Wine Quality Machine Learning Project

By: Nic Landry, Annie Dang, William Edge, Erik Martinez, Abigail Loy, Wenhao Zhao, Vlad Potapenko

Abstract:

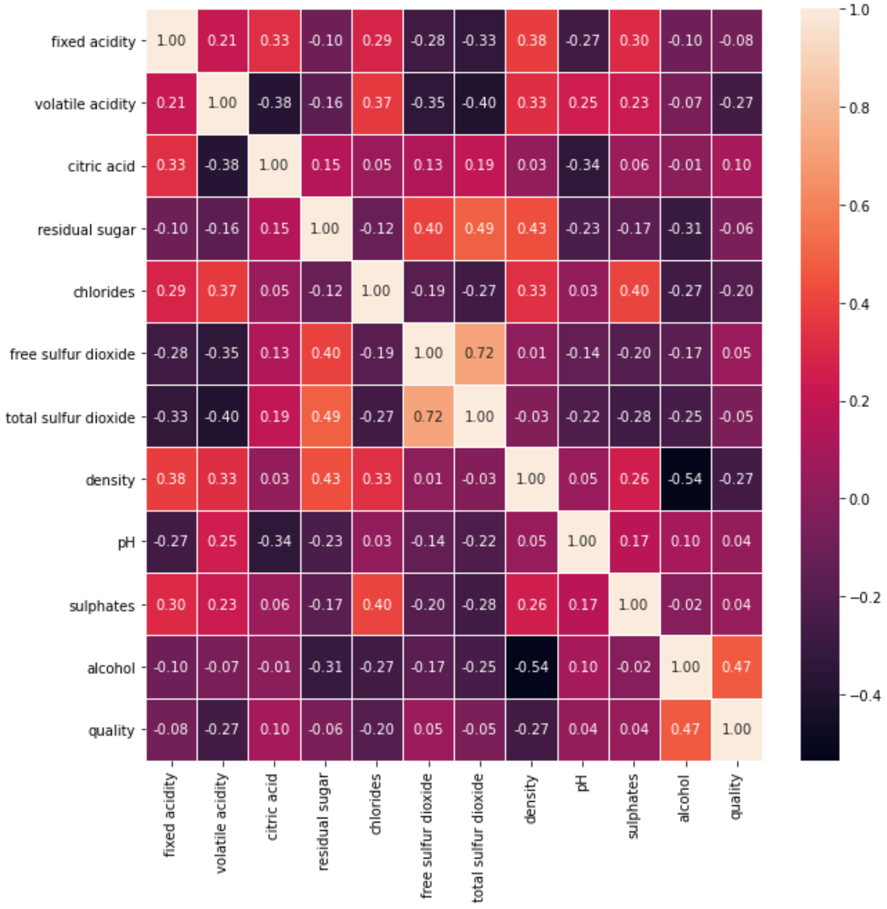
In order to produce higher quality wine, it's beneficial to understand what in the chemical makeup of a wine generates its quality. Through various machine learning models, we explore the features inherent to wine that allow for proper classification of quality in a "poor, decent, good" level of classification. Through analysis, we have discovered that focus on fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, and alcohol can lead to accurate wine quality classification. Particularly, these features through a specialized Random Forest model can accurately predict wine quality, regardless of color, approximately 83% of the time. With this information, wine producers may potentially engineer their production to achieve higher quality control.

Introduction

As global wine consumption has grown—with a nine percent climb from 2000 to 2018—so has the desire to produce high-quality wine. While production process improvements such as temperature-controlled fermentation involving stainless steel tanks have led to quality jumps across the industry, a question remains: what features inherent to a wine define its quality?

For the purpose of this project, our team has set out to answer this question with a focus on machine learning principles. Analyzing a dataset comprised of 5320 observations of distinct red and white Vinho Verde wines, we develop a machine learning model in hopes of accurately classifying wine by its chemical composition. In effectively doing so, we hope to distinguish the characteristics in wine that determine its quality, an accomplishment that is self-evident in its implication for wine producers.

In our provided data, both red and white wines are observed with respect to the following features: fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, alcohol, and quality. These features provide the basis for our determination of significant characteristics, and we take as an initial assumption their importance in wine quality classification.



Methodology

In order to validate the dataset's variables as valuable predictors of quality, we employ various machine learning classification models. These models test the relevant features' ability to delineate between quality levels, with greater success in classification reflecting positively on the features' relationship to quality.

To compare feature combinations, a Kendall's Tau Test was conducted. Kendall's Tau is a non-parametric test for assessing the relationship between two variables, with one variable, quality in this case, being ordinal. The test selects features based on statistically significant correlation coefficients between a given feature and a wine's quality. By keeping only significantly correlated features, we expect Kendall's Tau to make our models more efficient.

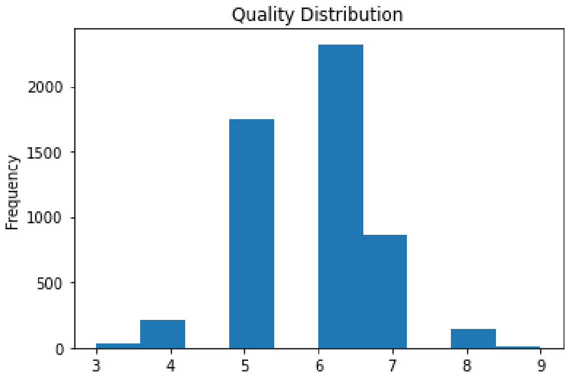
After performing the test, we discovered that all variables inherent to the dataset are useful for quality prediction, and thus we chose to run on our models with a full feature set.

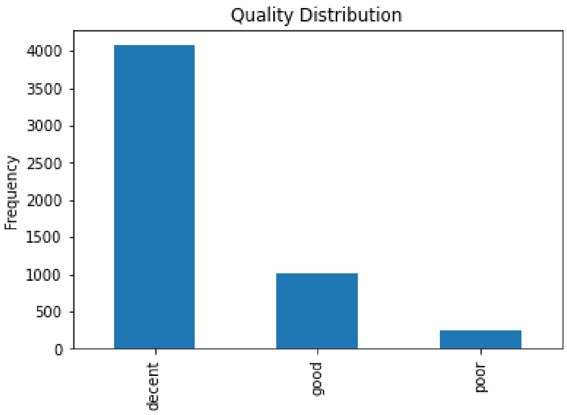
To test features under various machine learning algorithms, our group conducted train-test splits of the data for the following models (all models were ran with Random States equal to 1111 and a split where test samples comprised 30% of the total data):

- ***Random Forrest****: An ensemble model that uses averaging of randomized decision trees to carry out prediction.*
- ***Logistic Regression****: A linear model that classifies through quality level probability calculation.*
- ***K-Nearest Neighbors (KNN)****: A model that uses data points closest to a new point in order to classify it*

Accompanying these models, a Support Vector Machine model, as well as more specialized Random Forest and KNN models were run parallel to provide greater appreciation for model parameters and cross validation.

Presented with data that separates wine quality on a 3-9 scale, we reclassify quality into three levels representing more comprehensive scores such as "poor, decent, good." In the adjusted classification system, scores of 7 or higher equate to "good;" scores greater or equal to 5 and less than 7 equate to "decent;" and scores of 4 or below equate to "poor." The "poor, decent, good" scale shall be referred to as the *Alternate Scale*. Though we run our models under both systems, we focus our attention on the tri-level classification as our data collection fails to provide sufficient observations for each quality level on the 3-9 scale. The inherent underrepresentation in our data opens the door to biased results. By broadening our classification range, we create a more conservative model that may perform well when tested against outside data.



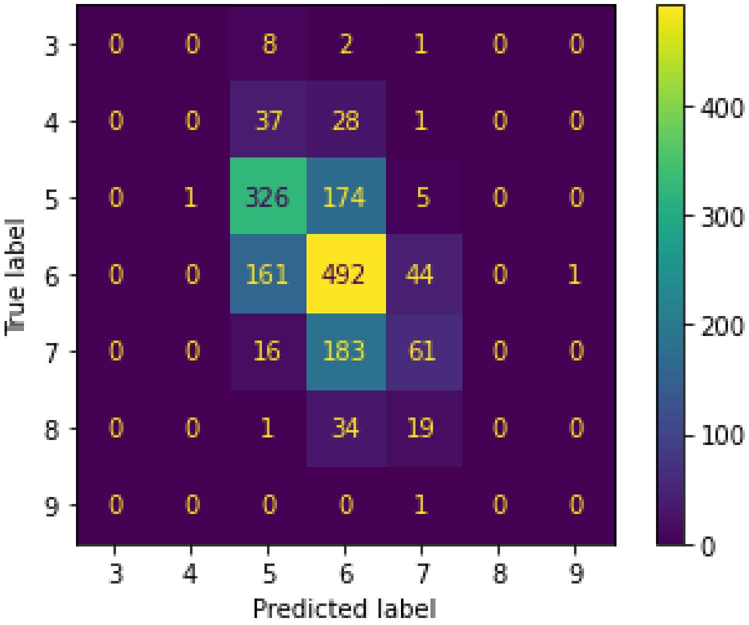


To recognize differences between red and white wines, models were run on datasets where the colors are separate as well as where they are combined. Working in the Python programming language, we used the SciKit Learn, Numpy, Pandas, Seaborn, SciPy**,** Time, and Matplotlib modules for our analysis.

Results and Analysis

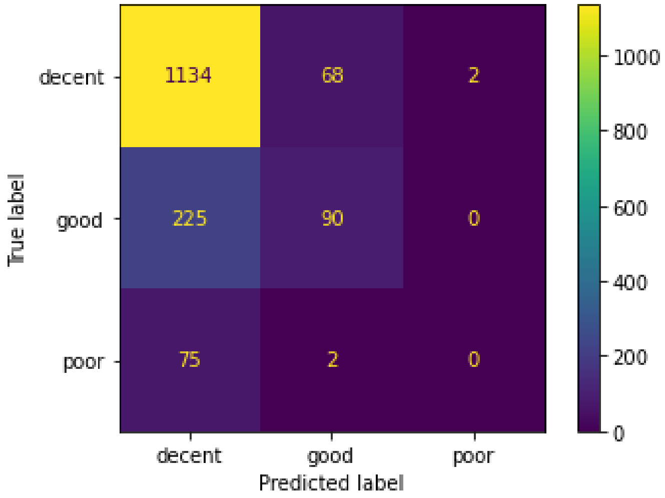
All Wines

Combining both red and white wine data sets, we recognized a sizeable jump in prediction accuracy when moving from the 3-9 quality scale to the 3-level scale of “poor, decent, good.” While the Random Forest model provides an accuracy score of 56.69% with the 3-9 scale, accuracy shoots to 78.95% with the alternate scale. This notable increase is again demonstrated with the Logistic Regression and K-Nearest Neighbors models, where the alternate scale sees accuracy scores of 76.69% and 75.75% for the test set, increases of nearly 21% and 30%, respectively.



Through reconfiguration of the Random Forest model with quadruple the estimators (400) and tuned hyperparameters, the RF model produced an improved accuracy score of 82.31%. This is the best score among any model for the composite dataset.

Judging by our results, when wine quality is rated on a 3-9 scale we have approximately a coin-flip's chance of correctly classifying a given sample. This makes sense as the lack of both high and low observations in our data fail to give the models proper representations of the extreme quality levels. Without proper representation, the models have more difficulty delineating between wines found at the lower and higher ends of the quality spectrum. Such difficulty is highlighted by confusion matrices for the Logistic Regression Model provided in the appendix. These plots demonstrate the difficulty in classifying "poor" or low-quality wines through a miss-rate of 100% as all 77 true "poor" wines we mislabeled as "decent" (75 out of 77) or "good" (2 out of 77) . This phenomenon is common among the other models, particularly the specialized Random Forest where all True "poor" samples were labeled as "decent."



White Wine Only

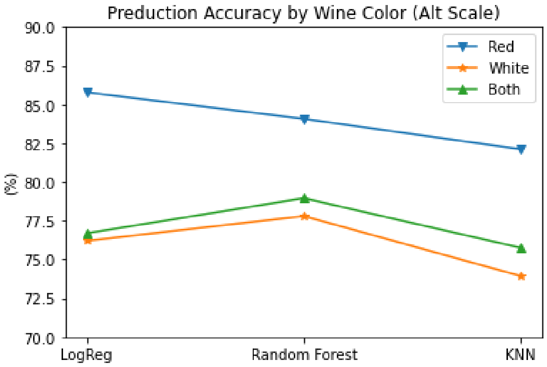
Looking strictly at white wines, our models have less success in accurately classifying the alternate scale while performing comparably to "all wines" models with the 3-9 scale. Under the alternate scale, the Logistic Regression, Random Forest, and K-Nearest Neighbors models score 76.20 , 78.55%, and 73.93% respectively. As a point for comparison, the Logistic Regression model for white wine under the 3-9 scale scored 53.74%. The dramatic increase in efficacy is seen throughout the models as we move from 3-9 scaling to the alternate method. Intuitively, this makes sense due to regrouping providing more evenly distributed quality observations, and thus easier classification given the data. When working with the specialized Random Forest Model outclassed all other models with an accuracy score of 83.69, trumpeting the benefits of parameter tuning.

Red Wine Only

As seen with the white wine data, when attention is solely focused one wine color, our models perform more admirably. Logistic Regression, Random Forest, and K-Nearest Neighbors models with respect to red wine score 85.78%, 84.07%, and 82.11% under the alternate scale. These figures, as before, considerably outperform those of the 3-9 scale. Using the SVM model on the red wine data, we produced an accuracy score of 81.48%, falling in line with the regular models. The specialized Random Forest, however, again beat out the field with an accuracy of 88.33%, the highest score among any combination of inputs.

From the results of our testing, we feel confident in our models' ability to classify wine quality given our designated features. With percentages settling in the high 70s to low 80s when classifying among combined red and white wine types, the models seem adept at distinguishing between quality without first understanding the wine color. This characteristic makes our models more robust to testing outside data as well as suggests that wine quality, independent of wine color, has a distinct connection to the features. Understanding this connection, wine producers can incorporate these features in production quality control efforts and product development.

The fact accuracy scores improve when models focus on only one color of wine suggests that it is better to attempt classification separately if color is known. This partition by color allows for less ambiguous relationship calculations and in turn predictions as it removes variability from the models by narrowing the range over which predictions must be made. Furthermore, as white and red wine models have distinct levels of prediction power under the same algorithms, our results suggest that our features impact wine quality differently based on color even if a general connection can be established in the case where wine colors are not known.

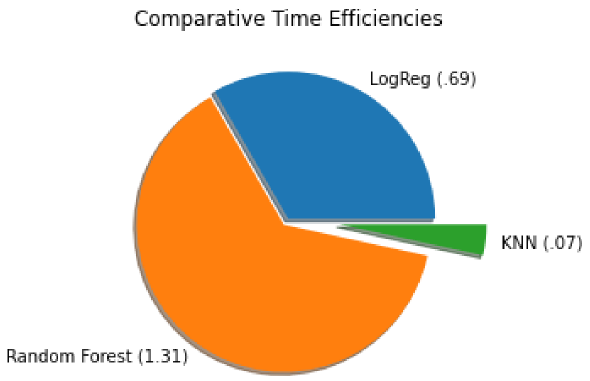


Other Considerations

While our models correctly predict wine quality at a decent rate, there is much to be said for the models' inability to classify wines on the high and low ends of the quality scale. As mentioned earlier, the lack of observations pertaining to extreme qualities deprives the models with sufficient information to create accurate classification parameters on those ends. This problem leads us to believe it is possible that these models can thus be accurate only with respect to our used dataset. If scores abnormal to our dataset are, in fact, more common to the general population of quality scores and wines, then our models may fail to into account the proper distribution and thus become error-prone. Methods to correct for the imbalance are duplication of observations, however this was not employed for our analysis.

In terms of feature selection and the use of the Kendall Tau Test, our assumption of a monotonic relationship between features and the target array (quality) may be a bit presumptive. Preliminary analysis does suggest monotonicity for features such as sulphates, but validation of monotonicity for all features would require more analysis which, for the sake of time, has been left out. Models could be further developed through cross-validation as well as parameter tuning, some of which has been done and provided with the code in the appendix.

To check for the efficiency with which the various machine learning models operate, we timed the Logistic Regression, (Basic) Random Forest, and K-Nearest Neighbors to see the speed of their calculation. Though our dataset is small, our analysis suggests more time commitment from the Random Forest model in the case of substantially larger datasets. With times of 1.31, .69, and .07 seconds corresponding to Random Forest, Log. Reg., and KNN, the large disparity between KNN and Random Forest (x18.7) and the notable difference between LogReg and Random Forest (x2) gives merit to balancing computational speed with accuracy in the case of greater amounts of data. There may be instances where the LogReg may be a more practical model as it approaches Random Forest in terms of accuracy while running twice as fast.



Conclusion

Following our analysis, our team believes wine quality can be effectively classified through its features such as fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, and alcohol. This relationship is demonstrated through high accuracy/classification scores using Logistic Regression, Random Forest, and K-Nearest Neighbors models. With scores in the 70s to 80s, wines can be classified irrespective of color, though such segregation does provide better results. This suggests that chemical composition varies in its effect based on wine color, though a general connection outside of color may exist. Further analysis is needed to distinguish the relationship of color, quality, and composition.

****NOTE****

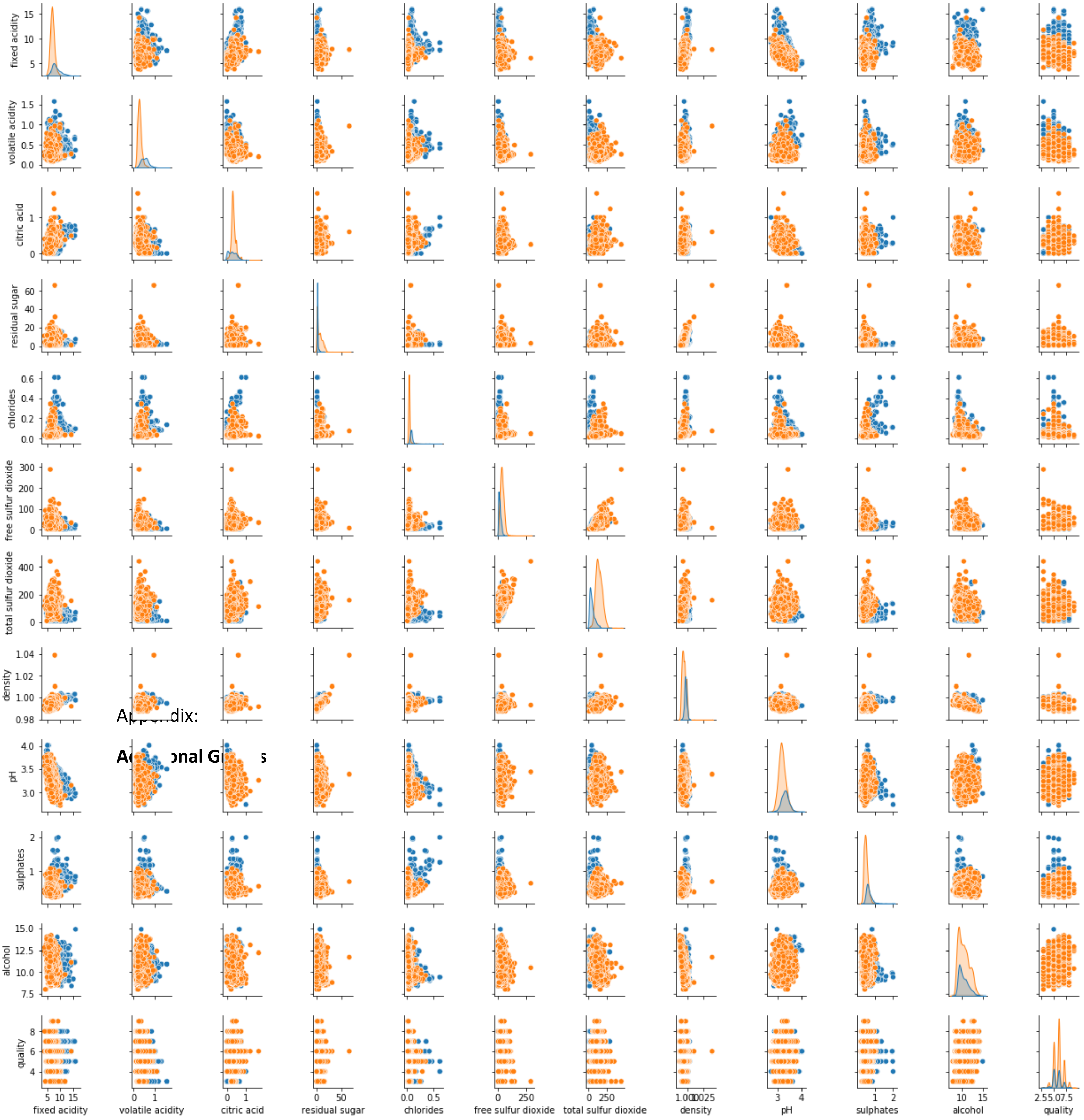
All code, including graphs, data-cleaning, etc, can be found in the Team1 GitHub Repository:

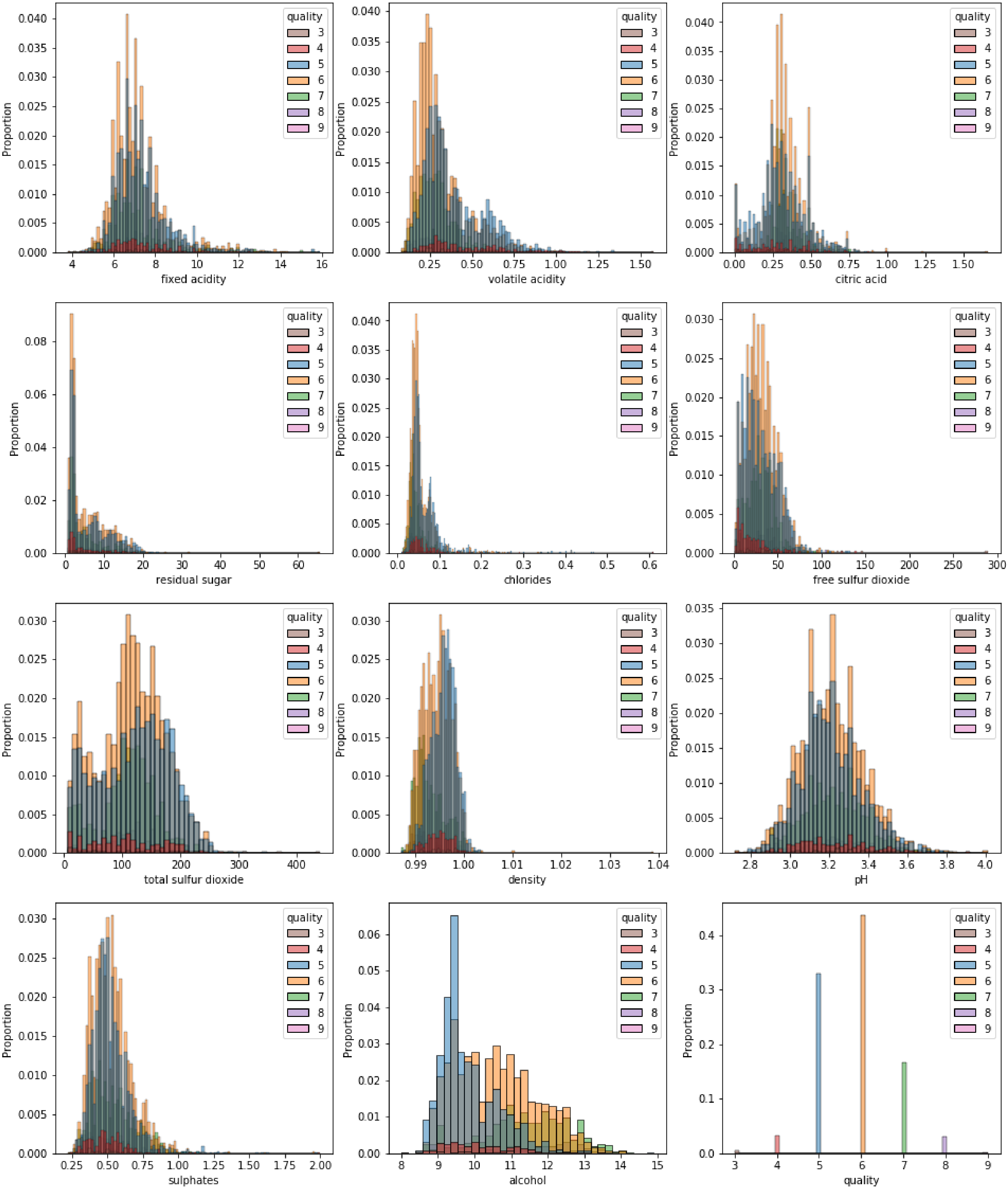
vladimirpotapenko/TTS-ML-Wine-Project-Team-1-: TTS ML Group Project - Team 1 (github.com)

This report does not encompass all work completed for the project, so please visit the GitHub Repo

Appendix:

Additional Graphs





Proportion
0.000
0.005
0.010
0.015
0.020
0.025
0.030
0.035
0.040
4
6
8
10
12
14
16
fixed acidity
0.25
0.50
0.75
1.00
1.25
1.50
volatile acidity
0.00
0.25
0.50
0.75
1.00
1.25
1.50
citric acid
0.00
0.02
0.04
0.06
0.08
0
10
20
30
40
50
60
residual sugar
0.0
0.1
0.2
0.3
0.4
0.5
0.6
chlorides
0
50
100
150
200
250
300
free sulfur dioxide
0
100
200
300
400
total sulfur dioxide
0.99
1.00
1.01
1.02
1.03
1.04
density
2.8
3.0
3.2
3.4
3.6
3.8
4.0
pH
0.25
0.50
0.75
1.00
1.25
1.50
1.75
2.00
sulphates
0.01
0.02
0.03
0.04
0.05
0.06
8
9
10
11
12
13
14
15
alcohol
0.0
0.1
0.2
0.3
0.4
3
4
5
6
7
8
9
quality
quality
3
4
5
6
7
8
9