Red and White Wine Quality

Training a Wine Quality Classification Model

Jessica Barber-Bauman
School of Engineering Graduate
Program in Software
University of St. Thomas
St. Paul, MN, US
barb8975@stthomas.edu

Saeedeh Hosseini
School of Engineering Graduate
Program in Electrical Engineering
University of St. Thomas
St. Paul, MN, US
saeedeh.hosseini@stthomas.edu

Pa Xiong Vang
School of Engineering Graduate
Program in Software
University of St. Thomas
St. Paul, MN, US
vang1901@stthomas.edu

ABSTRACT

The purpose of this study is to employ machine learning algorithms in making the most accurate prediction of human wine taste preference given several physicochemical characteristics of red and white Vinho Verde wine. Experiments were conducted on datasets containing information about red and white Vinho Verde wine quality taken from the University of California Irvine (UCI) Machine Learning Repository. This study explores the use of classification algorithms, including logistic regression, k-Nearest Neighbors (k-NN), support vector machine (SVM), Naïve Bayes, Random Forest, Ensemble, Extra Trees, and Gradient Boosting, as well as Deep Neural Network (DNN), to accurately predict the sensory quality of each wine instance. Random Forest and DNN models proved to be two of the more accurate machine learning classification predictors.

KEYWORDS

UCI Machine Learning Repository, Wine Classification, Random Forest, Deep Neural Network

1 Introduction

Wine certification and quality assessment are key to the wine production and sales processes. Laboratory tests are conducted to determine the wine sample's physicochemical properties, while human taste testers evaluate its sensory quality element. It is important to note that for the purposes of these datasets, "each sample was evaluated by a minimum of three sensory assessors (using blind tastes), which graded the wine in a scale that ranges from 0 (very bad) to 10 (excellent). The final sensory score is given by the median of these evaluations" [1]. How a wine's physicochemical properties relate to human taste preference is still somewhat unknown; therefore, the use of machine learning (ML) techniques to better understand this relationship is very valuable to the growth of the wine industry. One of the most important uses of ML technology in this area is the implementation of classification algorithms to predict the sensory quality of wine samples using the chemical attributes determined in laboratory tests as input [2][5]. Previous research studies have determined that regression techniques and support vector machine (SVM) produce higher accuracy models for the red and white Vinho Verde wine quality datasets from the University of California Irvine (UCI) Machine Learning Repository [1]. More recent research is showing that neural networks and random forest yield even more accurate and promising results [2][3][4][5]. When they were first added to the UCI Machine Learning Repository, the red and white Vinho Verde wine quality datasets consisted of 178 rows and 13 chemical attributes [3]. Wine producers in the Vinho Verde region have contributed several more data instances via the Viticulture Commission of the Vinho Verde region (CVRVV), allowing us access to one of the largest datasets containing both physicochemical and sensory data on wine samples [1].

This study focuses on training the most accurate predictor of sensory quality based on the physicochemical properties input for each wine sample. In order to obtain models with the most computational efficiency, feature reduction techniques were performed on the 12 predictor variables. The process of obtaining a model that achieved the highest accuracy while using only significant predictors, allowed us to develop a better understanding of machine learning techniques and models. Experiments were performed using logistic regression, k-Nearest Neighbors (k-NN), support vector machine (SVM), Naïve Bayes, Random Forest, Ensemble, Extra Trees, Gradient Boosting, and Deep Neural Network techniques in order to determine the optimal classification model for the red and white Vinho Verde wine quality datasets.

2 Background

Vinho Verde is a class of wines named for the demarcated region of northwestern Portugal in which it is produced. The grapes of that region are "known for producing not only light and fresh wines, but also complex, structured and mineral wines" [6]. The original datasets were collected between May 2004 and February 2007 using only protected designation of origin samples that were tested by the official certification entity (CVRVV). The CVRVV is an inter-professional organization with the goal of improving the quality and marketing of Vinho Verde wine. The data were recorded by a computerized system (iLab), which manages the process of wine sample testing from producer requests to laboratory and sensory analysis [1].

3 Red and White Wine Datasets

3.1 Dataset Characteristics

The red and white Vinho Verde wine quality datasets were downloaded from the UCI Machine Learning Repository. They contain data about several physicochemical input variables and a sensory output variable (quality) related to the red and white variants of the Portuguese "Vinho Verde" wine.

The dataset has 11 variables/attributes that are all continuous and numerical. The 11 predictor variables include fixed acidity, volatile acidity, citric acid, residual sugar, chloride, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, and alcohol. The target variable, quality, is a categorical, ordinal variable that serves as a label in this dataset. The classes in the dataset are ordered and not balanced. Information regarding the year, age, brand, or price of the wine was not included in this dataset. The following table provides useful "real-world" information on each variable [7]:

Table 1: Description of Predictor Variables

Predictor Variable	Description
fixed acidity	most acids involved with wine are fixed or nonvolatile (they do not evaporate readily)
volatile acidity	the amount of acetic acid in wine, which at too high of levels can lead to an unpleasant, vinegary taste
citric acid	found in small quantities, citric acid can add "freshness" and flavor to wines
residual sugar	the amount of sugar remaining after fermentation stops. It is rare to find grapes with < 1 g/L and wines with > 45 g/L are considered sweet
chlorides	the amount of salt in the wine
free sulfur dioxide	the free form of SO_2 exists in equilibrium between molecular SO_2 (as a dissolved gas) and bisulfite ion; it prevents microbial growth and the oxidation of wine
total sulfur dioxide	amount of free and bound forms of SO_2; in low concentrations, SO_2 is mostly undetectable in wine, but at free SO_2 becomes evident in the nose and taste of wine
density	the density of wine is close to that of water depending on the percent alcohol and sugar content
pH	describes how acidic or basic a wine is on a scale from 0 (very acidic) to 14 (very basic); most wines are between 3-4 on the pH scale
sulfates	a wine additive which can contribute to sulfur dioxide gas (SO_2) levels, which acts as an antimicrobial and antioxidant
alcohol	the percent alcohol content of the wine

3.2 Data Cleansing

The imported datasets consist of 1,599 red wine samples and 4,898 white wine samples, respectively. Combined, there are a total of 6,497 entries in both datasets. First, the individual datasets were checked for potential missing values and duplicate entries. The results show that there are no missing values in either dataset; however, 240 duplicate rows were detected in the red wine dataset and 937 duplicate rows were detected in the white wine dataset—a total of 1,177 duplicate rows in the combined dataset. In the next step, all duplicated entries were removed. The removal of duplicate rows left us with a red wine dataset of size (1359, 12) and a white wine dataset of size (3961, 12) prior to our train/test split. Finally, in order to detect any potential outliers in either dataset, the datasets were normalized using min-max scalar, then boxplots were drawn for every normalized feature. Data normalization allowed us to plot all features in one graph and visualize the outliers in a more distinctive way. As shown in Figure 1, many of the features in the combined dataset contain outliers that could adversely affect our ability to train accurate, high performing classification models. The Z-score statistical method was used to identify and eliminate these outliers. The acceptable Z-score threshold was set to 3; therefore, each of the rows containing calculated Z-score values above and below 3 standard deviations were eliminated. The total number of outliers detected in the combined dataset was 510 values, leaving the fully cleansed dataset at a size of (4810, 13).

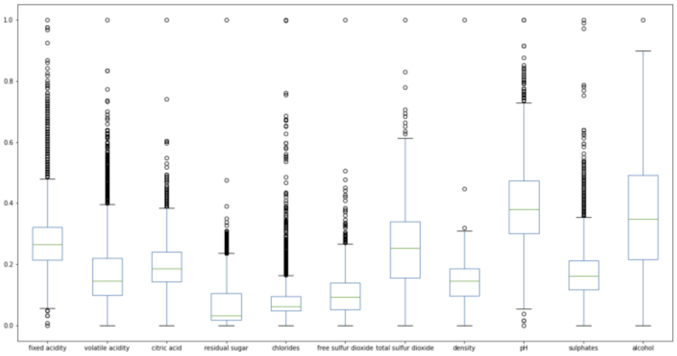

Figure 1: Boxplot Visualization of Combined Dataset Feature Outliers

3.3 Data Exploration

After data cleansing, the red wine dataset consisted of 1,208 samples and the white wine dataset consisted of 3,602 samples with the same initial 12 features per data entry. Although the red and white wine datasets each contain a different number of observations, the datasets share a similar wine quality distribution, as shown in Figures 2 and 3. The bulk of wine quality scores in both datasets fall in the range of 5 to 7. There are no observations below a quality score of 3 and observations above a quality score of 9. Futhermore, the red wine dataset includes no samples that achieved a quality score of 9. Upon further analysis, we calculated the mean quality score of the red wine dataset to be very close to the mean quality score of the white wine dataset. The proportion of high-quality wines is slightly higher in the white wine dataset when compared to the red wine dataset. Another interesting discovery was that although white wines are traditionally regarded as having a lower alcohol content, the mean alcohol level was similar in both datasets. Below are some additional data exploration findings:

- There are no negative or out-of-range values in either dataset.
- pH levels specifying the wine sample's acidity range from 2.8 to 3.8 (acidic).
- The median residual sugar content of white wines is 4.8 g/L which is significantly higher than that of red wines, 2.2 g/L.

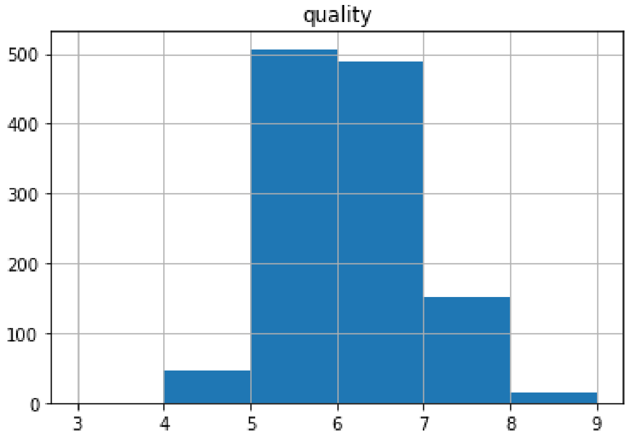

Figure 2: Red Wine Dataset Quality Distribution

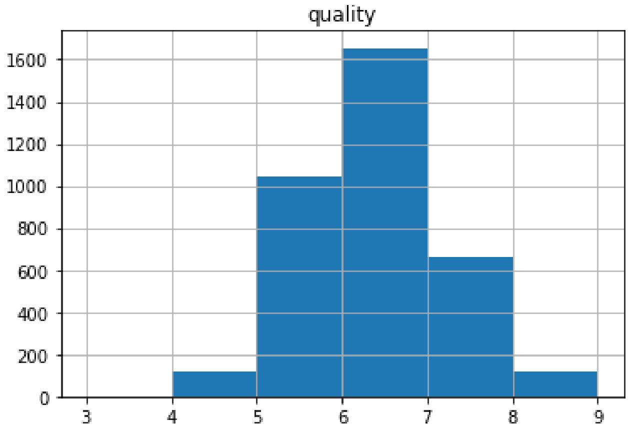

Figure 3: White Wine Dataset Quality Distribution

To determine the relationships between each variable and wine quality, a correlation heatmap was made. There are some noticeable differences in the way that certain variables interact depending on the variety of wine. However, there are no strong correlations between features, specifically between "quality" and other features. Below are some additional observations from examination of the combined dataset heatmap (Figure 4):

- In both datasets, the "quality" feature has the highest correlation with alcohol (moderate positive correlation).
- The correlation between alcohol and sugar content is much higher for white wines (moderate negative correlation, -0.46) than it is for red wines (weak positive correlation, 0.16). This would indicate that "boozy" red wines contain less sugar than less "boozy" red wines while on the other hand, "boozy" white wines have more sugar that less "boozy" white wines. Upon closer inspection of the sugar content in the combined dataset summary statistics table, one can see that the average residual sugar in red wines is much less than the average residual sugar in white wines.

In order to gain meaningful insight from consideration of the combined dataset heatmap, we focused on the relationships indicated by the correlation between "alcohol" and "quality" (moderate positive correlation, 0.48) and the correlation between "density" and "alcohol" (relatively strong negative correlation, -0.7). Below are some comments on those relationships:

- Even though the strongest correlation between "quality" and another feature is the alcohol-quality correlation, the correlation itself is not very strong and therefore, we cannot rely solely on a wine sample's alcohol level to predict the quality of the wine sample.
- Based on the density-alcohol correlation, which is the highest overall correlation between features in our dataset, when the wine sample has a higher percentage of alcohol, the density level of that particular wine sample is typically lower. This observation is based on the fact that a wine sample's density represents the conversion rate of natural sugars to alcohol. The same observation explains the negative correlation between "alcohol" and "residual sugar content".

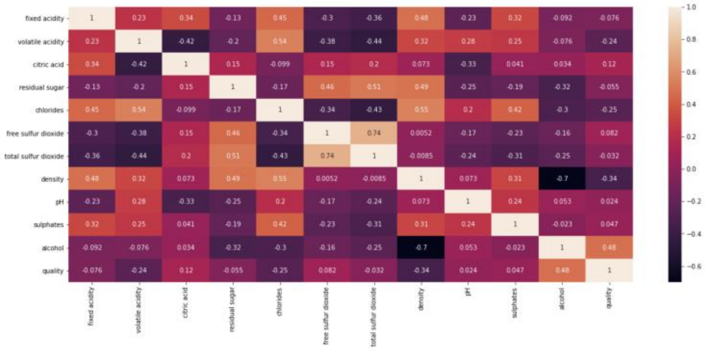

Figure 4: Heatmap of Combined Dataset Feature Correlations

3.4 Data Preparation

When working to prepare our fully cleansed data for training our model experiments, we decided to combine the red and white Vinho Verde wine datasets. A categorical variable "color" was added as the first column in the new dataset and One Hot Encoding was then used to convert the "color" variable to a binary variable. The extra column created in this process was removed. We began training our models using a combined dataset and continued to do so until we read some of the existing literature regarding these datasets. The recent publications regarding work on these datasets supported the approach of training two different models—one based on the red wine dataset and the other based on the red wine dataset [2][5]. In an effort to increase the accuracy of our models, the data was separated back into their respective datasets based on wine color and models were trained using the two separate red and white wine datasets. This separation yielded higher accuracy models. We continued with the separate dataset approach, citing the difference in the taste of red and white wines as an additional motivator. The focus moving forward was to optimize the performance of two different models—one for each wine color.

4 Modeling Experiments

Using "quality" as the target variable, we chose to train a classification model instead of a regression model due to the non-continuous and ordinal nature of the wine quality rankings.

4.1 Optimizing Model Fit and Performance

Grid Search was used to tune hyperparameters in the k-NN and Random Forest classifiers. K-fold cross validation was used to assess how the accuracy of each model's test results would generalize to an independent dataset.

4.2 Dimension Reduction Techniques

Feature selection methods were utilized to determine whether all input variables were relevant to our model. Then each classifier was trained using only the significant features. First, the Automatic Backward Elimination script was used to train models with only significant predictors. This technique eliminated the following variables from the red wine model: "volatile acidity", "chlorides", "free sulfur dioxide", "total sulfur dioxide", and "pH". Only the "density" variable was removed when backward elimination was performed on the white wine dataset. While improving the computational efficiency of each model, Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA), and kernel Principal Component Analysis (kPCA) dimension reduction techniques did not produce a significant increase in model accuracy. We speculate that this is due to the relatively low correlation between features in our dataset.

4.3 Dimension Reduction Techniques

Each model was trained on the complete set of dataset features and then retrained on the reduced sets of features determined by various dimension reduction techniques. The accuracies of each modeling classification technique were compared to determine the model with the highest individual classification performance by measure of accuracy.

4.3.1 Logistic Regression. The Logistic Regression class in sci-kit learn was used to create a classifier that would interpret the probability of each training instance belonging to a certain class and then convert that value to a classification label. Due to the multiclass nature of the datasets, the Logistic Regression training algorithm defaulted to use of the cross-entropy loss technique to minimize the multinomial loss fit across the entire probability distribution. Prior to the use of any dimension reduction techniques, the following mean model accuracies were determined:

Table 2: Accuracy of Logistic Regression Model

Dataset	Mean Model Accuracy	Std. Dev. of Model Accuracies
Red Wine	0.5728	0.0312
White Wine	0.5407	0.0222

4.3.2 k-Nearest Neighbors. The KNeighborsClassifier class in sci-kit learn was used to create a classifier that would predict the class of each new instance based on the majority class amongst the k neighbors. GridSearchCV was then used to determine the optimal number of k neighbors, weights, algorithm, p, and metric. Prior to the use of any dimension reduction techniques, the following mean model accuracies were determined:

Table 3: Accuracy of k-NN Model

Dataset	Mean Model Accuracy	Std. Dev. of Model Accuracies
Red Wine	0.5870	0.0254
White Wine	0.5486	0.0222

4.3.3 Support Vector Machine. The SVC class from sci-kit learn was used to create a classifier that would determine the subset of training points that lies on the border between each of the classes and is important to defining the decision boundary. These points are the support vectors. The accuracy of multiple models trained using each kernel type (linear, rbf, ply, and sigmoid) was calculated and the model with the highest accuracy was selected. The 'rbf' kernel produced the highest accuracy model for both datasets. Prior to the use of any dimension reduction techniques, the following mean model accuracies were determined:

Table 4: Accuracy of SVM Model

Dataset	Mean Model Accuracy	Std. Dev. of Model Accuracies
Red Wine	0.5751	0.0370
White Wine	0.5605	0.0340

4.3.4 Naïve Bayes. The GaussianNB class from sci-kit learn was used to create a classifier that assumes each feature is statistically independent. Prior to the use of any dimension reduction techniques, the following mean model accuracies were determined:

Table 5: Accuracy of Naïve Bayes Model

Dataset	Mean Model Accuracy	Std. Dev. of Model Accuracies
Red Wine	0.5396	0.0435
White Wine	0.4772	0.0180

4.3.5 Random Forest. The RandomForestClassifier class from sci-kit learn was used to create a classifier that builds an ensemble of decision trees using randomly selected subsets of the training data. GridSearchCV was then used to determine the optimal number of decision tree estimators, criterion, and whether bootstrapping should be used. The red wine dataset was trained using 177 decision tree estimators and the white wine dataset was trained using 101 decision tree estimators. Prior to the use of any dimension reduction techniques, the following mean model accuracies were determined:

Table 6: Accuracy of Random Forest Model

Dataset	Mean Model Accuracy	Std. Dev. of Model Accuracies
Red Wine	0.5835	0.0473
White Wine	0.5767	0.0412

4.3.6 Extra Trees. The ExtraTreesClassifier class from sci-kit learn was used to create a classifier that implements a meta estimator that fits a number of randomized decision trees (called extra trees) on various sub-samples of the dataset and uses averaging to improve the model's predictive accuracy and control over-fitting. GridSearchCV was then used to determine the optimal number of decision tree estimators, criterion, maximum features algorithm, and whether out-of-bag samples and bootstrapping should be used. The red wine dataset was trained using 176 decision tree estimators and the white wine dataset was trained using 196 decision tree estimators. Prior to the use of any dimension reduction techniques, the following mean model accuracies were determined:

Table 7: Accuracy of Extra Trees Model

Dataset	Mean Model Accuracy	Std. Dev. of Model Accuracies
Red Wine	0.5988	0.0324
White Wine	0.5759	0.0365

4.3.7 Gradient Boosting. The GradientBoostingClassifier class from sci-kit learn was used to create a classifier that builds an additive model in a forward stage-wise fashion, optimizing differentiable loss functions. GridSearchCV was then used to determine the optimal loss function, criterion, and maximum features algorithm with the use of the default 100 boosting stages. Prior to the use of any dimension reduction techniques, the following mean model accuracies were determined:

Table 8: Accuracy of Gradient Boosting Model

Dataset	Mean Model Accuracy	Std. Dev. of Model Accuracies
Red Wine	0.5728	0.0397
White Wine	0.5450	0.0322

4.3.8 Deep Neutral Network (DNN). The keras module from tensorflow was used to create a deep neural network. It consisted of three dense layers each with tanh activation followed by relu activation for the output layer. Adam optimizer was used for the purpose along with Sparse Cross Categorical Entropy as the loss function. We wanted the training process to run for a fixed number of iterations, so we used the epochs argument and set epochs to equal 100. We found the described combination to be the one that resulted in the best accuracy. Epochs more than 100 had no impact on the DNN and the gradients seem to converge at almost around 100 epochs.

Table 9: Accuracy of DNN Model

Dataset	Mean Model Accuracy*
Red Wine	0.9666
White Wine	0.7305

*subject to change due to randomness

4.4 Experimentation Results

Our first approach was to train a classification model on the combined dataset. After taking the time to understand the nature of our two datasets, we achieved a higher model accuracy when training two separate classification models on fully cleansed data (duplicates and outliers removed).

Table 10: Accuracy of All Models

	Red Wine	White Wine
Model	Accuracy	
Logistic Regression	0.6364	0.5199
k-NN	0.5895	0.5467
SVM	0.6446	0.5486
Naïve Bayes	0.5702	0.4662
Random Forest	0.6143	0.5449
Ensemble	0.6419	0.5439
Extra Trees	0.6446	0.5421

Gradient Boosting	0.6171	0.5458
Log. Regression w/ Backward Elimination	0.5799	0.5423
k-NN w/ Backward Elimination	0.5562	0.5498
SVM w/ Backward Elimination	0.5622	0.5510
Naïve Bayes w/ Backward Elimination	0.5538	0.4863
Random Forest w/ Backward Elimination	0.5775	0.5680
Log. Regression w/ PCA	0.6171	0.5180
k-NN w/ PCA	0.6088	0.5254
SVM w/ PCA	0.6364	0.5273
Naïve Bayes w/ PCA	0.5620	0.5106
Random Forest w/ PCA	0.5840	0.5393
Log. Regression w/ LDA	0.6309	0.5254
k-NN w/ LDA	0.6226	0.5208
SVM w/ LDA	0.6364	0.5328
Naïve Bayes w/ LDA	0.6336	0.5310
Random Forest w/ LDA	0.5758	0.5069
Logistic Regression w/ kPCA	0.6281	0.5227
k-NN w/ kPCA	0.5868	0.5180
SVM w/ kPCA	0.6198	0.5439
Naïve Bayes w/ kPCA	0.6088	0.5208
Random Forest w/ kPCA	0.6501	0.5264

4.4.1 Optimization of Two Quality Classes on DNN and Random Forest. When the red and white wine datasets were kept separated the accuracy of the models did increase. From here, we wanted to see if we could increase the accuracy of the models even more with minimum error by plotting using PCA to look at the data then dividing the data into quality groups. When we divided the data into three quality groups, the accuracy of the models did increase but the individual accuracy of the quality group classes was very low. Next, we divided the data into two quality groups (group 1- bad and group 2- good) and got both high accuracy of the models and high accuracy of the individual quality groups classes. Since most of the overall wine quality is 6, any row belonging to quality 6 that is greater than the mean of all rows in quality 6 is changed to quality 6.5. Group 1 = quality 6 or less and Group 2 = quality 6.5 or greater. Training our models based on a reduced two quality groups/clusters, allowed for a better model accuracy and performance for both DNN and Random Forest.

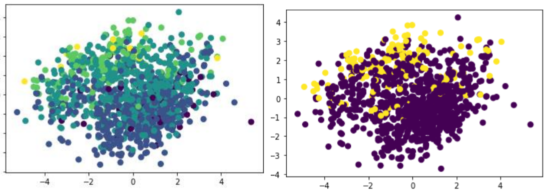

Figure 5: Plotting using PCA to show distribution of classes-Red Wine Dataset: (left) before two quality groups & (right) after two quality groups.

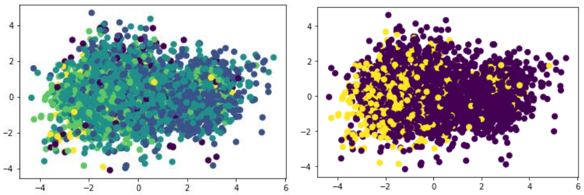

Figure 6: Plotting using PCA to show distribution of classes-White Wine Dataset: (left) before two quality groups & (right) after two quality groups.

As we desired to learn additional ML techniques and further optimize the performance of our models, we decided to also explore both SVM and deep neural network (DNN) after plotting using PCA to show the distribution of classes. According to Kumar, the deep neural network approach could also be used to predict the wine taste preferences. As shown below in Table 11, the mean accuracy for DNN decreased, but the accuracy for SVM increased.

Table 11: Accuracy of DNN and SVM Models After Quality Group Implementation

Dataset	Model	Mean Model Accuracy
Red Wine	DNN	0.9722*
White Wine	DNN	0.9814*
Red Wine	Random Forest	0.8796
White Wine	Random Forest	0.8284

*subject to change due to randomness

4.5 Interpretation of Results/Conclusion

One of the objectives of our study was to determine which of the several chemical properties are most useful in training the most accurate wine quality classification model. After data exploration and comparing the accuracies of different models trained using various dimension reduction techniques, we found that there is no strong correlation between features (specifically between quality and other features) and feature selection only afforded small increases in accuracy. This leads us to believe that the sensory quality of wine determined by human perception can only begin to be accurately predicted with the use of all physicochemical properties in the red and white Vinho Verde wine datasets.

Our primary objective in completing this analysis was to utilize ML algorithms to make the most accurate prediction of human wine taste preference given the physicochemical attributes of a sample. Our models show that performing Random Forest and DNN techniques on a fully cleansed and clustered dataset both perform equally well and provide high accuracy predictions.

5 Discussion

Given additional time and resources we would have liked to have spent more time honing our DNN model to see if we could achieve an even higher accuracy with minimal error by increasing the number of iterations and using a different hidden layer network with different activation functions, as shown in a study from April 2020 [2]. It is important to note that apart from all of the attributes provided in the dataset, there are still a lot of significant features that can impact human perception of wine quality, such as: type of grapes, age, producer, and region. We believe the inclusion of these attributes in this analysis would have allowed us to train a more accurate model. It would be interesting to see how other attributes contribute to the sensory quality of wine besides and/or in comparison to the provided chemical attributes. As technology and machine learning techniques progress, there will be more methods for increasing the accuracy and performance of a model trained with this wine quality dataset, as well as other ways to predict human wine taste preferences. As it pertains to consumer decision making, at present there are multiple applications that consumers can download and utilize as an aid in their wine selection process. These applications are developed by using machine learning and artificial intelligence methods on wine datasets and consumer preferences.

ACKNOWLEDGMENTS

UCI Machine Learning Repository

REFERENCES

[1] Paulo Cortez, Antonio Cerdeira, Fernando Almeida, Telmo Matos and Jose Reis. 2009. Modeling wine preferences by data mining from physicochemical properties. In Decision Support Systems, Elsevier, 47(4):547-553. DOI: https://doi.org/10.1016/j.dss.2009.05.016

[2] Sachin Kumar, Yana Kraeva, Radoslava Kraleva and Mikhall Zymbler. 2020. A Deep Neural Network Approach to Predict the Wine Taste Preferences. Intelligent Computing in Engineering (pp.1165-1173). DOI: 10.1007/978-981-15-2780-7_120

[3] Garima Agrawal and Dae-Ki Kang. 2018. Wine Quality Classification with Multilayer Perceptron. International Journal of Internet, Broadcasting and Communication Vol.10 No.2 25-30.

[4] Dale Angus. 2019. Modeling Wine Quality from Physicochemical Properties. Standard Project. Corpus ID: 209521363

[5] Yesim Er and Ayten Atasoy. 2016. The Classification of White Wine and Red Wine According to Their Physicochemical Qualities. International Journal of Intelligent Systems and Applications in Engineering 4(Special Issue-1):23-23 DOI: 10.18201/ijisae.265954.

[6] Comissão de viticultura da região dos vinhos verdes (CVRVV). 2020. Vinho verde: Like no other wine in the world. Retrieved from: https://www.vinhoverde.pt/en/homepage

[7] Sagnik, Roy. 2020. Red wine quality: A to Z. Retrieved from: https://www.kaggle.com/sagnik1511/red-wine-quality-a-to-z