Quality of Wine Prediction Using Classification UC San Diego - ECE 225 Report

Ehsan Ziaeikajbaf*, Sepehr Foroughi Shafiei†
Jacob school of engineering, M.S.Eng
UC San Diego, CA
Email: *eziaeika@ucsd.edu, †seforoug@ucsd.edu,

***Abstract*—One of the most important facts about white wine is it is known as the preventer of heart diseases. On the other hand, some people think red wine helps people to have a longer lifespan. In this project, we would like to predict the quality of red and white wine based on the feature classes that were provided in our data-set such as pH, alcohol, residual sugar and chlorides. Our data-set is from UCI and consists of 12 attributes for each data entry. We are planning to start some exploration throughout our data and make sure to have a balanced data-set to avoid over-fitting and under-fitting problems. After data cleaning, we will try to find out which features have the most impact on our prediction by visualizing data. Then we are planning to use three machine learning algorithms: Random Forest, Logistic Regression and SVM to make classification. At the end, we are going to conclude our data exploration by comparing the result of each model and create the best graphs possible to compare our results. Also we will use different prediction metrics like F1-score and accuracy to obtain better results.**

I. Introduction

The United Sates is one of the most high consumer and importer of wine in the world. Based on International Food and Agribusiness Management Review, The United states is the fourth largest wine producer. When it come to wine, some people prefer red wine than the white wine. The question is which one is better? One group of people might say they like red wine because it has a better test and it's healthier and on the other hand some people might not be agree with that. Wine is in the alcohol family. There are a lot beneficial facts about wine. Companies would like to know which one is better to make their client like the product and recommend the product to other people based on scientist facts. My team is going to use some supervise machine learning techniques to predict which one is more popular based on the data that was gather from north of Portugal winery "Vinho Verde". We are going to figure out which wine is better by comparing correlated chemical components of wine. We are going to do that by figuring out what's the good ratio to make an amazing wine.

II. DATASET

1) Data set description:
The data-set was downloaded from UC Irvine machine learning data-set [1]. Our Dataset divided into 2 individual datasets. First one is Red wine Dataset consist of 1599 data and second one is White wine dataset consist of 4898 data. Both of the datasets contains 12 features. The Table I contains some of the information that we used for this research paper.

Table I
Dataset Basic Description

	Name	Description
1	fixed acidity:	Most acids involved with wine or fixed or nonvolatile
2	volatile acidity:	The amount of acetic acid in wine
3	citric acid:	Found in small quantities
4	residual sugar:	The amount of sugar remaining after fermentation stops
5	chlorides:	The amount of salt in the wine
6	free sulfur dioxide:	It prevents microbial growth and the oxidation of wine
7	total sulfur dioxide:	Amount of free and bound forms of S02
8	density:	The density of water, alcohol and sugar content
9	pH:	Describes how acidic or basic a wine is
10	sulphates:	Acts as an antimicrobial and antioxidant
11	alcohol:	The percent alcohol content of the wine
12	quality:	Output variable (score between 0 and 10)

2) Data Visualisation:
In this part we can look at different relations among different features, it is very useful to have an insight of coloration of features as it helps us in determining the effect of different elements in starting stages only. We created all the Bar-Graphs for each of our features in both white and red wine dataset and we displayed only three of them in this paper.

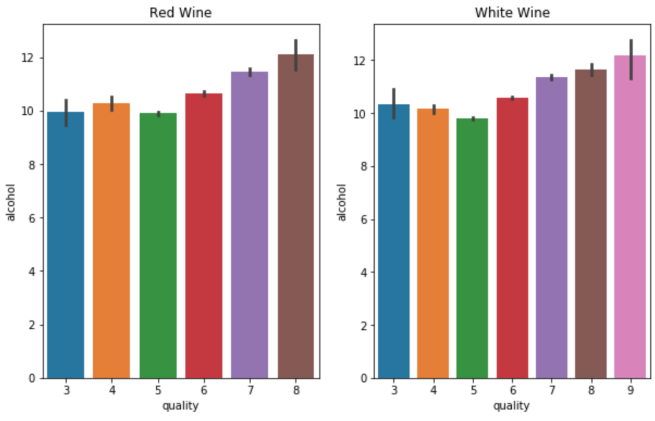

Figure 1. **Alcohol vs. Quality**

Based on Figure 1 we can say this distribution improves with increase in Alcohol quantity for both red and white wine.

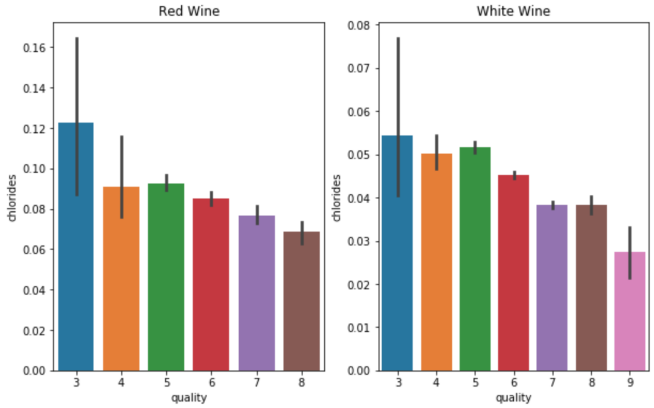

Figure 2. **Chlorides vs. Quality**

From Figure 2 we can clearly see that decrease in chlorides corresponds to better wine quality for both of the wins. Also based on the values of the chloride we are getting the ratio of chloride of white wine is half of the red wine so basically red wine has chloride double of the white wine to reach the same quality.

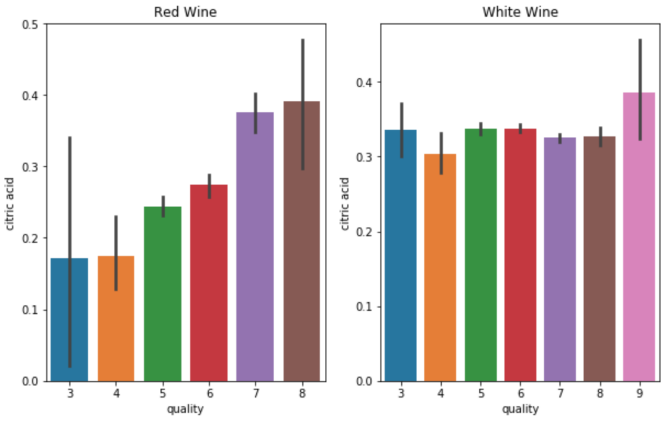

Figure 3. **Citric Acid vs. Quality**

In the Figure 3 for Citric Acid vs quality we can note that once the value of citric acid increases the quality of red wine improves and when the values reaches 0.4 we are getting the best quality of red wine but for white wine our graph is mostly constant compare to the red wine. We can observe that increasing citric acid won't change the overall result for white wine.

III. Models

In this part of our analysis, we studied the data preprocessing and data cleaning technique in order to prepare our dataset for our Machine Learning algorithm then we explained the training models with their behaviors.

1) Data pre-processing:

At first, we obtained the basic information from each of our datasets to find out if there is any null or empty or zero data within our columns and it turned out, in both of the datasets we did not have any null data. Then we looked at our tables to find out how is our data within each feature is look like. The reason behind exploring and understanding data was to gather as much as information in order to create and predict the best accuracy models.

Second, we observed how well our datasets statically distributed. For doing that we used the Panda data to describe the function. In this method, we were able to gather the main statistical information: Count, Mean, Standard Deviation, Minimum, Maximum, and 25,50, and 75 percent percentile. This information's obtained from all of 12 feature classes and performed on both of the datasets. This technique helped us to understand how our dataset actually look likes and helped us with our training model analysis. After researching some cleaning techniques, we looked at the problems that often happen on datasets.

• **Outlier Problem:** Outlier is one of the big problems that can happen during every analysis. Outlier happens when some of our data fall very far from the majority of the dataset and outside of the range of natural or expected values. In this case, the outlier can cause noise in our training models which will affect our prediction. After carefully review the distribution since we have limited number of data and in order to get high accuracy, we need all the data. We used the actual data that was gathered from real wine in both dataset. We did not observe any outlier therefore for this part we did not have to use any outlier techniques.

• **Correlation:** Since we have 12 features and classes for these datasets, each feature has its own range therefore, we needed to rescale the range to unify all the features. As a result, we scaled down all the features to be between the range of [0,1] and to find the relations between each feature and to get the significant perspective that how our features correlated, therefore, we graphed the correlation heatmap for red wine and white wine dataset classes. Based on Figure 4, red wine correlation was between -0.7 and 0.7 and for white wine was between -0.8 and 0.8. Figure 5. In both cases, if the correlation reaches toward the maximum it means that those two features were linearly correlated, and when goes toward the minimum it means that they did not have any relevance. As a result, we observed that the top features that correlated with the quality class which is the main feature for White Wine are: Alcohol, PH, Sulphates, and Free Sulfur Dioxide and for Red Wine are: Alcohol, Citric Acid, Sulphates, and Residual Sugar.

Third, after all of the exploration and gathering information, we prepared our dataset for the training model. In this part, we decided to choose quality as our predictive task. Since quality for both datasets are in the range between 3 to 8 and 3 to 9, for using quality as a predictive task for our models we had to classify the quality between 0 and 1. For that purpose, we used a bucketing system where we chose the range of 6.1 to 8 as good quality and below 6.1 as a bad quality for the Red Wine dataset. For White Wine the range of 6.1 to 9 represent

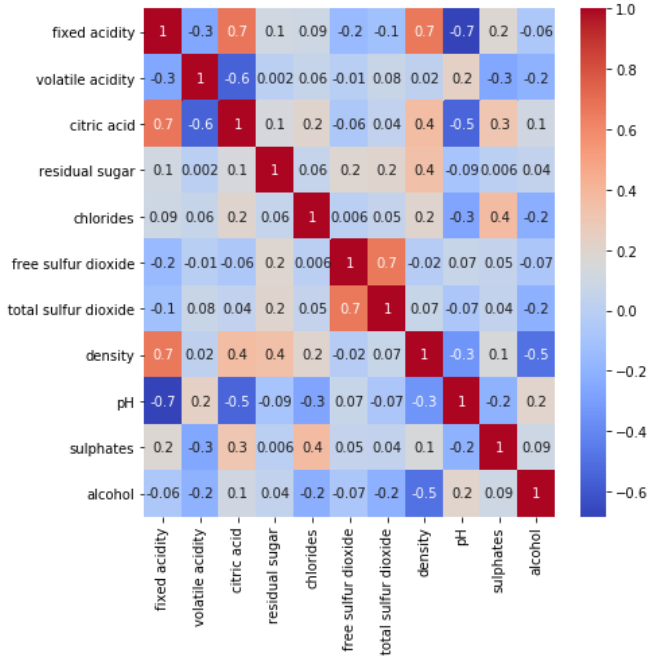

Figure 4. **Red Wine Correlation Map**

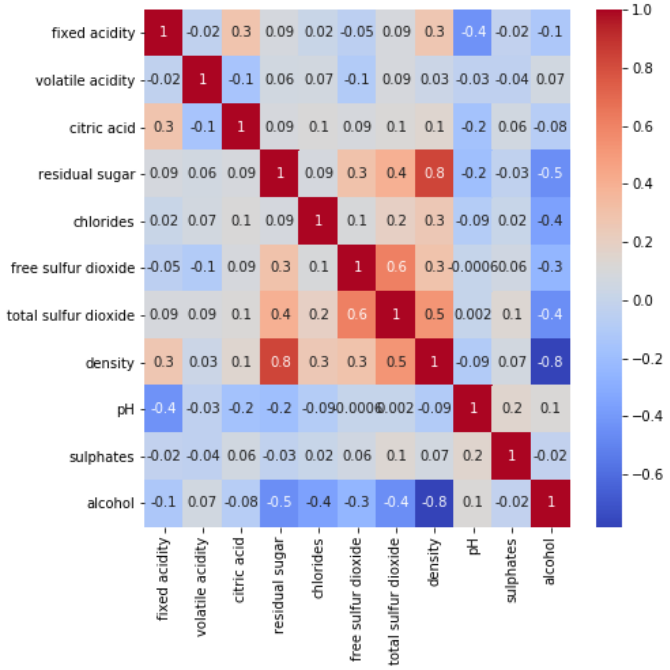

Figure 5. **White Wine Correlation Map**

good quality and anything bellow 6.1 is not good.

$$PredictingQuality(y) = \begin{cases} Good(1) & \text{if x is between [6.1,8]} \\ Bad(0) & \text{Otherwise} \end{cases}$$

After the classification, we graphed the bar plot to observe the number of good and bad wines as it can be seen in Figure 6.

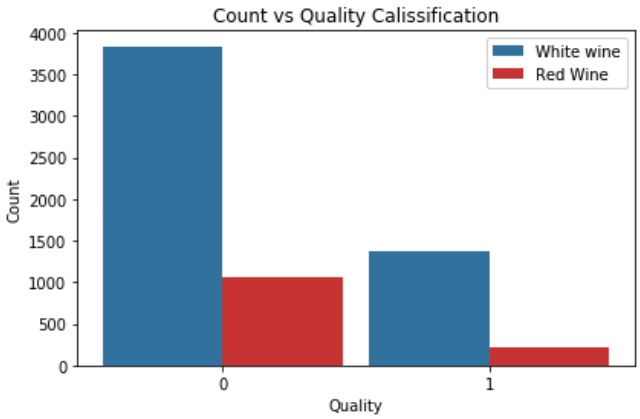

Figure 6. **Quality Classification**

Finally, in order to have a proper prediction, we split our datasets into train and test set by 25 percent which means that the dataset is divided into 2 parts. We used 75 percent of data for the training and 25 percent was used to test the data so we can predict the training result against the test data.

2) Logistic Regressionc :

After going over all the preprocessing, now we are ready to design and train our data. The first model we decided to use was Logistic Regression. we used this model as our baseline model. Since we classified our quality as 0 and 1 (binary) Logistic Regression would be the closest training model to use. This model always trains True or False classes. In our case, we classified between good and bad. Logistic regression always explains the relation between one class to another class or multiple nominals. We used all the default settings for our model since we needed the baseline results. Another observation about Logistic Regression is model fit. Considering that we were fitting 11 features into the Logistic Regression, this model increases the amount of variance with having more features, and also, we had a limited amount of data in our datasets, as a result, this can cause an overfitting problem[2].

3) Random Forest:

One of the classic classification training models is Random Forest. This classifier consists of a large number of decision trees that work together as a group. A decision tree is an intuition that simplified how each tree works and technically is 0 and 1 classier within each branch. The prediction is true or false and to count all the decisions, we cover separately each one of them. Each single of these decision trees has its own decision and vote. In the end, the decision of each tree which is either 0 or 1 becomes the majority of vote then after that, we can conclude that the majority votes are our final prediction. Random Forest is a powerful model because it works best in uncorrelated models. Since the decision trees are separate and they are not continuous, it works perfectly for the models with many features and uncorrelated. The

key for Random Forest is that when it tries to build each individual tree ,it creates an uncorrelated forest based on the randomness that we chose. In our case since we have 12 features and also many of the features did not have correlation, Random Forest worked perfectly. Another big factor that Random Forest helps us with, that always uses the averaging to control over-fitting problem and improve accuracy. In order to choose the perfect estimator, we had to perform experiments at first, we chose 50 estimators then we increased that to 100, 150, 200, 250, and 300. We recorded the results. The number estimator means the number of decision tree in the forest. We used same fitting technique as Logistic Regression for choosing predictive task and rest of the feature. At the end 200 has the best accuracy [3].

4) Support Vector Machine(SVM):
At last, in order to explore more complex and better accuracy, we chose another classifier called Support Vector Machine. SVM is a supervised learning model that can be used for both classification and regression. This training model is responsible to find the best possible decision boundary between classification classes, which in our case are 0 and 1. Between each class can be an infinite number of the line but SVM makes sure that to choose the best possible line that maximizes the margin between two classes. The line is intuition that we use but in general, they are hyperplane. SVM chooses the hyperplane that has the greatest margin between the nearest point of each class to the line. It can be possible that sometimes it cannot find a good decision boundary as our data point from each class either very close or very mixed. This model works really fine with smaller and cleaner datasets. SVM uses the subset to find the decision boundary, it is more efficient. Therefore, we used SVM as our third and final training model [4].

IV. Results

In this section, we go through all the results we gathered from each model. The result of each part is based on the metrics that we believed would be the best tools to evaluate our models. Choosing proper metrics to evaluate each result is a major point in every research paper.

In this analysis, we chose the Accuracy and F1 Score to evaluate our models. Accuracy is the basic metric tool that is the result of a total number of correct predictions over the total of all the predictions. In other words, the sum of true positives and the sum of true negatives over all the predictions. F1 score is one of the greatest metrics to conclude the results. F1-score is a combination of precision and recall Figure 9 which is going to balance them out by taking the harmonic mean. F1 score is a good metric when we have imbalance classes because it uses both False-negative and False positives in the calculation.

we calculated the result based on classification reports. In the Logistic Regression Model, the Accuracy for Red wine was 88.5 percent and for White wine was 79.1percent and for F1-score metrics, our result for Red wine was 87 percent and

$$\text{F1-score} = \left(\frac{\text{Recall}^{-1} + \text{Precision}^{-1}}{2}\right)^{-1} = 2 * \frac{(\text{Precision} * \text{Recall})}{(\text{Precision} + \text{Recall})}$$

Figure 7. **F1-Score Formula**

for White wine was 76 percent. We observed that the accuracy and F1-score for our red wine were higher than white wine. This can be resulted due to lack of data on red wine which was 3 times less than White wine also the quality of Red wine with the features that we classified was more accurate. another reason we can observe here that Logistic Regression is a perfect model for the smaller datasets.

Afterward, we evaluated the Random Forest Classifier. For Random Forest we were able to obtain an accuracy of 88.75 percent for Red wine and an accuracy of 87.8 percent for White wine. This can show a huge improvement in our White wine with almost 10 percent increase and results for F1-score was noted 87 percent for Red wine and the same number for White wine. We observed that our F1-score also improved by 10 percent since Random Forest is a great model for the uncorrelated dataset. It proved that we could get a better result. Finally, we evaluated the SVM model. We did not achieve any improvement for our Red wine accuracy and the result was 88.75 percent and our White wine was 81.87 percent. the results showed that we had a little improvement for our White wine accuracy compare to Logistic Regression. However, Random Forest still was better than SVM. Comparing our F1-score, SVM was able to achieve 86 percent for Red wine and 79 percent for White wine which we had the same conclusion as accuracy.

Below is our bar plot for comparing the result of metrics for each model

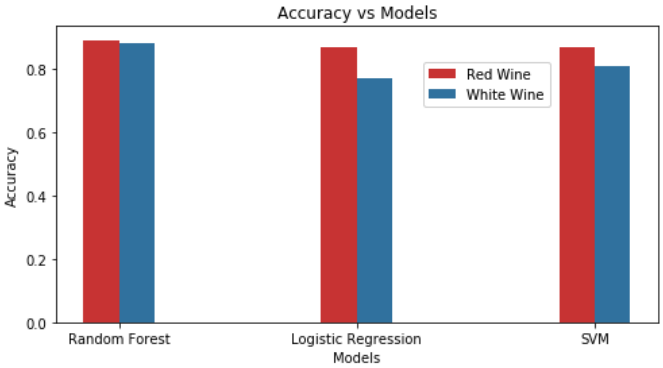

Figure 8. **Accuracy**

V. Conclusion

The capability of a using machine learning to investigate various chemical components on a wine dataset to predict wine quality was the motivation behind this paper. Based on the results that we collected we achieved a higher accuracy for Red Wine. One of the reasons that we couldn't get a better result was that we only had 1600 data available for

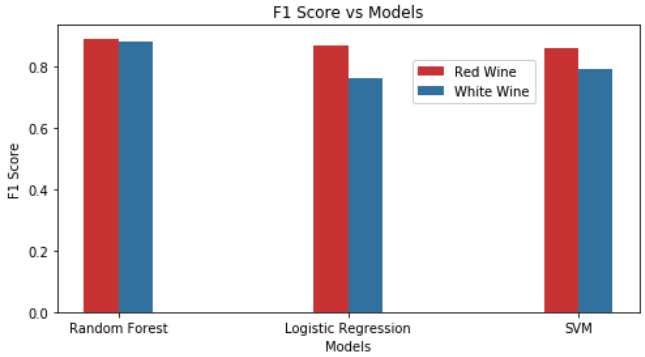

Figure 9. **F1-Score**

red wine and that could result into underfitting if the model or algorithm was low variance but high bias. We used about 3 times more data for white wine than the red wine. This advantage for white wine give us a better model in terms of realistic dataset. One thing we had to pick was choosing the right machine learning technique base on our specific dataset. This is one biggest challenge in machine learning because choosing the right algorithm can give us a more robust scores. For example in this research we end up using SVM because of the complexity of the data which increased the efficiency of the model however, Random Forest Classification gave us better result since we had higher number of uncorrelated features. Based on our models we were able to achieve decent results which can be used for some other wine dataset to predict whether it is a good or bad wine with an accuracy of 89%. Based on comparison between red and white wine we found out red wine has a higher quality and it's healthier. For the future study we think by adding more dataset into our current data we can improve the accuracy. We also think it will be useful to detect outlier and remove it which can improve the accuracy of our model.

REFERENCES

[1] D. Dua and C. Graff, "UCI machine learning repository," 2017. [Online]. Available: http://archive.ics.uci.edu/ml

[2] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[3] "3.2.4.3.1. sklearn.ensemble.randomforestclassifier¶." [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

[4] R. Gandhi, "Support vector machine - introduction to machine learning algorithms," Jul 2018. [Online]. Available: https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms