissipate from the wine if left open into the air under normal conditions(ambient environmental conditions i.e. room temp and atmospheric pressure)
critic acid	an organic acid chemical, it is gaseous
residual sugar	typically fructose & glucose in wine. More sugar typically registers as a sweeter taste in food.
chlorides	chemicals that contain chlorine
free sulfur dioxide	free sulfur dioxide molecule, not bound to anything else
total sulfar dioxide	free and bound sulfur dioxide molecule
density	weight per volume
pH	a measure of acidity (concentration of Hydrogen Ions) lower = more acidic
sulphates	chemicals that are salts which have a (SO4) anion, like sodium sulphate (Na(SO4 ^2))
alcohol	% ethanol concentration
quality	Ranking given by a judge (unknown, anonymous as provided by data set)

Data Set Continued: Explaining the Dependent variable

- Each value of the dependent variable is used as a class (range is 3 through 8 for quality value- an ordinal number). Therefore, there are 6 classes of quality in this data set: one could consider the classes as follows; very poor (3), poor(4), neutral(5), good(6), very good(7), or excellent(8).

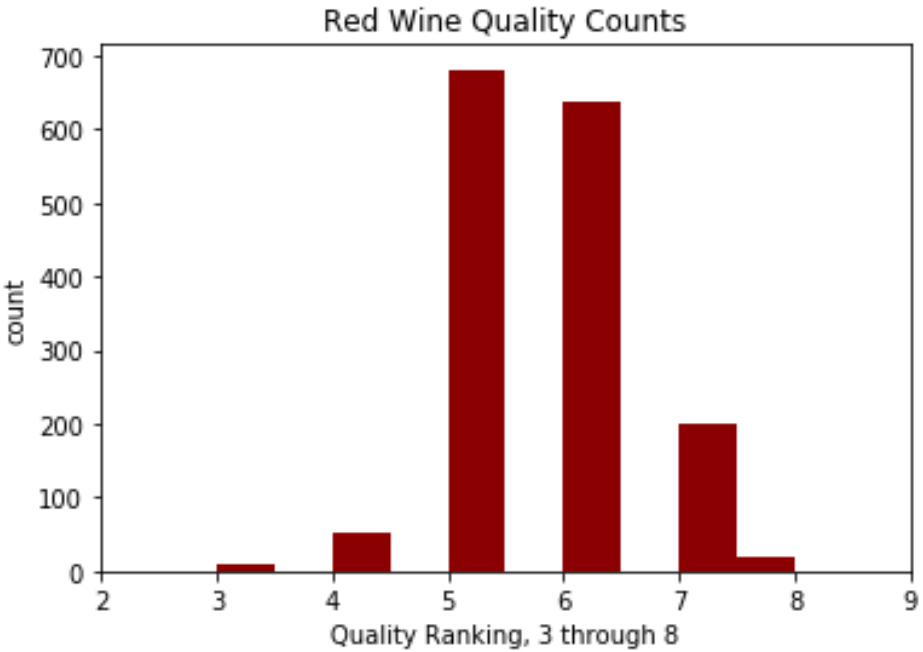
Research Question

Can a classification model (decision tree) be used to predict the quality of red wine using chemical data?

Data prep and cleaning

- This data set required no cleaning and prep. Data was checked for NA’s or other items that would be deemed troublesome.

Findings –Exploration of Dependent Variable



The quality variable ranges from 3 to 8
From the table below, we can see the frequencies of the quality counts in order of most frequent to least frequent.

Quality Ranking	Frequency
5	681
6	638
7	199
4	53
8	18
3	10

Most of data is between 5 and 7

Findings –Exploration of Independent Variables

Summary table of independent variables

	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

The variables above are very important to the quality of the wine and sensory characteristics that a consumer would experience.

Findings- Correlated Values

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
fixed acidity	1.000000	-0.256131	0.671703	0.114777	0.093705	-0.153794	-0.113181	0.668047	-0.682978	0.183006	-0.061668	0.124052
volatile acidity	-0.256131	1.000000	-0.552496	0.001918	0.061298	-0.010504	0.076470	0.022026	0.234937	-0.260987	-0.202288	-0.390558
citric acid	0.671703	-0.552496	1.000000	0.143577	0.203823	-0.060978	0.035533	0.364947	-0.541904	0.312770	0.109903	0.226373
residual sugar	0.114777	0.001918	0.143577	1.000000	0.055610	0.187049	0.203028	0.355283	-0.085652	0.005527	0.042075	0.013732
chlorides	0.093705	0.061298	0.203823	0.055610	1.000000	0.005562	0.047400	0.200632	-0.265026	0.371260	-0.221141	-0.128907
free sulfur dioxide	-0.153794	-0.010504	-0.060978	0.187049	0.005562	1.000000	0.667666	-0.021946	0.070377	0.051658	-0.069408	-0.050656
total sulfur dioxide	-0.113181	0.076470	0.035533	0.203028	0.047400	0.667666	1.000000	0.071269	-0.066495	0.042947	-0.205654	-0.185100
density	0.668047	0.022026	0.364947	0.355283	0.200632	-0.021946	0.071269	1.000000	-0.341699	0.148506	-0.496180	-0.174919
pH	-0.682978	0.234937	-0.541904	-0.085652	-0.265026	0.070377	-0.066495	-0.341699	1.000000	-0.196648	0.205633	-0.057731
sulphates	0.183006	-0.260987	0.312770	0.005527	0.371260	0.051658	0.042947	0.148506	-0.196648	1.000000	0.093595	0.251397
alcohol	-0.061668	-0.202288	0.109903	0.042075	-0.221141	-0.069408	-0.205654	-0.496180	0.205633	0.093595	1.000000	0.476166
quality	0.124052	-0.390558	0.226373	0.013732	-0.128907	-0.050656	-0.185100	-0.174919	-0.057731	0.251397	0.476166	1.000000

Using the corr() function in Python the above table was created. There are some correlations greater than 0.5 magnitude, highlighted in yellow above. Anyone with a chemical background should not be surprised to see these values, but that discussion is beyond the scope of this project. What is important to note is that quality does not have a correlation greater than 0.5 with any other variable. Alcohol content is correlated at 0.476 with quality-therefore, one could say quality is moderately positively correlated with alcohol content.

First Classification: Explanation of model

.

- I created a training set with 2/3 of the data to make the decision tree classification model- in Python this is created using “DecisionTreeClassifer” in the “sklearn.tree” package.

- A test set using the remaining 1/3 of the data will be used to determine the accuracy of the model by seeing how well the model predicts the quality rank in the test set.

    - The accuracy is simply the correct answers divided by the total answers(sum of wrong and correct ranks).

First Classification Decision Tree Results:

- The decision tree classification model predicted the correct quality rank with accuracy of ~55% of the time. Considering this has 6 rank classes, this model is much better than simply guessing- which would result in a correct answer 1/6 or ~17% of time.

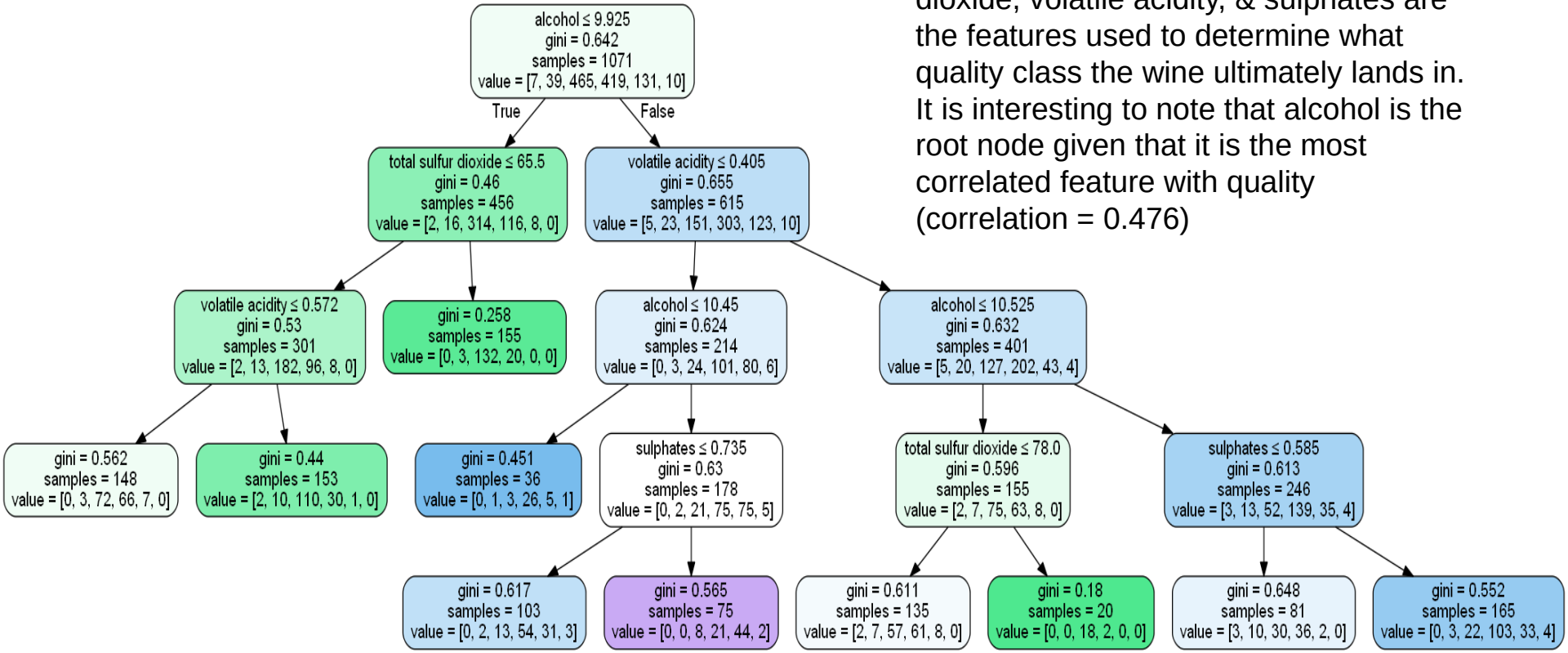
- Accuracy score of the classification model: Python Screenshot

```
accuracy_score(y_test, predictions)
```

```
0.5511363636363636
```

Findings: Decision Tree Visual

From the decision tree print out, we can see that alcohol content, total sulfur dioxide, volatile acidity, & sulphates are the features used to determine what quality class the wine ultimately lands in. It is interesting to note that alcohol is the root node given that it is the most correlated feature with quality (correlation = 0.476)



Findings: Ideal classification cutoffs to create binary classification

- A binary model can be created using arbitrary cutoffs. Using the same method for the multiclassification model in terms of the training and test set, the following accuracies were found:

Rank Cutoff Accuracy

3	0.9924953095684803
4	0.9530956848030019
5	0.699812382739212
6	0.8949343339587242
7	0.9831144465290806

From the analysis, we can see that a binary classification model increases the accuracy significantly, the lowest accuracy being ~0.7 or 70% accuracy, which occurs at a split of 5, where <= 5 is a “poor quality” wine and >=6 is a “good quality” wine.

The sensitivity of each cut off was not tested. At a cutoff of 7, the model has an accuracy of 98%, but as most of the data is under 7, it is not a surprise. It would be good to see how accurate it predicted the 8 quality ranking.

Limitations of Model

- The sensitivity & specificity were not analyzed here. Overall accuracy was the desired measure. Other measures (e.g. ROC-AUC) could also be executed.
- Another model, such as logistic/ordinal regression, may have a higher accuracy and provide magnitudes of the variables to see what is more important.
- Random forests or boosting techniques (ada boost, Xgboost) could have also be executed to improve the model.

Acknowledgements

- I used information from the lecture, and chemical knowledge that I have obtained from my chemistry degree and food industry experience to explain the dependent variables.

- Most likely this analysis or something similar has been done before, given the free availability of the data set, but I did not use anyone's research or projects for my model.

References

None, other than information from the lectures.