Aadharsh Pannirselvam and Harris Lencz
Professor Nussbaum
Introduction to Data Science II
9 May 2023

Project Report

Introduction to the Dataset

This dataset originally comes from a scientific article published on June 9, 2009 by Paulo Cortez at the University of Minho, Portugal and other researchers describing a data-driven approach to predicting human taste preferences in wine based on its physicochemical properties.[1] Using different testing methods, Cortez and the researchers compiled a dataset of 11 objective measurements (such as pH, density, citric acid content, and others) for 1599 red and 4808 white variants of the Portuguese "Vinho Verde" wine for a total of 6497 samples of wine. In this particular context, the quality of the wine samples was analyzed by a panel of wine experts, with quantitative measurements taken to understand the characteristics that merited the ratings.

This dataset takes on increasing importance given the growth of the wine industry and efforts by wine producers to continually improve wine quality with new technology. Broadly speaking, wine quality could only previously be assessed in a subjective manner by humans, which also may have been expensive, time-consuming, and inefficient. However, this dataset and others like it now enable more objective predictions and measurements of wine quality to be determined with ease. Moreover, in fact, as we will explain further later in this report, predicting wine quality is not the only machine learning task that this data can be used for; we will predict the color, not quality, of each wine sample based on its physicochemical characteristics.

Literature Review

The data was donated to the UCI Machine Learning Repository on October 7, 2009 and has since amassed over 2 million web hits. Much of the research involving it has, as in the original paper mentioned above, investigated methods to determine wine quality from its physicochemical characteristics using a variety of statistical frameworks. For instance, a group of researchers at Barcelona Tech University used hybrid fuzzy logic techniques to predict human wine test preferences of the wine samples.[2] Another group of researchers from Hanyang University in South Korea used a decision tree to predict wine quality.[3] Moreover, a third group of researchers from the Institute of Engineering and Management in Saltlake, India used support vector machines, random forest classifiers, and multilayer perceptron models to predict wine quality.[4]

---

[1] P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis.
Modeling wine preferences by data mining from physicochemical properties,
https://archive.ics.uci.edu/ml/datasets/wine+quality. In Decision Support Systems, Elsevier, 47(4):547-553, 2009.
[2] Nebot, Àngela et al. “Modeling Wine Preferences from Physicochemical Properties using Fuzzy Techniques.” *International Conference on Simulation and Modeling Methodologies, Technologies and Applications* (2015), https://www.scitepress.org/papers/2015/55519/55519.pdf.
[3] Lee, Seunghan et al. “Assessing wine quality using a decision tree.” *2015 IEEE International Symposium on Systems Engineering (ISSE)* (2015): 176-178, https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7302752&tag=1.
[4] Shaw, Bipul et al. “Wine Quality Analysis Using Machine Learning.” *Advances in Intelligent Systems and Computing* (2019), https://link.springer.com/chapter/10.1007/978-981-13-7403-6_23.

It must also be noted that similar datasets involving wine and its physicochemical characteristics have been used in a variety of other contexts from production economics to food chemistry. For instance, in 2008, Juan-Carlos Ferrer and researchers at The Pontifical Catholic University of Chile used a wine dataset to create a model to optimize the schedule of wine grape harvest operations.[5] In addition, in 2007, Isabel M. Moreno and researchers at the University of Sevilla in Spain used a probabilistic neural network model to differentiate two kinds of red wines based on metal content.[6]

While these are all interesting problems to investigate, we chose to analyze a problem not yet fully explored by the literature: classifying the color of each sample of wine (red or white) based on the physicochemical characteristics of that sample mentioned above.

We decided this would be excellent practice at a binary classification task for which we could implement two of the primary methods we learned in class: Logistic Regression and k-Nearest Neighbors (kNN) classification. In addition, we felt this would be a more interesting problem to investigate than merely predicting the quality of each wine sample, given the class imbalance towards higher quality wines in the dataset. In fact, there are no entries rated lower than a 3/10, further evidence that our pool of wine samples is highly skewed. Thus, we opted to predict color, not quality, for each wine sample.

Exploratory Data Analysis

Figure 1: Correlation Matrix

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality	color
fixed acidity	1.000000	0.219008	0.324436	-0.111981	0.298195	-0.282735	-0.329054	0.458910	-0.252700	0.299568	-0.095452	-0.076743	-0.486740
volatile acidity	0.219008	1.000000	-0.377981	-0.196011	0.377124	-0.352557	-0.414476	0.271296	0.261454	0.225984	-0.037640	-0.265699	-0.653036
citric acid	0.324436	-0.377981	1.000000	0.142451	0.038998	0.133126	0.195242	0.096154	-0.329808	0.056197	-0.010493	0.085532	0.187397
residual sugar	-0.111981	-0.196011	0.142451	1.000000	-0.128940	0.402871	0.495482	0.552517	-0.267320	-0.185927	-0.359415	-0.036980	0.348821
chlorides	0.298195	0.377124	0.038998	-0.128940	1.000000	-0.195045	-0.279630	0.362615	0.044708	0.395593	-0.256916	-0.200666	-0.512678
free sulfur dioxide	-0.282735	-0.352557	0.133126	0.402871	-0.195045	1.000000	0.720934	0.025717	-0.145854	-0.188457	-0.179838	0.055463	0.471644
total sulfur dioxide	-0.329054	-0.414476	0.195242	0.495482	-0.279630	0.720934	1.000000	0.032395	-0.238413	-0.275727	-0.265740	-0.041385	0.700357
density	0.458910	0.271296	0.096154	0.552517	0.362615	0.025717	0.032395	1.000000	0.011686	0.259478	-0.686745	-0.305858	-0.390645
pH	-0.252700	0.261454	-0.329808	-0.267320	0.044708	-0.145854	-0.238413	0.011686	1.000000	0.192123	0.121248	0.019506	-0.329129
sulphates	0.299568	0.225984	0.056197	-0.185927	0.395593	-0.188457	-0.275727	0.259478	0.192123	1.000000	-0.003029	0.038485	-0.487218
alcohol	-0.095452	-0.037640	-0.010493	-0.359415	-0.256916	-0.179838	-0.265740	-0.686745	0.121248	-0.003029	1.000000	0.444319	0.032970
quality	-0.076743	-0.265699	0.085532	-0.036980	-0.200666	0.055463	-0.041385	-0.305858	0.019506	0.038485	0.444319	1.000000	0.119323
color	-0.486740	-0.653036	0.187397	0.348821	-0.512678	0.471644	0.700357	-0.390645	-0.329129	-0.487218	0.032970	0.119323	1.000000

[5] Juan-Carlos Ferrer et al. "An optimization approach for scheduling wine grape harvest operations." *International Journal of Production Economics*, Volume 112, Issue 2, 2008, Pages 985-999, ISSN 0925-5273, https://www.sciencedirect.com/science/article/pii/S092552730700285X.

[6] Isabel M. Moreno et al. "Differentiation of two Canary DO red wines according to their metal content from inductively coupled plasma optical emission spectrometry and graphite furnace atomic absorption spectrometry by using Probabilistic Neural Networks." *Talanta*, Volume 72, Issue 1, 2007, Pages 263-268, ISSN 0039-9140, https://www.sciencedirect.com/science/article/pii/S0039914006007168.

Figure 1: This correlation matrix shows the correlation (r) between all variables in the dataset. Free sulfur dioxide and total sulfur dioxide appear to be collinear with a high correlation r value ($r = 0.72 > 0.70$, a common threshold for collinearity). This makes sense given their underlying chemistry: a higher free sulfur dioxide content implies a higher overall sulfur dioxide content, all other factors equal.[7] It also must be noted that total sulfur dioxide, one of our predictors, seems to be highly correlated ($r = 0.70$) with the color of the wine sample, our binary response variable. We will further investigate this relationship later in the report.

Data Modeling Technique 1: Logistic Regression

We used forward-model selection to identify a logistic regression model, with the default regularization level given by C=1, that would best predict the color of a wine sample from its physicochemical characteristics.

First, we split our dataset into 80% training data and 20% testing data. We ensured that the split was stratified such that our testing data consisted of an equal proportion of white to red wine. Then, in forward-model selection, we obtained logistic regression seven models, as in Figure 1.

Figure 2: Logistic Regression Models obtained from Forward-Model Selection

Model	Features	Accuracy	Precision	Recall	F1 Score	Feature Added
Model 1	total sulfur dioxide	0.906154	0.926441	0.951020	0.938570	
Model 2	volatile acidity, total sulfur dioxide	0.944615	0.955823	0.971429	0.963563	volatile acidity
Model 3	volatile acidity, total sulfur dioxide, sulphates	0.963846	0.972644	0.979592	0.976106	sulphates
Model 4	fixed acidity, volatile acidity, total sulfur ...	0.967692	0.977597	0.979592	0.978593	fixed acidity
Model 5	fixed acidity, volatile acidity, total sulfur ...	0.981538	0.985772	0.989796	0.987780	pH
Model 6	fixed acidity, volatile acidity, chlorides, to...	0.983077	0.986789	0.990816	0.988798	chlorides
Model 7	fixed acidity, volatile acidity, chlorides, fr...	0.981538	0.985772	0.989796	0.987780	free sulfur dioxide

Figure 2: This table shows the features and performance of each logistic regression model as measured by accuracy, precision, recall, and F1 score. As expected from the high correlation seen between total sulfur dioxide and color in the correlation matrix above, the first feature chosen in forward model selection was total sulfur dioxide. This one feature alone allows for the prediction of wine color with approximately 90% accuracy, which we would like to use as our primary performance metric. When volatile acidity and sulphates are added to the model as predictors, accuracy reaches about 96%. There are diminishing returns on accuracy and each of the other metrics once more features are added to Model 7, which indicates that those additional features (like pH and chlorides, for instance) are increasingly less important, and make the model worse in all measures, as opposed to trading one off for the others. Nonetheless, this figure helps to identify Model 6 as the logistic regression model with the highest accuracy at 98.3%.

[7] Moroney, Maureen. "Total Sulfur Dioxide – Why it Matters, Too!" Midwest Grape and Wine Industry Institute, Iowa State University, https://www.extension.iastate.edu/wine/total-sulfur-dioxide-why-it-matters-too/.

As noted in our code file, the equation for Model 6 is logit(p_whitewine) = 32.3 -1.25 * fixed acidity - 8.15 * volatile acidity - 3.80 * chlorides + 0.05 * total sulfur content - 6.17 * pH - 5.41 * sulphates

Figure 2: Confusion Matrix for Logistic Regression Model 6

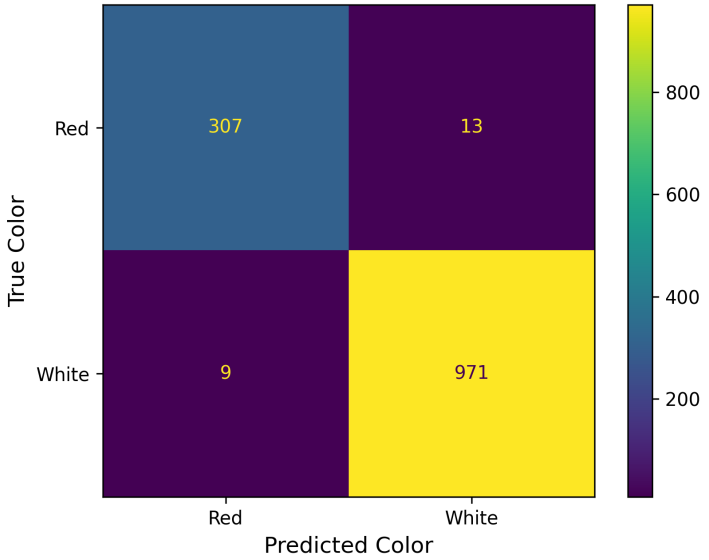

Figure 2: This confusion matrix reveals the counts of true positives, true negatives, false positives, and false negatives where positive indicates that the wine sample is white and negative indicates that the wine sample is red. There were 971 true positives for white wine, 307 true negatives, 9 false negatives, and 13 false positives. Thus, as we saw earlier in Figure 1, test accuracy = (971 + 307)/(971 + 307 + 9 + 13) = 1278/1300 = 98.3%.

Data Modeling Technique 2: kNN Classification

We also used k-Nearest Neighbors (kNN) classification to predict the color of a wine sample from its physicochemical characteristics.

Again, we split our dataset into 80% training data and 20% testing data. We ensured that the split was stratified such that our testing data consisted of an equal proportion of white to red wine. Then, to determine the best k-value for our model, we measured the test accuracy corresponding to k-values ranging from 1 to 50, as shown in Figure 3.

Figure 3: Test Accuracy of kNN Classifier for Different K Values

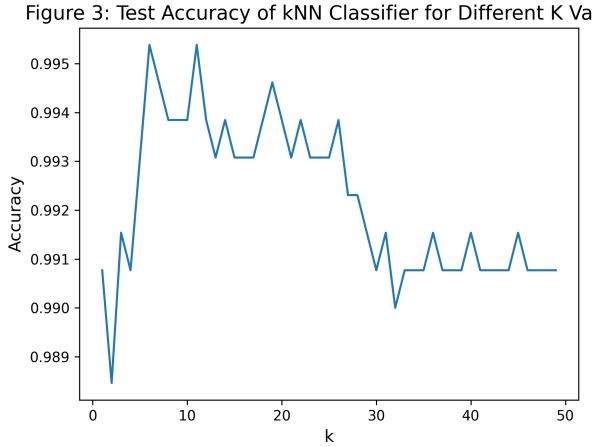

Figure 3: This graph shows the test accuracy of the kNN classifier across models with k-values ranging from 1 to 50. The graph has two equal peaks, one at k=6 and another at k=11.

Since accuracy is maximized equally at both k=6 and k=11, we chose to compute 2 kNN models and compare information presented in their respective confusion matrices, shown below.

Figure 4: Confusion Matrix for kNN Model with k=6

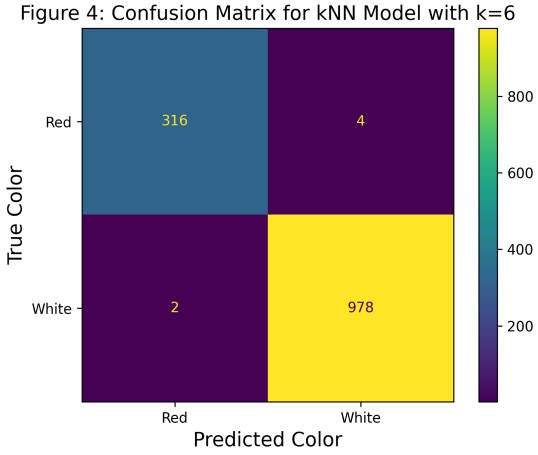

Figure 5: Confusion Matrix for kNN Model with k=11

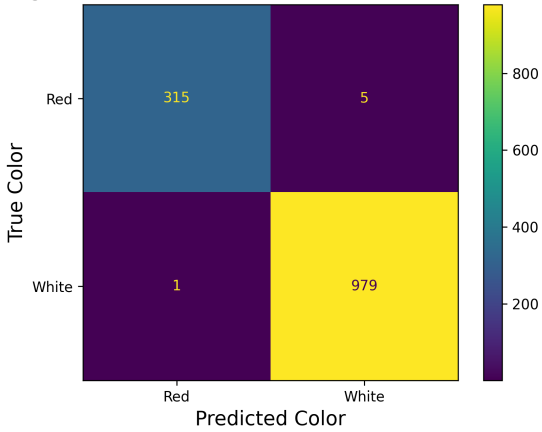

Figure 4: This confusion matrix reveals the counts of true positives, true negatives, false positives, and false negatives where positive indicates that the wine sample is white and negative indicates that the wine sample is red. There were 978 true positives for white wine, 316 true negatives, 2 false negatives, and 4 false positives. Thus, test accuracy = (978 + 316)/(978 + 316 + 2 + 4) = 1294/1300 = 99.5%.	Figure 5: This confusion matrix reveals the counts of true positives, true negatives, false positives, and false negatives where positive indicates that the wine sample is white and negative indicates that the wine sample is red. There were 979 true positives for white wine, 315 true negatives, 1 false negatives, and 5 false positives. Thus, test accuracy = (979 + 315)/(979 + 315 + 1 + 5) = 1294/1300 = 99.5%.

As verified above, the kNN model at k=6 and k=11 both have the same accuracy: 99.5%. Moreover, the confusion matrices above also reveal that the counts of true positives and true negatives for the model at k=6 and k=11 only differ by 1 on a scale of hundreds, meaning that they are almost virtually the same. However, going from k=6 to k=11 changes slightly more significantly the false prediction balance of the kNN model: at k=6, 2/6 = 33.3% of the false predictions are false negatives, while at k=11, 1/6 = 16.7% of the false predictions are false negatives. In other words, the kNN model, when going from k=6 to k=11, is a little less likely to predict a false negative than a false positive in circumstances where it is incorrect.

Conclusion

In Data Modeling Technique 1: Logistic Regression, our best performing model (Model 6) had an accuracy of 98.3%.

In Data Modeling Technique 2: kNN Classification, our best performing model(s), equally well-performing at k=6 and k=11, had an accuracy of 99.5%.

Overall, we can say that both techniques produced models with the ability to predict the color of a wine sample from its physicochemical characteristics with very, very high accuracy. This is exciting news and has great promise to researchers and those working with wine! However, one limitation with getting such high accuracy metrics is that the headroom for fine-tuning our models is rather limited, which could prove troublesome when applying the model to new data.

We can also say that our kNN Classification model(s) had a slightly greater accuracy than did our Logistic Regression model, even with many standardized dimensions. We suspect this might be due to there being distinct subcategories within the two groups of wine, white and red, that lend themselves well to non-parametric analysis via a kNN model. Perhaps future research could investigate this matter further and shed light on the extent to which non-parametric approaches may outperform parametric approaches for similar datasets.

Team Statement:

We split the work evenly, as we each contributed to aspects of the code and the report. We collaborated nicely and enjoyed working through how to present technical information in a clear and compelling manner.

Dataset Citation:

P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis.
Modeling wine preferences by data mining from physicochemical properties,
https://archive.ics.uci.edu/ml/datasets/wine+quality. In Decision Support Systems, Elsevier, 47(4):547-553, 2009.