# Project Milestone 3

Harlan Wittlieff

Data Science, Bellevue University

DSC 630: Predictive Analytics

Dr. Brett Werner

April 24, 2022

# Introduction

With the almost infinite options winemakers have at their disposal when creating a new wine many find themselves asking, “What incorporates a successful wine?” Is it alcohol content, sweetness, another metric, or a combination of multiple characteristics? Knowing the foundation of a good wine can turn a failing winery into a thriving one. In today’s competitive environment, unlocking the formula for a successful wine is the advantage that can set a winery above the competition.

To investigate this question, a dataset was selected from UCI’s machine learning repository. The “Wine Quality Data Set” contains 4898 instances of white wine and 1599 instances of red wine of the “Vinho Verde” variety. The dataset contains eleven total characteristics for each individual wine. These features are listed below.

- Fixed Acidity
- Volatile Acidity
- Citric Acid
- Residual Sugar
- Chlorides
- Free Sulfur Dioxide
- Total Sulfur Dioxide
- Density
- pH
- Sulphates

- Alcohol

In addition to the features listed on the previous page the data set contains an output variable, quality. The quality metric is built from sensory data from wine experts. These experts graded each wine based on a scale from 0-10 with 0 equating to a “very bad” score and 10 being “excellent.” The final quality metric is the median from at least three of these scores for each wine.

From the combination of features a model will be built to predict a wine’s quality. The initial model to be evaluated will be k-nearest neighbors. This model was selected based on the assumption that features of a good wine will fall into groups. In other words, good wines will have similar characteristics. Based on the success of this model, regression and random forest may also be considered. All models will be built with a training and test set of data. The existing data will be randomly split placing eighty percent of the data into the training set and twenty percent into the test set to evaluate the model.

After the model is completed, it must then be evaluated. The primary tool for understanding the model’s success will be a confusion matrix of the model’s predictions on the test data set. The confusion matrix will be used to understand the model’s overall accuracy in addition to identifying the individual areas where the model succeeds and fails. Priority will be given to models that successfully identify high ranking wines (wines with a quality score of 7 or higher) as these wines contain the characteristics that need to be identified by the model. Misclassification of a poor-quality wine (below 6) as a high-quality wine will also be considered more harshly.

Based on the success of this model, additional types of wines may be investigated to further expand the overall scope and applicability of the model. This additional insight will allow the model to be more relevant to wineries that create wines outside of the initial project's scope.

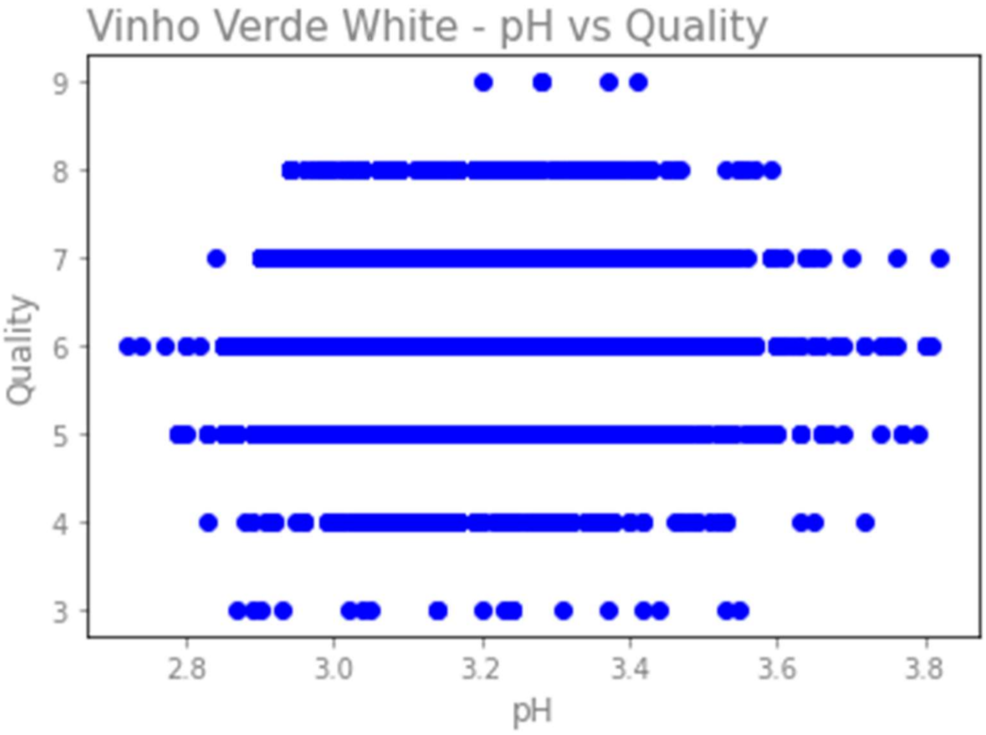
If the model successfully predicts the quality within the test data set, there are still a few risks that warrant investigation. The first risk is that the quality metric is subjective. This metric is based on the opinion of an expert and may not translate well to other populations. Secondly, the subjective nature of the quality metric brings the risk of inconsistency. If the quality metric was not created in a consistent fashion, any model may have difficulty in predicting it. Additionally, just because a wine is considered high quality does not mean the wine will be a financial success. Other factors (marketing, economic conditions at the time, etc.) will play a role in determining an individual's wine's overall acceptance within a population.

Should this original project plan for the evaluation of wine quality not work out as anticipated, a contingency plan is in place. This contingency plan consists of investigating additional dataset to uncover new features of wine and potentially additional correlations through the investigation of the relationships between the new and existing features and their impact on the wine's quality. The lack of unique identifiers could cause this contingency plan to be more challenging.

## Preliminary Analysis

I am confident that I should be able to identify the key characteristics that make up a superior Vinho Verde wine using the Wine Quality Data Set. As I investigate the relationships between specific variables and their impacts on the overall quality of wine, plotting the variables against wine quality on a scatter plot will be most useful as this visualization technique allows for immediate visual cues to identify correlation. Additionally, I plan to gain insight from the overall distribution of key variables through leveraging histograms.

As an example, insight can be gained into pH's impact on quality by creating a scatter plot of the two variables. As the below chart demonstrates, white wines that achieved an extremely high rating of 9 for quality had a narrow pH range of approximately 3.2-3.5.



At this time I do not anticipate needing to adjust the data or my driving questions. I may however need to adjust my model and evaluation choices. Based on the success of the KNN model, additional models may be explored such as decision trees and regression models.

Overall, I feel that my original expectations relative to this project are still reasonable. I look forward to the data cleaning and model construction for project milestone four.

# References

Cortez, P., Cerdeira, A., Almeida, F., Matos, T., & Reis, J. (n.d.). *Wine Quality Data Set*. UCI Machine Learning Repository: Wine quality data set. Retrieved March 25, 2022, from https://archive.ics.uci.edu/ml/datasets/Wine+Quality

P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis.

Modeling wine preferences by data mining from physicochemical properties.

In Decision Support Systems, Elsevier, 47(4):547-553. ISSN: 0167-9236.