Final Report
Predicting Wine Quality Based on Physicochemical Properties

Brandon Lavinsky

Abstract

Many different fields in science leverage machine learning (ML) to aid in understanding data and predicting important trends. Oenology, the study of wine, is one of those fields in which machine learning can be applied to determine how to refactor, price, and evaluate different wines with minimal human involvement. In the context of this project, machine learning was used to determine if the physicochemical properties of wine can effectively determine its quality (rating 0-10). Through extensive data analysis and ML modeling, it was determined that physicochemical properties are mediocre at predicting an exact quality rating (0-10) for both red and white wines using the available data due to the data's limited quality distribution. The top performing Random Forests (RF) model produced **70.75%** and **67.43%** accuracy for both wines respectively. Modifying the framework to rate wines on a scale from 0 to 2 (0 - poor, 1 - moderate, 2 - good) however, accuracy for the same RF model increased to **86%** and **84.2%** respectively.

Environment & Objective

The environment for this project can be seen as a subset of oenology (the study of wine), but mainly focusing on the impacts of machine learning in the field. Like many other subsets of science such as biology and chemistry which utilize machine learning, oenology is no different and can benefit from data analysis and machine learning modeling.

The core objective of this research project is to leverage data analysis, machine learning, and the physicochemical properties of red and white wines to create a framework for assessing the quality of wines without the need of human involvement i.e. taste tests, sommelier ratings, etc. By eliminating the need of humans in the rating process, wineries can quickly and cost effectively create ratings by (1) not having wait through a lengthy taste testing process and (2) paying sommeliers (wine experts) to individually rate their wines.

Research Questions

The primary research questions to be answered through this project are:

1. Can the physicochemical properties of wine be effectively used to rate the quality of wine?
2. Which of the/combinations of physicochemical properties are most impactful at determining wine quality?
   (a) Are these properties different or the same for red and white wines?
3. Which data analysis/machine learning techniques and models are most successful at predicting wine quality?

Related Work

There are many examples of work directly related to modeling wine quality, price, and characteristics. In particular, Cortez et al. were the original authors/creators of the wine quality based on physicochemical properties dataset and provided the initial analysis/modeling of the data back in 2009 (1). Similarly, Kotsiantis et al. used supervised machine learning to analyze the organoleptic properties of matured wine distillates to eliminate human involvement and subjectivity in the wine refinement process (2).

Project Infrastructure & Setup

Development Environment

The main development environment for this project is a Jupyter Notebook running Python 3.9 through Anaconda.

The main Python libraries used include:

- Pandas
- Numpy
- Matplotlib/Seaborn
- Scikit-learn
- Tensorflow/Keras

Datasets

The datasets used in this project consist of two separate sets: one for red wine and the other for white wine. The datasets are obtained from UCI (3). Both datasets consist of the same 11 attributes & 1 output. The red wine set contains 1599 samples while the white wine set has 4898. A sample from the red wine dataset can be observed in Table 1. The attributes are: fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, and alcohol. The output is the quality of the wine ranging from values 0 through 10.

Table 1: Red Wine Data Samples.

fixed acidity	*volatile acidity*	*citric acid*	*residual sugar*	*chlorides*	*free sulfur dioxide*	*total sulfur dioxide*	*density*	*pH*	*sulphates*	*alcohol*	***quality***
7.4	0.7	0	1.9	0.076	11	34	0.9978	3.51	0.56	9.4	**5**
7.8	0.88	0	2.6	0.098	25	67	0.9968	3.2	0.68	9.8	**5**
7.8	0.76	0.04	2.3	0.092	15	54	0.997	3.26	0.65	9.8	**5**
11.2	0.28	0.56	1.9	0.075	17	60	0.998	3.16	0.58	9.8	**6**

Data Preprocessing

Neither dataset requires any data preprocessing other than some standardization for certain models. Aside from this, all entries are: labeled, non-null, and numerical.

Feature Selection

Given that many of the attributes in both the red and white datasets are either highly correlated with one another or not correlated with the target quality label, feature selection was applied to produce the ideal configuration for each model. The feature selection process will be discussed further in the *Experimental Results* Section.

Experimental Vision

Experiment Selection

In order to effectively answer the research questions previously defined in Section *Environment & Objective: Research Questions* a varying number of feature selection, hyperparameter tuning, and machine learning modeling techniques were explored in this research. These techniques are listed below.

Feature Selection Techniques

- Feature Correlation
- Chi^2 Test
- Forward Feature Selection

Hyperparameter Tuning

- GridSearch
- RandomSearch

Models

- Linear Regression
- Naive Bayes
- Logistic Regression
- K Nearest Neighbor
- Decision Trees
- Random Forest
- Support Vector Machines (SVM)
- AdaBoost
- Simple Neural Net

The utilization of the aforementioned techniques will be used to answers all of the target research questions.

Initial Hypothesis

Without any knowledge about how the physicochemical properties effect wine quality, I initially speculated that KNN and SVMs would perform well. Both models generally exhibit good performance and with fine-tuning, high accuracy could be achieved. The one property I expect to be a good predictor is residual sugars. The sweetness of a wine should be telling of its quality. The *Experimental Results* section provides insights on this initial hypothesis.

Experimental Results

Below are a set of experiments and results in the context of this research followed by the answers to the proposed research questions.

Data Distribution by Quality

When first diving into analysis & modeling it is very important to visualize the distribution of the wine quality for both red and white wine. Figure 1 displays this distribution of quality across the entire dataset. From the figure, it becomes evident that the majority of ratings for both wines fall between the 5-7 quality, heavily favoring 5s and 6s. Additionally, neither the red or white wine set contained values below a 3 or any 10 quality ratings. From this visualization, it is evident that the data lacking a wide variation of quality values could hinder its classification potential especially for future use with new data.

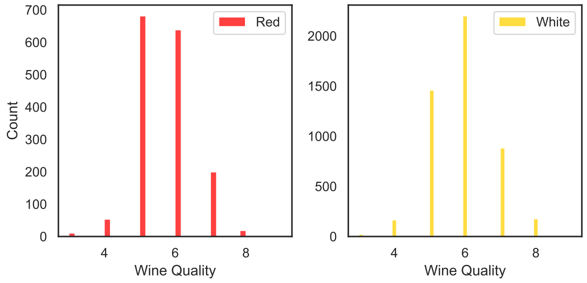

Figure 1: Distribution of quality for red & white wines.

Feature Correlations

In order to determine which features were highly correlated with the quality variable, two correlation heat-maps one for red and the other for white wine were created. Figure 2 displays the correlation for red wine while Figure 3 displays the correlation for white wine. From Figure 2 the features most correlated with quality are: alcohol, followed by sulphates, citric acid, and volatile acidity. Similarly, for Figure 3 alcohol is the most correlated, but is instead followed by pH, and then sulphates. Surprisingly, residual sugars had very little correlation to the quality.

Visualizing Most Highly Correlated Features

Visualizing the two top correlated features with quality namely alcohol/sulphates and alcohol/pH for red and white

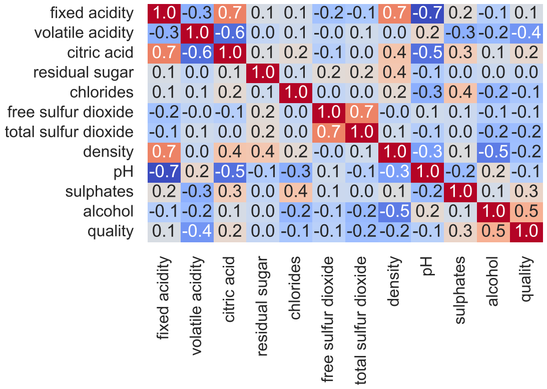

Figure 2: Feature correlations for red wine.

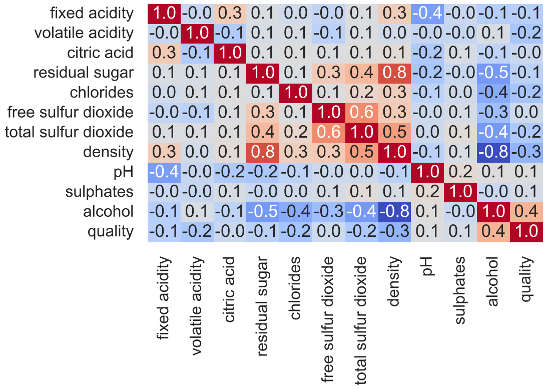

Figure 3: Feature correlations for white wine.

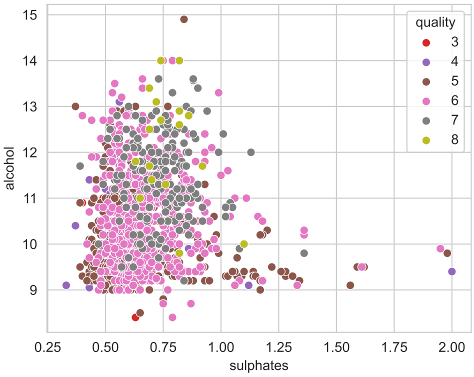

Figure 4: Relationship between alcohol & sulphates on red wine quality.

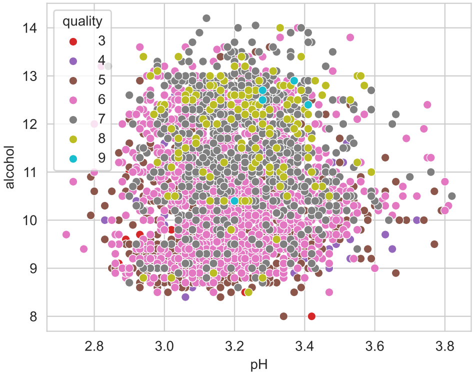

Figure 5: Relationship between alcohol & pH on white wine quality.

wine respectively as shown in Figure 4 and 5 provides some interesting findings. First, there is little separation between each of the quality ratings. Specifically, there is little to no distinction between the 5 and 6 ratings and only marginal separation between 7 and 8. Secondly, the distinction of quality labels is more apparent for red wine compared to white. The effects of alcohol and sulphates on quality depicts a much clearer trend for red wine whereas alcohol and pH in white has much more variation. This little distinction between qualities and data variation may make predicting exact ratings difficult especially for white wine.

Chi^2 Best Features

In conjunction to a creating correlation heat-maps, a Chi^2 test to determine strong features was conducted. According to the Chi^2 test for red wine the top 4 selected features were: total sulfur dioxide, free sulfur dioxide, alcohol, and volatile acidity. For white wine the top 4 selected features were: total sulfur dioxide, free sulfur dioxide, residual sugar, and alcohol. Contrary to the correlation heat-maps, sulphates was determined the 8th best feature and pH the 10th for red and white respectively.

Forward Feature Selection

In order to determine the ideal configuration of features on top of the correlation and Chi^2 test, forward feature selection was applied to each model. For weak/inaccurate baseline models this process was done manually testing the top correlating & Chi^2 features. Conversely, for strong/accurate base models, an exhaustive search of all possible feature combinations was conducted to create the most accurate configuration. While this process is relatively time consuming given the possible 2047 feature combinations, with reasonably sized datasets the selection took between 5 and 15 minutes depending on the complexity of the model.

Machine Learning Modeling

Train & Test Sets The training and testing set were partitioned such that **75%** of the data was used in training and the remaining **25%** was used for testing. This configuration (75/25) was the same for both wine types.

Procedure To construct the best models possible, the procedure illustrated in Figure 6 is followed. First, the model type is selected. Next, a base model including all 11 features & default hyperparameters is created. Default hyperparameters refer to the preset values defined by scikit-learn. Different methods of forward selection are then applied as discussed in *Forward Feature Selection*. Depending on the complexity of the model, either RandomSearch or GridSearch is conducted to tune the hyperparameters. Finally, the base model, model with only feature selection applied, and the feature selection + hyperparameter tuned model are compared choosing the most accurate as the best model.

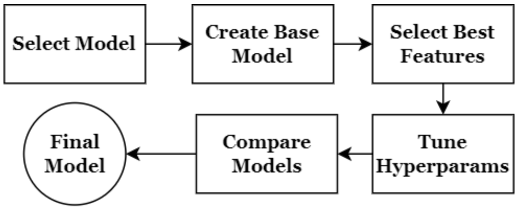

Figure 6: Model selection procedure.

Model Performance Given a plethora of models were tested, only a general synopsis of performance is provided for each. The order of accuracy will be reported as red first and then white. Features will be labeled numerically 0 to 10 moving left to right in the columns of Table 1 such that fixed acidity is 0, volatile acidity is 1, citric acid 2, residual sugars 3, all the way to sulphates at 9 and alcohol at 10.

- **Linear Regression:** Even with feature selection and hyperparameter tuning the linear model was extremely poor with **36.5%** and **28.7%** accuracy. The data cannot be learned through a linear relationship.
- **Naive Bayes:** For red wine, Gaussian Naive Bayes with a var_smoothing set to $1e^{-22}$ and features [7, 8, 9, 10] performed best with **57%** accuracy. For white wine, Categorical Naive Bayes using all features and an alpha=0.01, and fit_prior=True resulted in **49.7%** accuracy. An improvement to linear regression, but still very poor.
- **Logistic Regression:** For red wine, features [0, 1, 3, 4, 9, 10] and default hyperparameters resulted in **59.75%** accuracy. For white wine, features [1, 2, 3, 7, 9, 10] and default hyperparameters resulted in **54.37%** accuracy.
- **K Nearest Neighbor:** Using GridSearch and exhaustive forward feature selection resulted in features [1, 2, 8, 9, 10] and hyperparameters {’algorithm’: ’brute’, ’leaf_size’: 1, ’n_neighbors’: 1, ’p’: 2} with an accuracy of **65%** for red wine. For white wine, features [1, 3, 7, 8, 9, 10] and hyperparameters: {’algorithm’: ’auto’, ’leaf_size’: 1, ’n_neighbors’: 1, ’p’: 1} resulted in an accuracy of **60.2%**.
- **Decision Trees:** For decision trees the base model using default hyperparameters and all features resulted in the best accuracy of **62%** and **58%** for both wine types.
- **Random Forest:** Similar to decision trees, random forest also performed best with the default hyperparameters, but in its case, the best features were [1, 3, 4, 6, 9, 10] for red and [0, 1, 2, 4, 5, 6, 9, 10] for white with accuracy **70.75%** and **67.43%**. This model was the top performer for both red and white wine.
- **Support Vector Machines:** For red wine, features [0, 1, 2, 5, 7, 9, 10] and hyperparameters {’C’: 1.0, ’degree’: 1, ’gamma’: ’scale’, ’kernel’: ’rbf’} resulted in **62.5%** accuracy. For white wine, features [0, 1, 2, 5, 7, 9, 10] and hyperparameters: {’C’: 0.7625, ’degree’: 1, ’gamma’: ’scale’, ’kernel’: ’rbf’} resulted in **55.76%** accuracy. All data was standardized using StandardScaler before input into the SVM model.
- **AdaBoost:** For red wine, features [1, 2, 4, 8, 9, 10] and hyperparameters {’learning_rate’: 0.3272, ’n_estimators’: 16} produced **58%** accuracy. For white wine, features: [1, 3, 6, 8, 10], with default hyperparameters had an accuracy of **50.04%**.
- **Simple Neural Net: Architecture**: input 3-red or 4-white layer → dense 11 relu layer → 1% dropout layer → dense 4 relu layer → dense 10 softmax output layer. With **Optimizer**: Adam; **Learning rate**: 0.01; **Loss**: sparse categorical crossentropy. Data was standardized using StandardScaler before input into the network. The network was trained for 10 epochs with a batch size of 15. For red wine the best input features were [8, 9, 10] yielding **61.75%** accuracy. For white wine features [1, 7, 9, 10] were best yielding **52.9%** accuracy.

The accuracy breakdown comparison for each model for the red and white wine test datasets is illustrated in Figure 7 and Figure 8 respectively.

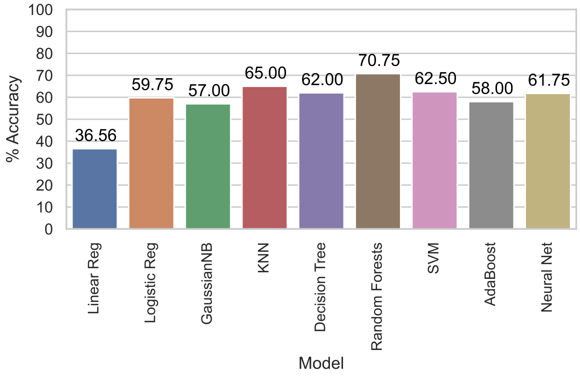

Figure 7: Red wine test set accuracy per model.

Modifying the Framework Given that the best models only resulted in 70.75% and 67.43% accuracy and the separation between each quality label is minimal, testing the outcome after modifying the target label seemed necessary. Specifically, the quality was changed such that values between 0-4 are marked as 0 or poor, values 5-6 are marked as 1 for moderate, and values 7-10 are marked 2 or good. After doing so, the KNN, Random Forest models and the neural network (now with 3 outputs) were re-tested with the newly modified quality label. The comparison of these models with the originals are displayed in Figure 9 and Figure 10. From the results presented in the figures, it is evident that by simply modifying the target quality label, that accuracy is greatly improved. This improvement is not surprising now that the 5 & 6 labels are combined; the two most prevalent

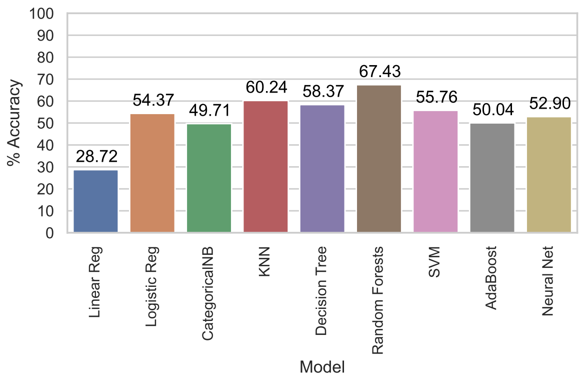

Figure 8: White wine test set accuracy per model.

classes with also the least distinction between each other. Prior to modifying the label, 5s would often be classified as 6s and vice versa, negatively impacting accuracy.

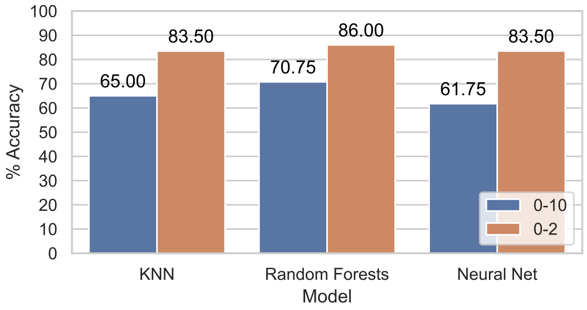

Figure 9: Red wine test set accuracy qualities 0-10 vs. 0-2.

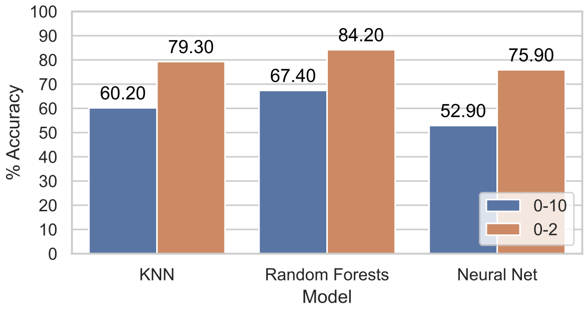

Figure 10: White wine test set accuracy qualities 0-10 vs. 0-2.

Answering the Research Questions

1. Can the physicochemical properties of wine be effectively used to rate the quality of wine?
   Rating the exact quality of wine (0-10) using just physicochemical properties only works moderately well given the best models produced 70.75% and 67.43% accuracy. A more relaxed rating scheme (0-2) however, works quite well with 86% and 84.2% accuracy using the exact same models.
2. Which of the/combinations of physicochemical properties are most impactful at determining wine quality?
   The most predominate properties for predicting quality are: alcohol, sulphates, volatile acidity, pH, and density.
   (a) Are these properties different or the same for red and white wines?
   The properties are generally the same for both wine types, however, white wines were impacted by residual sugars and density more, while red was effected more by pH.
3. Which data analysis/machine learning techniques and models are most successful at predicting wine quality?
   From this experimental evaluation, the top models for both red and white wine are: Random Forest, followed by KNN, and SVM for red and Decision Tree for white. Forward feature selection and feature correlations were both very beneficial to a models success. The Chi^2 test, on the other hand, was misleading with its top selections. Total & free sulfur dioxide were not very helpful in predicting quality.

Conclusion

Machine learning and data analysis/visualization are powerful techniques that can be applied to many scientific fields to aid in predicting trends. Specifically, in the context of this work, these techniques are leveraged in the field of Oenology to create a wine rating framework that can rate the quality of red and white wine using only their physicochemical properties, eliminating human involvement in the process. While a strict rating framework predicting the exact qualities (0-10) had mediocre performance, altering the framework to use poor, moderate, and good as ratings significantly improved the ability to rate wine quality with **86%** and **84.2%** accuracy for red and white wine respectively. From this analysis it is clear that the models are limited by the given data, due to the limited quality distribution. Gathering more ratings with qualities 0-4 and 7-10 has the potential to produce an even stronger wine rating framework.

References

[1] Cortez, P., Cerdeira, A., Almeida, F., Matos, T., & Reis, J. (2009). Modeling wine preferences by data mining from physicochemical properties. Decision support systems, 47(4), 547-553.

[2] Kotsiantis, S. B., Tsekouras, G. E., Raptis, C., & Pintelas, P. E. (2005, July). Modeling the organoleptic properties of matured wine distillates. In International Workshop on Machine Learning and Data Mining in Pattern Recognition (pp. 667-673). Springer, Berlin, Heidelberg.

[3] wine-quality. Index of /ml/machine-learning-databases/wine-quality. (n.d.). Retrieved March 26, 2022, from https://archive.ics.uci.edu/ml/machine-learning-databases/wine-quality/