Exploration Into the Quality of Vinho Verde Variations

Ed Pasfield

I. Abstract

This paper outlines the major correlations in the two data sets on Vino Verde (Red/White) through the use of linear regression, evaluating these thoroughly. It introduces these explanations into real world situations connecting them with their hypotheses defined in part of the introduction. The second part of this paper carries out the implementation of a binary classifier to determine whether an instance of a wine is deemed good/bad quality, again for both red/white variations and explores why they managed to do this to a reasonable standard. After the analysis of the results of both sections, the conclusion of the paper suggests that the white data set had better correlations and worked better at classifying due to its larger data set but not necessarily because of the model defined or the data. This is where more statistical facts are analyzed and explained such as, Kappa statistic, Mattehews Correlation Coefficient and the ROC/AUC. A final statement on how a future study into density of wines and how it could affect the industry leads these findings into a new possible direction.

II. Introduction

Vinho Verde originted in Minho province in the north of Portugal and is reknowned to be the countries most famous wine, the province used has been extended over the past 50 years to produce more wine. Vinho Verde translates to 'young wine' and it has many variations of rose, red, white and sparkling. The wine industry has been trying to keep up with modern techniques to improve not only the brewing, but the consumer market as well. One of the techniques which is being improved is the quality assessment, this is carried out to define the contributing factors and which are most influential and to scale the wines into a more interpretative form. This scale can be used to help pricing wines, and can also determine how to make wine that more people will prefer, a great development in consumerism for the wine industry. Firstly it is important to understand that taste is a sense which is subjective and therefore the human experts which influence when it is used in producing a quality assessment it is difficult to use it with scientific reliability. However what can be trusted on a more scientific level is the physical composition of Vino Verde which has 10 factors including density, alcohol and ph being some of the more important in terms of quality.

Due to developments in the field of data analytics there are many models in which the data on these wines can be plotted. A data model organizes elements of data and stratifies them to enable them to relate to one another [1]. The models that are used in this paper are a linear regression to predict numeric values from the data (alcohol levels, acidity, density, sulphates and residual sugars) and a binary classification model to predict one of two value of the possible 'quality' of the wine either good or bad. The factors which are deemed important and are used more in regression and classification models come from a logistic regression.

III. Data

The two data sets on the physical composition and quality of Vino Verde are based on red and white wines separately. The red data set has 1599 instances of Vino Verde whereas the white data set is significantly bigger with 4898. This means that the white data set will be more reliable as more accurate averages and therefore conclusions will be able to be extracted. The attributes for both datasets are identical including sulphate levels, acidity levels, alcohol levels and more along similar lines. These attributes are all on numeric scales and even though the quality is categoric by nature, the categories are integers. This led to the decision to use classifiers and regression as methods of analysis. After this was decided Data cleansing had to occurr this meant the attributes and data had to be extracted and be prepared for the different software used in the analysis process. The software used depended on what analysis was needed, for instance when dealing with classification WEKA was used due to its ease of use and interpretable outputs. However for data which is interpretable in itself such as averages, ranges and even basic plots like bar charts microsoft excel was more than competent.

The first step in cleansing the data was to convert it into a format where values had their own individual cell, for excels usage. This was done through the use of python where a program stripped all the semi-colons and quotation marks and replaced them with appropriate characters. This meant the two data sets where in comma separated value (CSV) format allowing them to be easily used by Microsoft's software. The second step was to prepare it for WEKA, this meant it needed to be in the correct Attribute-Relation File Format (ARFF) format. ARFF describes a list of instances sharing a set of attributes and was developed by the machine learning project at the department of computer science of the University of Waikato.[4] It's use is for the WEKA machine learning software and is similar to CSV; the only difference being in the attributes meaning it could be edited manually by adding the relation name, and defining the attribute types

TABLE I
SUMMARY OF DATA SETS

	RED			WHITE		
	min	max	mean	min	max	mean
fixed_acidity	4.6	15	8.29	3.8	10.3	6.85
volatile_acidity	0.12	1.33	0.525	0.08	0.65	0.27
citric_acid	0	0.76	0.261	0	0.74	0.33
residual_sugar	0.9	4.7	2.27	0.6	26.05	6.40
chlorides	0.012	0.147	0.079	0.009	0.092	0.043
free_sulfur_dioxide	1	57	15.69	2	112	35.1
total_sulfur_dioxide	6	165	44.93	10	344	137.84
density	0.99	1.0021	0.997	0.987	1.0029	0.994
pH	2.86	3.9	3.32	2.72	3.82	3.19
sulphates	0.33	1.22	0.64	0.22	0.97	0.49
alcohol	8.4	14	10.44	8	14.2	10.54
quality	3	8	5.64	3	9	5.9

along with where the data started. Things vital to making the file work with WEKA; specifically version 3.8.3. Once this had all been prepared there were four files, two in ARFF format and two in CSV. However the third step is to remove outliers/extreme values from the data as to not skew the results unfairly. This was done through the use of WEKA by using two filters on the data, one to find the InterQuartileRange (IQR). Once the IQR is applied to the data it identifies the outliers (data just outside the box plot created by the IQR) and then the extreme values (data which is deemed to far outside the box plot to be an outlier). The identification of these two new attributes allows them to be removed from the data sets entirely, white and red respectively. To remove these from the data sets the filter RemoveWithValues allows you to remove all the values which relate to the data in a class, more specifically in this scenario the data in outliers and extreme values. This will improve the prediction process in almost all cases, however it poses a question of whether it reduces the integrity of the results; which will be further explained. Below are summaries of the data sets after the data cleansing, and their attributes to give a greater understanding of the range and average of each attribute. As you can see in table 1, the data summaries give great context when discussing each attributes at a further point; an effective point of reference. Figure 1. in terms of quality gives an overview of its proportions in the form of histograms.

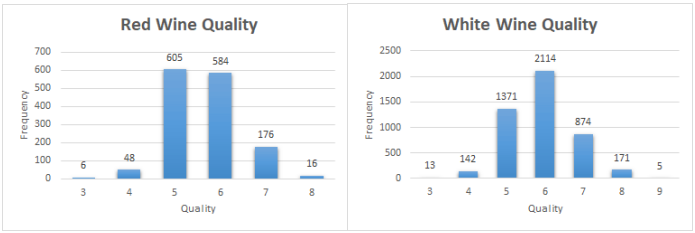

Fig. 1. Wine Quality Histograms

IV. HYPOTHESIS

After the major correlations were discovered hypotheses started to form, there were two major ones for each type of wine (White and Red). The whites hypotheses are; the lower the alcohol content the denser the wine; and as the residual sugar increases as should the density. The reds hypotheses are; the denser the wine the higher the fixed acidity will be; and there will be a strong correlation between total sulfur dioxide and the free sulfur dioxide showing as one increases as does the other, which should be obvious to an extent. The final hypothesis is a model will predict quality through the implementation of a classifier, which is the process of predicting a class of given data points [5]. It will be able to do this better for white wine as there is more data for it to train on. In concordance with these models there will be an evaluation of the models themselves and the data collected from them.

V. RESULTS

VI. LINEAR REGRESSION

Linear regression is a technique which allows a prediction of a value for a numeric attribute to occur, this is done by fitting a model to data which is then used for prediction. The data sets used have many attributes, to be able to implement linear regression the attributes being used have to have a correlation of sorts, having a better correlation induces a more accurate prediction. To determine these correlations in the data sets logistic regression was applied to the entirety of each data sets on WEKA, which on the visualize tab showed all the graphs of all the relations. From this the major correlations were noted and the hypothesis could be further formed or re-thought for each data set. These major correlations for the red wine and white wine could then be turned into linear regression models, to predict a numerical attribute based off the correlations alone. Even though WEKA was used to find and originally plot these following models, the excel graphs are clearer which is why they were used at the final output stage.

VII. RED WINE LINEAR REGRESSION

A. Density Vs Fixed Acidity (FA)

This models attributes had a correlation coefficient of 0.6748, which is not great however does show positive correlation to some extent. The model however did have a Root Mean Squared Error(RMSE):

$$\text{RMSE} = \sqrt{(SSres/n)} = \sqrt{(E[Y2]E[XY]2/E[X2])}$$

which equated to 0.0013 being really low which gives insight into the almost non existent spread in the values of y around the regression line[6]. At a simpler level shows the predictions are close to their real values. An attempt to improve the model was attempted to try and make this a log-linear model which turned out to be futile and ended up making the correlation worse.

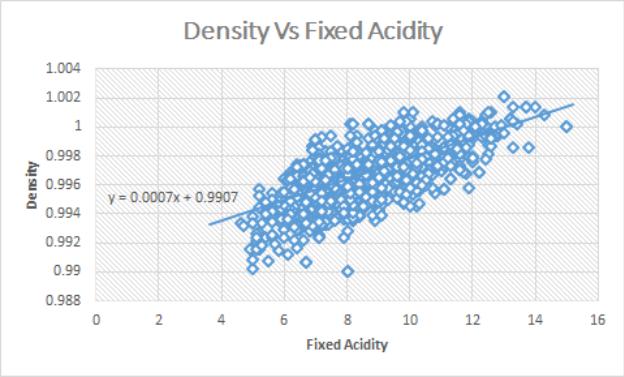

Fig. 2. Density Vs Fixed Acidity Model

$$y = 0.0007 \times fixed\_acidity + 0.9907$$

This figure does show the range of values plotted around the trend line(a representation of the model). The points are spread out due to the low correlation, however it does show that the denser the wine the more the fixed acidity increases even if it is not at a very fast rate.

B. Total Sulfur Dioxide Vs Free Sulfur Dioxide

This is a relationship which naturally should be a positive correlation if the assumption that the free sulphur dioxide is a contributing factor to the total sulfur dioxide; which it is. This model had a positive correlation of 0.7848 showing that the correlation is relatively positive. The RMSE was 0.4247 which in this case is really high and shows that even though the correlation was strong there was a big spread around the regression line meaning the predictions were not very close to their true values. This correlation was at its highest when finding the natural logarithm for both attributes, making it a log-log linear regression as to make the relationship clearer to interpret.

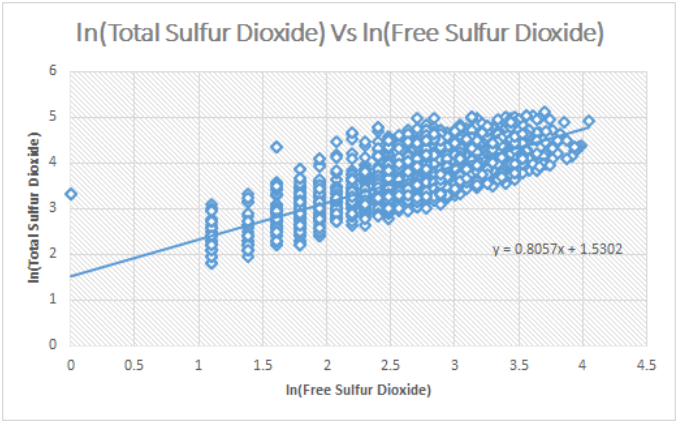

Fig. 3. ln(Total Sulfur Dioxide) Vs ln(Free Sulfur Dioxide)

$$y = 0.8057 \times ln(free\_sulphur\_dioxide) + 1.5302$$

This model (figure 3) does prove the hypothesis between the two attributes, and the assumption between the two is confirmed.

VIII. White Wine Linear Regression

A. Density Vs Alcohol

Figure 4. is a model which compares density and alcohol presents one of the better non-linear regressions in this exploration into the attributes. The relationship is more linear and better at predicting when the natural log of alcohol is put against density instead. This was a suspicion due to the shape of the visualization of the original WEKA outputted graph which had a more logarithmic nature. The correlation coefficient for this log-linear regression model is 0.8186; display a strong positive relationship. Alongside a RMSE of 0.0038 the predictions are very close to their real value solidifying the depiction of the correlation.

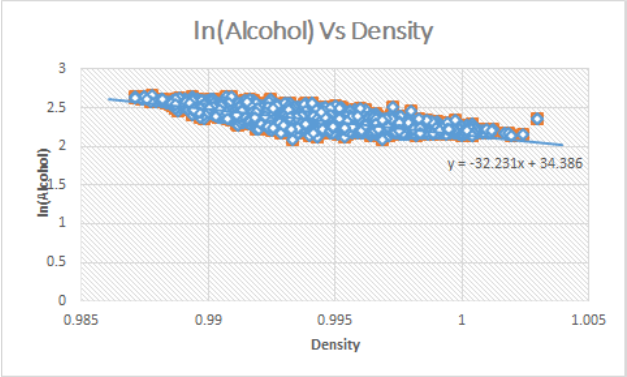

Fig. 4. Density Vs ln(Alcohol)

$$y = -32.2312 \times density + 34.3857$$

As shown in figure 4 the data points follow the trendline really closely, this is compared with figure 3 is the difference between a reliable strong positive relationship which would give accurate predictions and a weak untrustworthy one which would worse predictions. This justifies why the hypothesis of the lower the alcohol level the denser the wine, one could infer that the 'thinner' the wine the weaker the wine.

B. Residual Sugar Vs Density

The non-linear regression model shown in figure 5. which compares residual sugar against density is the best relationship between the attributes with a correlation coefficient of 0.8365. This was improved to this point by changing the model to a linear-cubic one, which was the most effective modification that could be made to provide the most linear, strongest coefficient and lowest RMSE (0.0038). The low RMSE means that the predictions are very similar to the original values proving the model works effectively. To produce figure 5. excel was used at the end point as it is a clearer graph, with a small bit of data modification making a new attribute through excels efficient formula copying to apply the cubic scaling to all instances.

$$y = 0.0015 \times residualsugar + 30.9728$$

The plot in figure 5. depicts the strongest correlation in in the data sets, this is a good model which not only

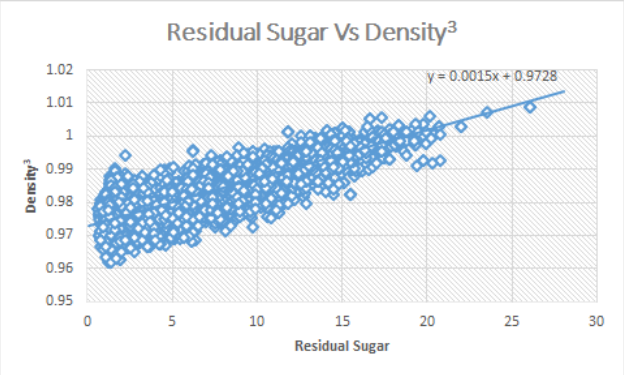

Fig. 5. Residual Sugar Vs Density^3

predicts accurately but does it reliably as well. It confirms the hypothesis that the more residual sugar the more dense the wine, proving the fourth and final linear regression to be true.

To summarise these linear regressions, these are most of the correlations in the data sets with a few extra more trivial ones e.g. ph against fixed acidity. The correlations might not have been as strong as hoped for, however the low RMSE's do show that most the models work effectively at predictions. Density is a common factor between lots of attributes in the data sets, if there was to be further study into this area it would be the relationship between density and the physical attributes of the wines. There is promise for some interesting inferences from that area, that could help with the consumerism of the wine if done effectively. The link between the two white linear regressions would also be interesting to look into a.k.a. the relationship between alcohol and residual sugar and if it does rely on density to link them; a multilinear regression model may be appropriate to explore this area to a significant level.

IX. Quality Classifier

When building a classifier to investigate the level to which the quality can be determined by the multiple chemical compositional attributes the attribute quality itself needed to be assigned the type categorical even though it only consisted of integers. This was because classifiers predict what class an instance will be placed into, and as quality was out of ten. However seven were only ever used in the white data set and six in the red, meaning that the classifier when originally made was attempting to put an instance into one of these seven/six bins, 3-9/3-8. This turned out to be problematic as there was not enough correlations between the attributes to predict this effectively. The quality of the red and white wine prediction accuracy's were 53.75% and 60.49% respectively. This was not good enough, even after trying many different models such as NaiveBayes. NaiveBayes uses the Bayes theorem to predict membership likelihoods for each class and uses that to assign the instance[7]. Because the near 60% accuracy was not good enough, the next logical step was to rethink how to split the classes that the instances were being put into, which lead me to a binary classifier. This only has two classes 'good' or 'bad' taking values from 6-9 and 3-5 respectively. This logistic classifier was compared against others, NaiveBayes, Decision trees and more. However the only one that competed was the nearest neighbour classifier WEKA defines as Instance Based Classifer(IBK), this was nearly the classifier used because it was more accurate at near 80%. However the logistic was used due to its lower RMSE and higher kappa statistic, which will be explained later. A logistic binary classifier was the final decision. Both data sets had no missing values, which meant when running the classifier there was no noise; meaning it was also robust. The decision to use a 10-fold cross validation also was made due to the increased amount of training it cause which was useful as the data sets were lacking slightly in instances, especially the red. The 10-fold cross validation split the data into 10 equal pieces and looped through each pieces to use as an evaluation set using the other 9/10's as the training data; the most reassuring training method possible for this model. This was now a reliable classifier with the best possible outputs to be evaluated.

X. Evaluation of Classifiers

A. Red

When running this classifier the data needed a small amount of preprocessing again, getting rid of the original quality attribute and the empty outlier and extreme values was also important. The summary of the stratified cross-validation is shown below in figure 6 which was produced by WEKA after analysis.

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        1077               75.0523 %
Incorrectly Classified Instances       358               24.9477 %
Kappa statistic                          0.4976
Mean absolute error                      0.3346
Root mean squared error                  0.411
Relative absolute error                 67.4155 %
Root relative squared error             82.5024 %
Total Number of Instances             1435

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
                 0.730     0.232     0.726     0.730     0.728     0.498     0.827     0.775     0
                 0.768     0.270     0.772     0.768     0.770     0.498     0.827     0.857     1
Weighted Avg.    0.751     0.253     0.751     0.751     0.751     0.498     0.827     0.820
```

Fig. 6. Red Quality Classifier Summary

The results here are positive, with an accuracy of 75.05% on such a subjective opinion of 'quality' of wine is impressive. The Kappa statistic - 0.4976, which takes into account guess work done by the classifier[8] shows that the predictions are strong as the higher the kappa the better. The RMSE is high at 0.411, this shows that the predictions are not very close to their true values which is strange because of the relatively high accuracy of the prediction itself. The confusion matrix in figure 7. shows the true positives/negatives and the false positives/negatives rates. The information in the confusion matrix primarily Shows us the divisions of how accurate the classifier. All of the rates can be used to calculate the detailed accuracy by class information in figure 6. The confusion matrix is also used to provide the Matthews Correlation Coefficient (MCC)

which describes the correlation between the prediction and the observation. It gives an indication of how good a prediction is on a scale between -1 and 1, where 0 is no better than random [10]. As the red classifier has an MCC of 0.498, there is room for improvement however it does hold up well.

```
=== Confusion Matrix ===

   a   b   <-- classified as
 479 177 |   a = 0
 181 598 |   b = 1
```

Fig. 7. Red Confusion Matrix

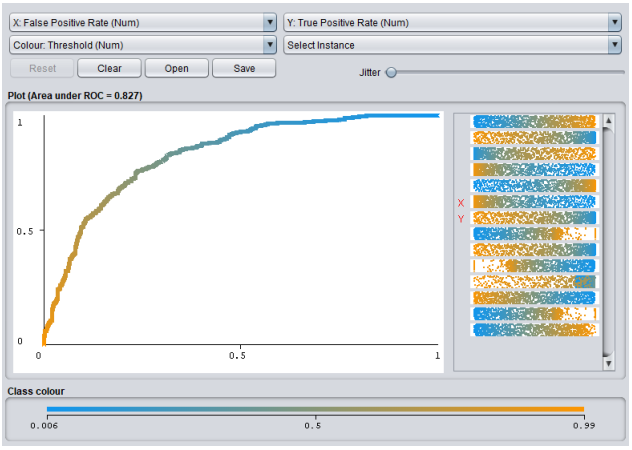

Fig. 8. Red ROC Curve

Figure 8. is a Receiver Operating Characteristic curve (ROC) which is a common way to visualize the performance of a binary classifier[9]. The tighter the curve holds to the top right corner the better the classifier is, this specific ROC for the red variation of vinho verde shows that it is really effective. To reinforce this the area under the curve is the percentage of the graph under the line, this quantifies the performance of the classifier and for this case it was 0.827 which quantifies the prior visualization assumption that the classifier was good. The ROC curve was produced by WEKA simply by selecting visualize threshold curve on the results list.

B. White

When implementing the white vino verde quality classifier, just as in the red one the original quality, the empty outlier and extreme values had to be removed. The summary of the classifier's stratified cross validation is below in figure 8. this shows all the information WEKA provided on the outcome of the classifier and can determine how effective it was. Figure 9. shows that the classifier correctly classified 75.23% of all instances, this is a good amount and means it works well.

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        3531               75.2878 %
Incorrectly Classified Instances      1159               24.7122 %
Kappa statistic                          0.3942
Mean absolute error                      0.3298
Root mean squared error                  0.4069
Relative absolute error                 75.1081 %
Root relative squared error             86.8427 %
Total Number of Instances             4690

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
                 0.484     0.117     0.665     0.484     0.560      0.404     0.800     0.649     0
                 0.883     0.516     0.780     0.883     0.828      0.404     0.800     0.889     1
Weighted Avg.    0.753     0.387     0.743     0.753     0.741      0.404     0.800     0.811
```

Fig. 9. White Quality Classifier Summary

The Confusion matrix in figure 10 is a simpler way of looking at the proportions of correctly/ incorrectly classified instances and can be used to calculate all information in the summary in figure 9 such as the RMSE which in this case was slightly better than the red wine at 0.4069. The Kappa statistic - 0.3942 in comparison to the red is much lower and therefore much worse with its guess work, a common problem in classification methods. The matrix also allows WEKA to provide the MCC - 0.404, this insinuates that the predictor is worse than the previous.

```
=== Confusion Matrix ===

   a    b   <-- classified as
 738  788 |   a = 0
 371 2793 |   b = 1
```

Fig. 10. White Confusion Matrix

The ROC curve for the white wine quality classifier (figure 11) interestingly shows that the classifier was similar in its effectiveness by looking at the shape and the scales to which they are bound. However figure 11 and figure 8 have a bigger difference in the AUC value, with the white wine ROC's AUC being 0.8004 meaning that when quantified it is deemed worse at prediction; not to take away from its effectiveness. This is interesting because the expectation was for the white classifier to be more effective due to its larger data set and therefore larger quantity of training data. An oversight which will have to be explored, but will probably be down to the much lower Kappa statistic and the guess work of the model. This ROC curve was produced in the same manner. Even though the accuracy is better the kappa and the MCC claim its the worse of the two classifiers.

XI. Conclusions and Context

A. Linear Regression Models

In conclusion these models show some really important relationships between attributes. In general the correlations between the white vino verde's attributes were stronger both being above 0.8 (quite strong), a common factor to both of these log-linear and linear-cubic regression models is density. The RMSE for both whites is really low aswell showing the accuracy of the predictions. The white wine hypotheses can be considered true due to these results. Secondly the red vino verde which were log-log and linear regression models show

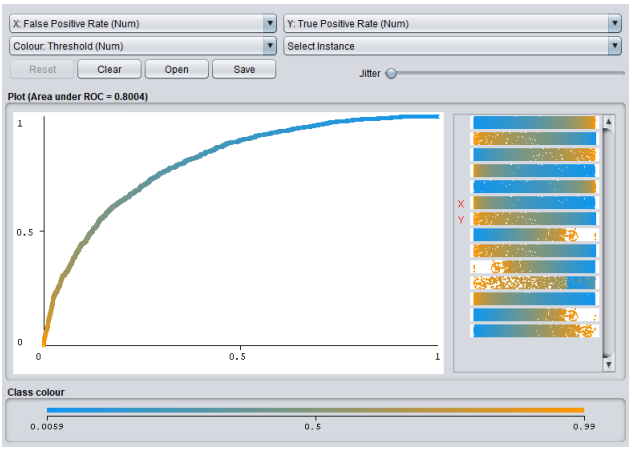

Fig. 11. White ROC Curve

that red wine is harder to predict numeric values for attributes based on their other components. The correlation coefficient for density against fixed acidity was not good enough and the RMSE for the total sulfur dioxide against free sulfur dioxide was way to high to consider the predictions accurate. This suggests that the hypotheses for the red wine are not confirmed, and also means that red vino verde's attributes are more likely to be more independent. Although the regression models worked as intended to show these relationships, the data for the red wine did not show what it was expected to, whereas the white performed as assumed.

What is an interesting topic to further study into is the impact density was on wines, as it is used in 3/4 main correlations between both data sets it is a good assumption that it does affect wine more than others. An exploration into other wine variations and whether density affects them in similar ways or even possibly more so could be enlightening for the production and consumerism of wine as a whole.

B. *Classifiers*

The conclusion that the classifiers can indeed predict quality of wine, did come true even though the scale of which quality was identified had to be changed. The binary classifiers were robust, accurate and worked to a very similar level. Table 2 below shows the comparison between the two on a statistical level:

TABLE II
SUMMARY OF VINO VERDE QUALITY CLASSIFIERS

Quality Classifier	Red Vino Verde	White Vino Verde
Correctly Classified Instances	75.05%	75.29%
RMSE	0.411	0.407
Kappa	0.498	4.7
MCC	0.498	0.404
AUC	0.827	0.8004

This table clearly shows that even though the amount of correctly classified instances and the RMSE are better for the white classifier, overall the red one worked better in terms of the MCC and the ROC/AUC. What can be assumed from comparing all of these is that they work to a very similar standard for different reasons. This is probably due to the different sizes in data sets, even though the red predictor works better the white classifier is trained more which is why it classifies at a higher accuracy. What has not changed and is affecting the classifier most is the non-scientific justification of quality, if it was a more set stratified scale for certain reasons the classifier would of had a much easier job at classification.

C. *Further Studies*

A study into why wine is defined at a certain quality is interesting, for other products there is more factual reasoning behind the level of its quality, thread count, fat percentage, etc. However for wine in particular does not have this as much, which is strange as it is a major import/export from countries all over the globe. If further studies were to be taken into this field a strong suggestion would be density against quality against many different wines. This was concluded as a good area to go into due to the mix between the linear regression models and classifiers produced in this paper, summarising that density is the biggest contributing factor in a wine and quality can be classified with enough specific data. To figure this relationship out could revolutionize the wine industry.

XII. ACKNOWLEDGMENT

The UCI machine learning depository deserves acknowledgment for this paper due to the data it provided and its usefulness and Paulo Cortez, University of Minho who was the source of this data. [11]

XIII. REFERENCES

[1] "What is a Data Model? — Center for Data, Analytics and Reporting", Cedar.princeton.edu, 2018. [Online]. Available: https://cedar.princeton.edu/understanding-data/what-data-model. [Accessed: 27- Dec- 2018].
[2] S. Ray, Modelling nitrogen and carbon cycles in Hooghly estuary along with adjacent mangrove ecosystem. The University of Burdwan: The University of Burdwan, 2015, pp. 306-307.
[3] S. Jrgensen, Fundamentals of ecological modelling, 2nd ed. Amsterdam: Elsevier, 1994.
[4] "Attribute-Relation File Format (ARFF)", Cs.waikato.ac.nz, 2008. [Online]. Available: https://www.cs.waikato.ac.nz/ml/weka/arff.html. [Accessed: 28- Dec- 2018].
[5] S. Asiri, "Machine Learning Classifiers Towards Data Science", Towards Data Science, 2018. [Online]. Available: https://towardsdatascience.com/machine-learning-classifiers-a5cc4e1b0623. [Accessed: 28- Dec- 2018].
[6] S. Holmes, "RMS Error", Statweb.stanford.edu, 2000. [Online]. Available: http://statweb.stanford.edu/ susan/courses/s60/split/node60.html. [Accessed: 29- Dec- 2018].
[7] R. Saxena, "How the Naive Bayes Classifier works in Machine Learning", Dataaspirant, 2017. [Online]. Available: http://dataaspirant.com/2017/02/06/naive-bayes-classifier-machine-learning/. [Accessed: 30- Dec- 2018].
[8] M. McHugh, "Interrater reliability: the kappa statistic", PubMed Central (PMC), 2012. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3900052/. [Accessed: 30- Dec- 2018].
[9] K. Markham, "ROC curves and Area Under the Curve explained (video)", Data School, 2014. [Online]. Available: https://www.dataschool.io/roc-curves-and-auc-explained/. [Accessed: 30- Dec- 2018].

[10] B. Matthews, COMPARISON OF THE PREDICTED AND OBSERVED SECONDARY STRUCTURE OF T4 PHAGE LYSOZYME. Elsevier Scientific Publishing Company, 1975.

[11] P. Cortez, "UCI Machine Learning Repository: Wine Quality Data Set", Archive.ics.uci.edu, 2009. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/Wine+Quality. [Accessed: 31- Dec- 2018].