A Study of the Physicochemical Constitution of Red Wine

Determination of the physicochemical constitution of highly rated red wine samples

Kumaran Suraj, Mayam Gautam, Nikhil Chate, and Rabael Shoaib

Harrisburg University of Science and Technology

Abstract

As wine consumption has become more mainstream, the demand for high quality produce has driven wine producers to turn to data-driven insights to alter their production schedules and methods. Using a novel dataset, showcasing the quality ratings of various vinho verde red wine samples and their associated physicochemical characteristics, various data models are built by viewing the problem as a regression problem at first and as a classification problem later. A random forest model is proposed as an effective decision making tool for wine makers to produce new wine samples that could potentially be highly rated. Insights based on a decision tree model delve into the specifics of the physicochemical constitution of highly rated red wine samples.

A Study of the Physicochemical Constitution of Red Wine

Determination of the physicochemical constitution of highly rated red wine samples

A wide range of audiences have taken to wine consumption of late. In order to support the rapid growth in demand, the wine industry is increasing its investment in new technologies, both in the production and sale of wine. Certification and assessment of the quality of new samples are imperative in this context as they prevent the illegal adulteration of wines and ensure the supply of high quality wine.

The quality evaluation phase of the certification process is important in improving wine making as a whole and also in differentiating different tiers of wines, which is useful in setting prices. The certification process includes both physicochemical and sensory tests. Physicochemical tests general include measurement of density, pH, and other chemical properties, while sensory tests are dependent on human experts.

IT advancements have facilitated the collection, storage, and processing of massive, highly complex datasets. These datasets have invaluable information, which can be utilized in facilitating the decision making process and optimizing the chance of success. Data-driven modelling approaches provide an avenue for extracting information, which is generally not clearly visible, from raw data.

The domain of supervised learning can be seen in the form of two problems: regression problems and classification problems. For the purposes of this study, several regression algorithms, such as linear regression, generalized linear model, and K-nearest neighbors regression, were explored by first viewing the problem as a regression problem. Taking another stab at the problem as a classification problem, several classification algorithms such as logistic regression, K-nearest neighbors classification, decision tree, and random forest were utilized.

Data-driven techniques are useful mainly in the production phase of wine making. Prior research has focused mostly on very small datasets, because of which the large scale implementation of derived insights and research findings has been limited. One study that stood out was the 2009 paper, ‘Modeling wine preferences by data mining from physicochemical properties’ authored by Cortez, Cerdeira, Almeida, Matos, and Reis. This study contributed a dataset of 1599 vinho verde red wine samples, which included 11 physicochemical characteristics of the samples along with their quality ratings as assessed by CVRVV. All prior studies have been able to establish with reasonable certainty that the physicochemical constitution of wine has an important role to play in determining its quality.

Our study attempts to take another look at the vinho verde red wine dataset to determine how different data-modeling approaches could potentially help wine makers alter their production schedules to manufacture superior quality red wine based on the physicochemical composition of previously rated red wines. Through our research, we try to re-validate the hypothesis that the physicochemical constitution of wine plays an important role in determining its quality and hence its perceived rating. The performance of various data-modeling techniques has also been explored.

Method

The 2009 paper, ‘Modeling wine preferences by data mining from physicochemical properties’, authored by Cortez, Cerdeira, Almeida, Matos, and Reis contributed one of the biggest available datasets in the field of study of the physicochemical constitution of wine. The study included two separate datasets, one for red wine samples and the other for white wine samples of vinho verde, a unique product from the Minho region of Portugal. Vinho verde wine is medium in alcohol, and accounts for 15% of the total Portuguese wine production.

The study limits itself to the analysis of the vinho verde red wine dataset, which included 1599 samples. The data was collected from May 2004 to February 2007 using only protected designation of origin samples that were tested at the official certification entity (CVRVV). The CVRVV is an inter-professional organization with the goal of improving the quality and marketing of vinho verde.

Table 1 presents a description of the different variables contained in this dataset and Table 2 presents the summary statistics of the dataset.

Sl. No.	Variable	Description
1.	Fixed Acidity	Grams of tartaric acid per dm3 of sample
2.	Volatile Acidity	Grams of acetic acid per dm3 of sample
3.	Citric Acid	Grams of citric acid per dm3 of sample
4.	Residual Sugar	Grams of residual sugar per dm3 of sample
5.	Chlorides	Grams of sodium chloride per dm3 of sample
6.	Free Sulfur Dioxide	Milligrams of free sulfur dioxide per dm3 of sample
7.	Total Sulfur Dioxide	Milligrams of total sulfur dioxide per dm3 of sample
8.	Density	Ratio of mass to volume of sample in g/cm3
9.	pH	Measure of acidity/alkalinity of sample
10.	Sulphates	Grams of potassium sulphates per dm3 of sample
11.	Alcohol	Percentage volume of alcohol in sample
12.	Quality	Rating of the sample by CVRVV

Table 1: Dataset Description

Sl.	Variable	Min	Q1	Median	Mean	Q3	Max
1.	Fixed Acidity	4.6	7.1	7.9	8.32	9.2	15.9
2.	Volatile Acidity	0.12	0.39	0.52	0.5278	0.64	1.58
3.	Citric Acid	0	0.09	0.26	0.271	0.42	1
4.	Residual Sugar	0.9	1.9	2.2	2.539	2.6	15.5
5.	Chlorides	0.012	0.07	0.079	0.08747	0.09	0.611
6.	Free Sulfur Dioxide	1	7	14	15.87	21	72
7.	Total Sulfur Dioxide	6	22	38	46.47	62	289
8.	Density	0.9901	0.9956	0.9968	0.9967	0.9978	1.0037
9.	pH	2.74	3.21	3.31	3.311	3.4	4.01
10.	Sulphates	0.33	0.55	0.62	0.6581	0.73	2
11.	Alcohol	8.4	9.5	10.2	10.42	11.1	14.9
12.	Quality	3	5	6	5.636	6	8

Table 2: Summary Statistics

Regression Approach

In order to determine whether the physicochemical characteristics affect the quality response variable, the problem at hand was viewed as a regression problem where the quality variable was seen as a continuous response to the mix of physicochemical characteristics as defined by a regression model. The frequency distribution of the quality ratings showed that most of the ratings are concentrated around 5 and 6 (Figure 1). For variable selection, a correlation plot was plotted (Figure 2), which shows the correlation of each and every variable in the dataset to every other variable in the dataset. The correlation plot indicates that the alcohol variable has the highest positive correlation with the quality response variable and the volatile acidity variable has the highest negative correlation with the response variable, indicating that these independent variables could potentially be highly significant in the regression model to be built.

Significantly high correlations were observed between several sets of independent variables – fixed acidity and citric acid, fixed acidity and density, fixed acidity and pH, volatile acidity and citric acid, citric acid and pH, free sulfur dioxide and total sulfur dioxide, and density and alcohol. The associated multi-collinearity can be removed through variable selection by checking the variance inflation factor of each of these variables in context. However, the stepwise procedure to eliminate variables was favored in our study to the method of variable selection based on VIF.

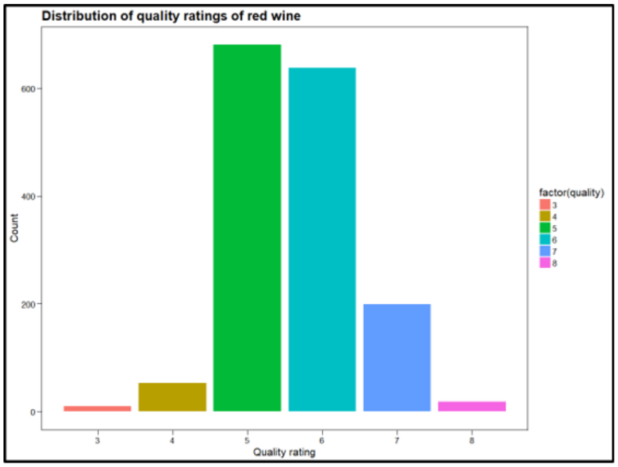

Fig. 1: Distribution of quality ratings

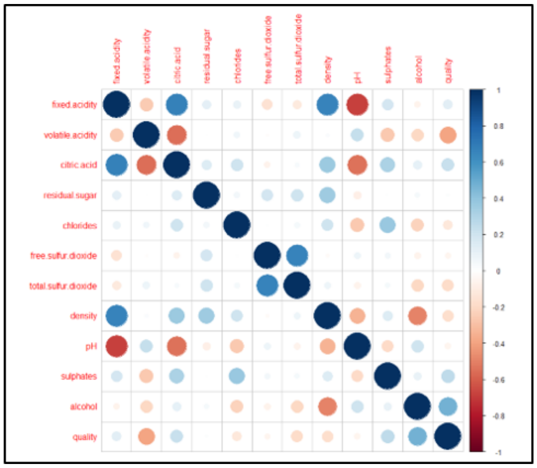

Fig. 2: Correlation plot of variables

Linear Regression

As the first approach to modeling, a multiple regression model, which regressed all the independent variables against the quality response variable, was built. Results from this model provided initial insights into the dependence of quality ratings on the physicochemical composition of red wine samples. Variables such as alcohol, volatile acidity, chlorides, total sulfur dioxide, sulphates, free sulfur dioxide, and pH, which were previously deemed to have high correlation with the response variable, stood out as significant variables in this model. A p-value of $2.2e^{-16}$ for the F-test indicated that quality ratings are in fact dependent on the physicochemical composition. However, the adjusted R^2 value of 0.3561 indicated that this model explains very little of the variation of quality ratings. To improve this model, a stepwise procedure was applied to the above linear regression model to eliminate insignificant variables and improve the adjusted R^2 value. The new model based on stepwise regression, despite encompassing just the above mentioned highly significant variables, did not provide much of an improvement over the previous model with an adjusted R^2 value of 0.3567.

Generalized linear model

The failure of linear regression model to explain the variation in the quality ratings to a satisfactory extent prompted the use of generalized linear models, coupled with stepwise procedures to eliminate insignificant variables. Two GLM models, one belonging to the Gaussian family and other belonging to the Poisson family were built to see if there was any improvement over the linear regression models. However, none of these provided any real improvement as can been seen from Figure 3 and Figure 4. The figures represent the predicted v/s actual plots for these two models. Figure 3 compares the Gaussian family GLM model, represented by triangles, to the linear regression model based on stepwise regression, represented by the red dots. Figure 4 compares the Poisson family GLM model, represented by triangles, to the Gaussian family GLM model, represented by the blue dots.

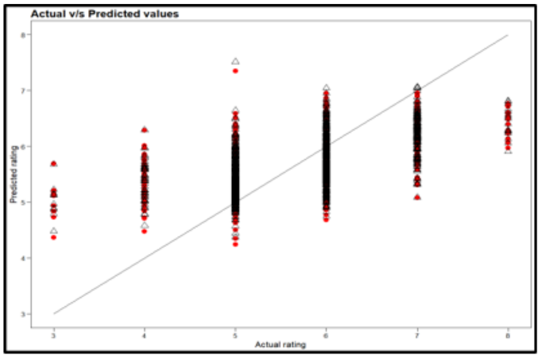

Fig. 3: Gaussian GLM v/s Stepwise LR model

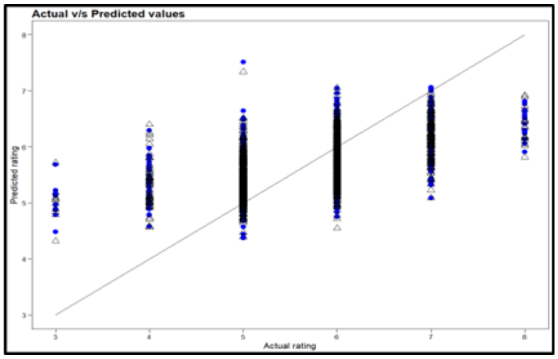

Fig. 4: Poisson GLM v/s Gaussian GLM

K-nearest neighbors regression

Further exploration of the feasibility of modeling wine quality ratings against physicochemical variables using regression was conducted by building KNN models, with 5 and 15 neighbors respectively. However, none of these models again provided any real improvement as can be seen from Figure 5, which compares the KNN model with 15 neighbors, represented by triangles, against the KNN model with 5 neighbors, represented by the green dots.

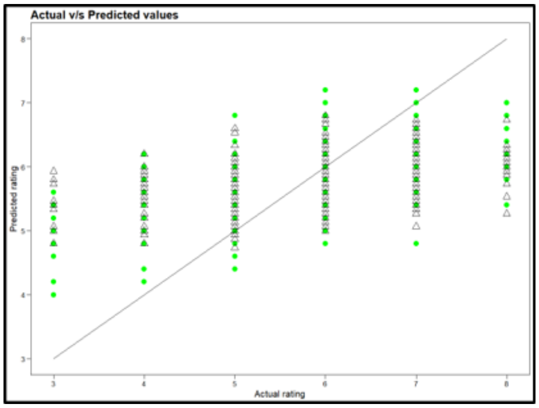

Fig. 5: KNN-15 v/s KNN-5

Classification Approach

None of the regression problems really gave a good fit because of the low variability among the physicochemical characteristics of individual wine ratings. The failure of traditional regression approaches to satisfactorily explain the variation in the dataset forced us to look at the problem from a different angle. Wine makers are not as interested in knowing in advance the exact ratings of their wine samples as to how much they are interested in knowing whether the sample will be perceived as good or bad. To this extent, samples from the dataset with ratings below 7 were classified as bad wine samples and samples with ratings 7 and above were classified as good wine samples. The new response variable, good_tag, was created as a binary classifier, which takes only 1/0 values. Figure 6 shows the distribution of good and bad wine samples in the dataset. The number of bad wines is far greater than the number of good wines, which clearly indicates that any classification model built will be better at predicting bad wines than at predicting good wines. Figure 7 presents the correlation plot of each variable against every other variable in the dataset. Notice that this correlation plot for the classification approach is just an extension of the correlation plot for the regression approach. Again, alcohol has the highest positive correlation with good_tag and volatile acidity has the highest negative correlation with good_tag, indicating that these variables could be potentially very important in the classification models to be built. Figure 8 shows the distribution of each independent variable in the classification problem split by good wine samples and bad wine samples. The distributions for the good wine samples are represented in cyan and those for the bad wine samples are represented in pink. Based on these distributions, it is easy predict in advance that volatility acidity, density, alcohol, and sulphate levels will be important variables in the classification problem. This is because the distributions and means of these variables for good wine samples and bad wine samples are relatively more separable.

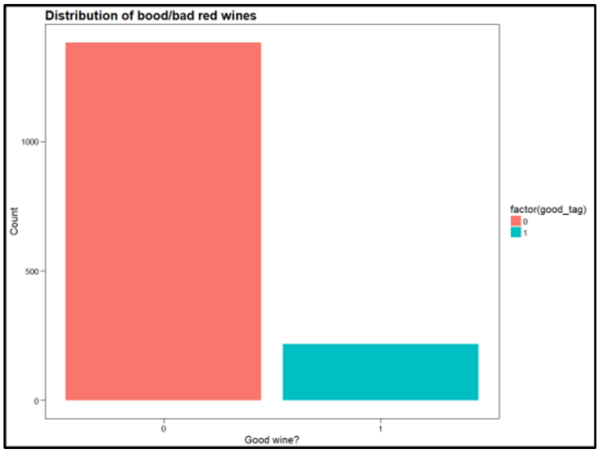

Fig. 6: Distribution of good and bad wine samples

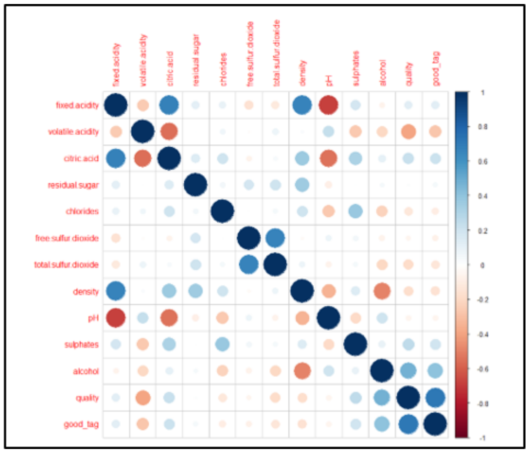

Fig. 7: Correlation plot of variables

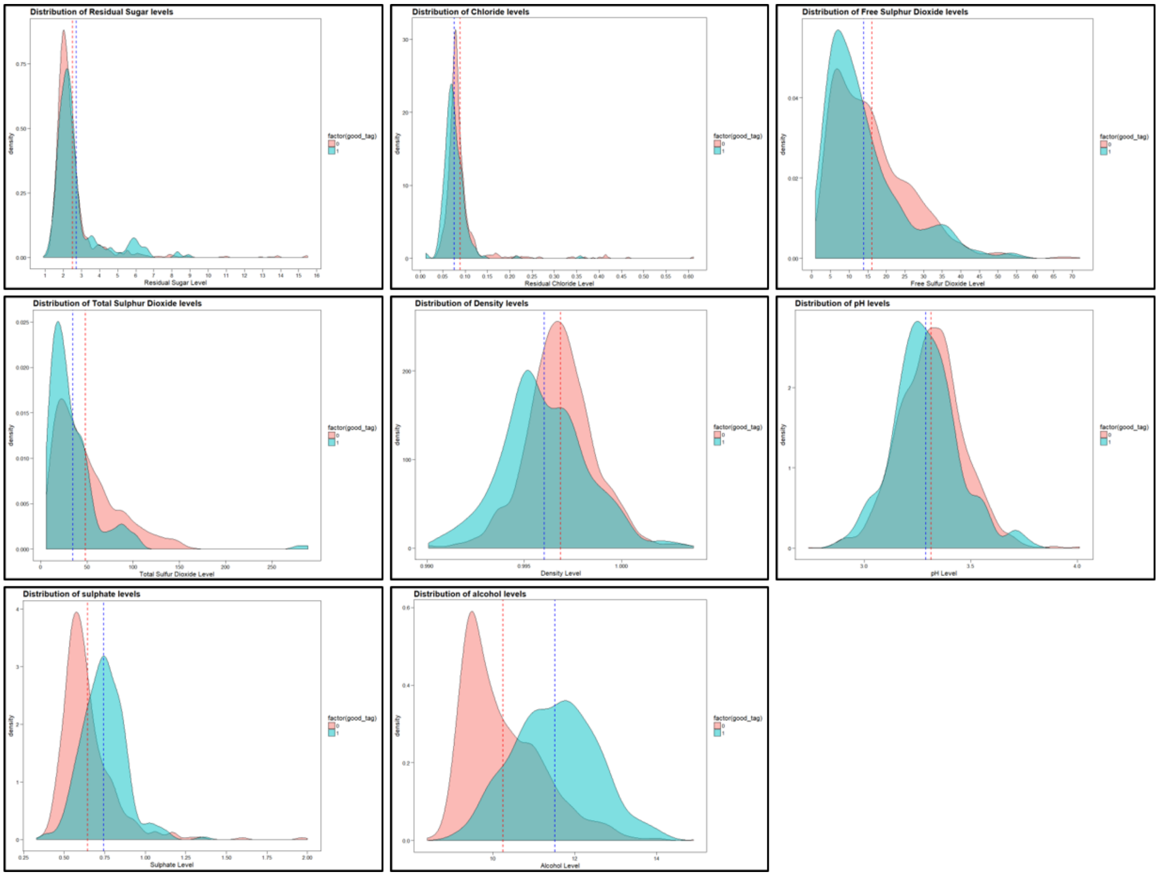

Fig. 8: Distribution of independent variables split by good/bad wine samples

Logistic Regression

The correlation plot and the distributions of variables, split by good and bad wine samples, clearly indicate that alcohol will be the most important variable in the classification model. Knowing this, as a first step towards building a classification model, a simple logistic regression model was built against the response variable, good_tag, with just alcohol as the independent variable. The ROC curve of this model in Figure 8 has an area under the curve of 82.2% with an error rate of 2.6%, clearly indicating that the classification approach is paying dividends. Next, a logistic regression model using all the independent variables was built after employing the stepwise procedure. This yielded better results as the area under the ROC curve improved to 88.2% with an error rate of 2.1% (Figure 9). As seen from the correlation plots above and the individual distributions, the highly significant variables in this model were alcohol, sulphates, volatile acidity, fixed acidity, residual sugar, total sulfur dioxide, density, and chlorides.

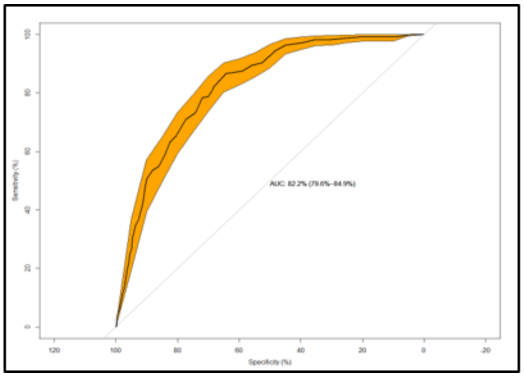

Fig. 8: ROC Curve of log model with alcohol

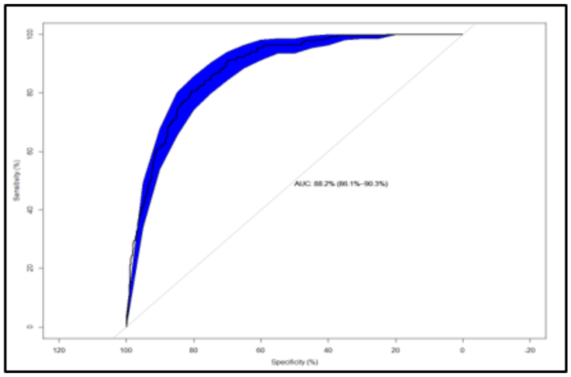

Fig. 9: ROC Curve of stepwise log model

K-nearest neighbors classification

Next, KNN classification models, with 20 and 10 neighbors respectively, were explored to determine if these models could do a better job than the logistic regression model in terms of prediction power. Table 3 and Table 4 represent the confusion matrices based on the KNN-20 and KNN-10 models. With classification error rates of just 13% and 12.44% respectively, the KNN-20 and KNN-10 models showcase better predictive power than the prior logistic regression model. However, the ability of these KNN models to accurately predict good wines is still pretty low. Besides these models are pretty hard to interpret in comparison to logistic regression models.

KNN-20	Predicted Bad	Predicted Bad
Actual Bad	1378	4
Actual Good	204	13

Table 3: Confusion matrix of KNN-20

KNN-10	Predicted Bad	Predicted Bad
Actual Bad	1368	14
Actual Good	185	32

Table 4: Confusion matrix of KNN-10

Decision tree

The need for a better algorithm to improve the prediction accuracy of good wine samples prompted the use of a decision tree classifier. The decision tree model had a superior classification error rate of just 7.817% and does a much better job at predicting good wines (Table 5). The decision tree also provides good interpretation of results in terms of the boundaries of various physicochemical constituents that separate the good wine samples from the bad ones (Figure 10). These insights could directly be utilized by wine makers in developing new samples that could be deemed to be of superior quality.

Decision tree	Predicted Bad	Predicted Bad
Actual Bad	1348	34
Actual Good	91	126

Table 5: Confusion matrix of Decision Tree

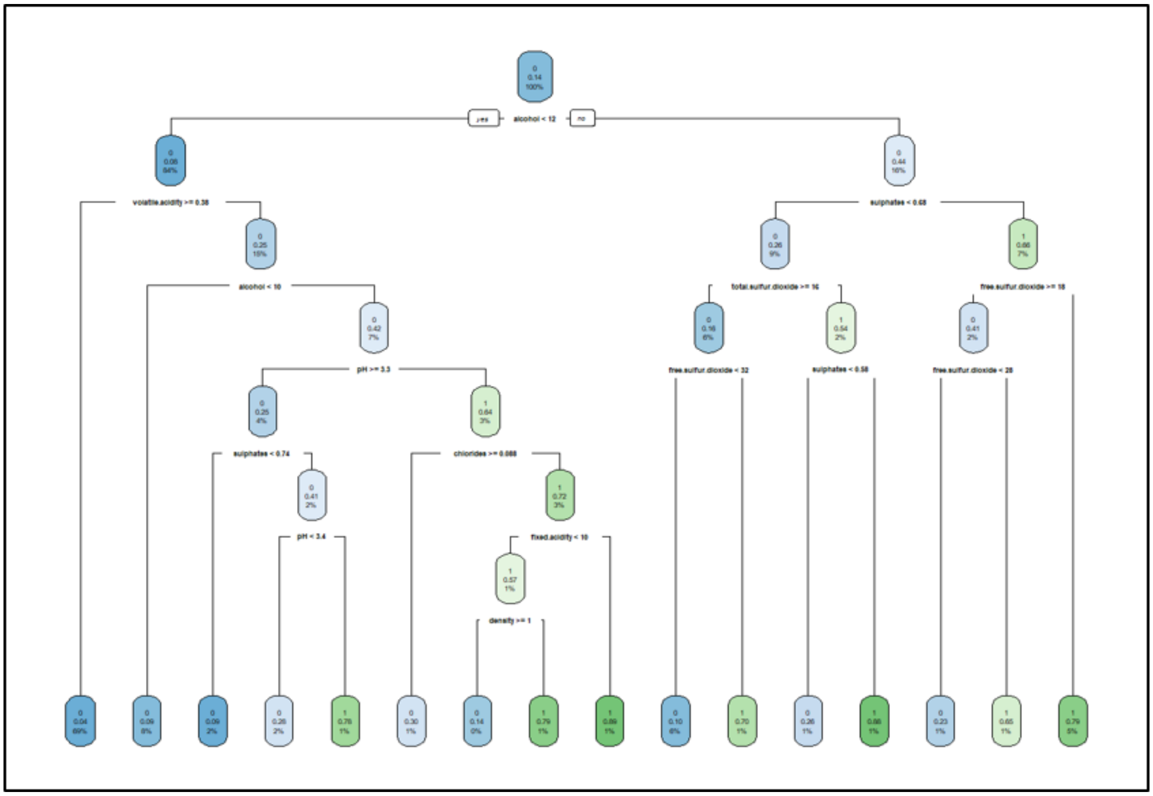

Figure 10: Decision Tree

Random Forest

Although the decision tree does an excellent job in terms of classification, it tends to over-fit the data, which reduces its predictive power when extrapolated to new samples. As a final approach, a random forest model, an ensemble model which takes the average result of one thousand decision trees built from subsets of the dataset, was employed. Despite having a higher classification error than the decision tree model, the random forest model does a better job at not over-fitting the training dataset and hence has better predictive power. This model is also more appropriate for building a platform around and scaling for commercial purposes, despite being more of a black-box model. Table 6 presents the confusion matrix of the random forest model.

Random Forest	Predicted Bad	Predicted Bad
Actual Bad	1351	31
Actual Good	97	120

Table 6: Confusion matrix of Random Forest

Results

The dependence of the quality ratings of red wine samples on their physicochemical constitution was validated using several data mining approaches. The research also yielded a random forest model which could be utilized for commercial purposes to aid wine makers in improving the quality of their produce. The decision tree model showcases some important properties of red wine samples and how they determine their associated quality ratings. The physicochemical constituents of red wine were ranked in Figure 11 in terms of their importance in influencing its quality rating.

The key findings from this study on how the physicochemical composition impacts the quality ratings of red wine samples are summarized below:

- Red wine samples with alcohol below 12% are rated poorly 92% of the time.
- However, samples with alcohol between 10% and 12%, which have volatile acidity less than 0.38 grams of acetic acid per dm^3 and a pH value less than 3.3 have a 36% chance of being classified as good wines.
- Also, samples with alcohol between 10% and 12%, which have volatile acidity less than 0.38 grams of acetic acid per dm^3 and a pH value greater than 3.4, with sulphate levels greater than 0.74 grams of potassium sulphate per dm^3 have a 22% chance of being classified as good wines.
- With regard to samples with alcohol greater than 12%, those with sulphates greater than 0.68 grams of potassium sulphate per dm^3 have a 34% chance of being classified as good wines.
- Finally, samples with alcohol greater than 12% and sulphates less than 0.68 grams of potassium sulphate per dm^3, which have greater than 16 milligrams of total sulfur dioxide per dm^3 but less than 32 grams of which is free sulfur dioxide, have a 30% chance of being classified as good wines.

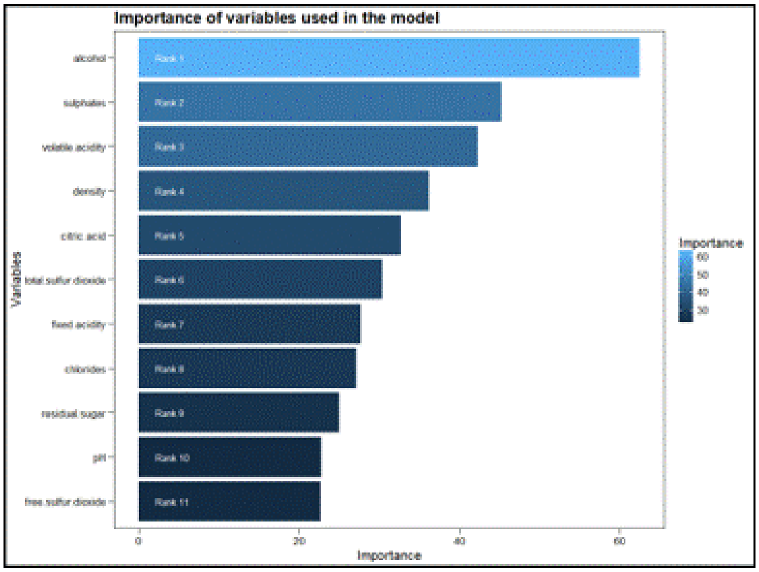

Fig. 11: Importance of physicochemical characteristics in determining red wine quality ratings

Discussion

The surge in the demand for wine over the past decade or so has seen an unprecedented growth in the wine industry**.** As wine consumption has become more mainstream, the demand for superior quality produce has driven wine makers to turn to data-driven insights to alter their production methods. Research in this area has mostly utilized datasets with a sample size. Using a relatively big dataset of 1599 samples from a 2009 study, showcasing the quality ratings of various vinho verde red wine samples and their associated physicochemical characteristics, various data-driven models were explored by viewing the problem first as a regression problem. The failure of regression models to adequately explain the variation in ratings and generate insights prompted viewing the problem as a classification problem. Both these approaches, however, were able to validate the notion that the physicochemical composition of wine samples affect their quality rating.

This study differs from prior research in this area by considering the problem as a classification problem, rather than a regression problem. This facilitated building a random forest model that could potential be scaled for commercial purposes to help facilitate wine makers in improving the quality of their produce. Individual constituent level boundary conditions with confidence limits were also determined to help producers quantify the risk associated with developing new samples.

Despite these contributions, the models developed here suffer from a serious case of over-fitting. Prioritizing generation of insights over building a predictive model for commercialization, the entire dataset was utilized for both training and validation purposes, except in case of the random forest model which has been proposed for commercial purposes. The random forest model could be further optimized through hyper-parameter tuning to improve its predictive power. Post commercialization, this model would need to be adjusted on a timely basis as perceptions of good and bad wine samples change and more data becomes available. Further study in this area could explore other approaches such as Neural Networks to build models that good do a better job at predicting good wines.

In conclusion, this study further validated the dependence of quality ratings of wine on the physicochemical composition of samples, generated tailor-made insights for wine makers to experiment and develop new samples, and helped develop a random forest model with substantially high predictive power to be scaled for commercial purposes to facilitate wine makers in knowing the market perception of their planned wine production schedules in advance.

References

P. Cortez, A. Cerdeira, F. Almeida, T. Matos, and J. Reis (2009). Modeling wine preferences by data mining from physicochemical properties, *Elsevier*, *47*, 547-553.

S. Ebeler (1999). Linking flavour chemistry to sensory analysis of wine, *Flavor Chemistry - Thirty Years of Progress,* Kluwer Academic Publishers, 409-422.

E. Turban, R. Sharda, J. Aronson, and D. King (2007). *Business Intelligence, A Managerial Approach*, Prentice-Hall.

I.H. Witten and E. Frank (2005). *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*, Morgan Kaufmann, 2nd edition

J. Ferrer, A. MacCawley, S. Maturana, S. Toloza, and J. Vera (2008). An optimization approach for scheduling wine grape harvest operations, *International Journal of Production Economics, 112(2):985–999, 2008.*

A. Asuncion and D. Newman (2007). UCI Machine Learning Repository, University of California, Irvine, http://www.ics.uci.edu/~mlearn/MLRepository.html, 2007.

L. Sun, K. Danzer, and G. Thiel (1997). Classification of wine samples by means of artificial neural networks and discrimination analytical methods. *Fresenius' Journal of Analytical Chemistry,* 359(2):143–149.

S. Vlassides, J. Ferrier, and D. Block (2001). Using Historical Data for Bioprocess Optimization: Modeling Wine Characteristics Using Artificial Neural Networks and Archived Process Information. *Biotechnology and Bioengineering*, 73(1).

A. Legin, A. Rudnitskaya, L. Luvova, Y. Vlasov, C. Natale, and A. D'Amico (2003). Evaluation of Italian wine by the electronic tongue: recognition, quantitative analysis and correlation with human sensory perception. *Analytica Chimica Acta*, 484, 33-34.

I. Moreno, D. Gonz´alez-Weller, V. Gutierrez, M. Marino, A. Came´an, A. Gonz´alez, and A. Hardisson (2007). 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*, 72(1):263–268.

P. Cortez, M. Portelinha, S. Rodrigues, V. Cadavez, and A. Teixeira (2006). Lamb Meat Quality Assessment by Support Vector Machines. Neural Processing Letters, 24(1):41–51.

H. Yu, H. Lin, H. Xu, Y. Ying, B. Li, and X. Pan (2008). Prediction of Enological Parameters and Discrimination of Rice Wine Age Using Least-Squares Support Vector Machines and Near Infrared Spectroscopy. *Agricultural and Food Chemistry*, 56(2):307–313.

CVRVV (2008). Portuguese Wine - Vinho Verde. Comiss˜ao de Viticultura da Regi˜ao dos Vinhos Verdes (CVRVV), http://www.vinhoverde.pt.