A machine learning approach using Random Forest and LASSO to predict wine quality

Ioannis Athanasiadis* and

Dimitrios Ioannides

University of Macedonia,
Egnatia str.156, Thessaloniki, Greece
Email: athang@uom.edu.gr
Email: dimioan@uom.edu.gr
*Corresponding author

Abstract: Quality assessment is a key factor for the wine industry, where the aim is to meet consumers' needs/demands and promote sales. Quality assessment is usually performed by experts and it is a time-consuming and expensive process. This paper proposes an alternative assessment using machine learning methods, such as the Least Absolute Shrinkage and Selection Operator (LASSO) and Random Forest to predict wine quality. Our data analysis is based on a real wine dataset provided by a well-known wine firm in Greece. For this purpose, we employ the LASSO method, which is particularly effective in selecting the best possible number of variables required. Additionally, the Random Forest method is used and its findings are contrasted to those derived by four (4) different M.L. methods, namely, Linear Discriminant Analysis (LDA), Classification and Regression Trees (CART), k-Nearest Neighbors (kNN) and Support Vector Machines (SVM), and using the well-known 10-fold cross-validation method. The results of our analysis show that the statistical technique of Random Forest proposed improves the accuracy of the prediction wine quality, up to almost 95%, compared to the rankings attributed by wine tasters.

Keywords: Random Forest, LASSO, machine learning, physicochemical properties, wine quality, prediction

Biographical notes: Ioannis Athanasiadis is a PhD student . He studied Mathematics and received his Master’s in Information Systems at University of Macedonia, Thessaloniki, Greece. He has several presentations in International Conferences. He currently works as Teaching Laboratory Staff in the Department of Economics of the University of Macedonia.

Dimitrios Ioannides is a Professor in Department of Economics, University of Macedonia, Thessaloniki, Greece. He worked also as Professor at the University of Patras. He was a visiting Professor at the University of Cyprus, Ren(France), Tomsk(Russia) and Davis, California (USA). He has several number of research papers published in reputed international and national journals, related to Statistics and Time series. He is author of three books in statistics on Greek and translated one in probability theory from English. Many of his works have been presented in International and National Conferences.

1 Introduction

The first efforts to classify and predict wine quality were mainly based on the weather of the year of harvesting, on ageing and quantity, among others. For example, Ashenfelter (2008) attempted to predict the quality and prices of Bordeaux red wines observing the weather that created the grapes and some other variables, such as historical reputation or quantity produced. It was, in effect, a kind of primitive predicting method depending on the weather and ageing processes rather than on statistical analysis and results derived from regression models, the aid of which makes it possible to predict both prices and quality.

On the other side of the Atlantic, in Chile, Beltran et al.(2008), based on information contained in wine aroma/bouquet chromatograms, measured using a commercial fast GC analyzer. Using principal component analysis (PCA) and Discrete Wavelet Transform (DWT) as feature selection techniques, they used Linear Discriminant Analysis (LDA) as well novel data mining algorithms methods, such as Radial Basis Function Neural Networks (RBFNN) and Support Vector Machines (SVM) for wine classification. Their sample was 100 wine labels of three different domestic wine types. For each wine, they made 10 wavelet observations and divided them into a 90% training set and a 10% testing set. The authors concluded that the best results were obtained using wavelet decomposition, as a feature extraction method, and SVM with a radial base function (RBF) type of kernel, as a classification technique. The weak aspects of their modeling were the small sample numbers, the wide variation percentages in the testing phase (between 37%-90%) and the fact that the authors depended on a commercial wavelet chromatography machine for their observations.

In a similar fashion but using a big dataset of wines (4898 white and 1599 red), Cortez et al. (2009) introduced three powerful and interesting classification methods, namely Multiple Regression (MR), Neural Networks (NN) and Support Vector Machines (SVM); these helped them make informed prediction on a wine label quality score based on its physicochemical properties. The authors concluded that the superiority of SVM over NN is probably due to differences in the training phase. The SVM algorithm guarantees an optimum fit, while NN training may fall into a local minimum. There is no doubt that the study by Cortez et al. (2009) was one of the most influential ones because of the big sample dataset of two wine types (white and red); however, there were a few open issues, such as tolerance and its influence on results, a single prediction metric (mean absolute deviation MAD), and absence of more layers in neural networks.

Abbal et al. (2019) studied the effect on wine quality of several variables including quality of soil, type of plants, meteorological conditions, various agronomic variables, vineyard parameters, grape characteristics, and enological parameters. Furthermore, the authors constructed a model validated for predicting quality scores and these scores were compared against the scores given by International Press for 49 modeled wine labels. This was an important study using many variables; however, its limitations had to do with the small sample and the low number of prediction metrics, such as correlation and marginal error.

The currently emerging use of “electronic datasets” increases with the big web market data and, consequently, affects any company that wants to remain a player in the field and seeking methods to achieve useful information from such big datasets. Athanasiadis and Ioannides (2014) attempted to make predict the quality score of the wine labels using multiple linear regression and logistic regression; logistic regression proved to have much better results. The analysis could be improved with the use of more metrics and, above all, with more data available. In recent years, we have been witnessing not only significant progress in data analysis but also the creation of many important tools that become available even to small firms. These tools of statistical analysis can be used with the safety required and the results derived could be used profitably by firms. Moreover, the wine quality achieved through this emerging technology can be ranked on objective criteria and, therefore, easily certified. An additional advantage of our data analysis and wine qualification and certification is the prevention of illegal adulteration of wine, ensuring a fair quality index in the wine market (Cortez, 2009).

Physicochemical and sensory tests are prerequisites for such certified wine. The former can be achieved by using certain chemical property indices and the latter can be achieved through human tasting. Taste is the least understood of human sensory perceptions and the relationships between physicochemical results and sensory analysis are difficult to recognize and still not fully understood. Legin, et al., (2003) and Smith & Margolskee, (2006) tried to introduce an ‘electronic tongue’, classify results with principal components analysis (PCA), and predict with Neural Networks using commercial software (NeuroSolutions® & Unscrambler®). Their sample was small (56 wine labels) and reported one prediction metric, namely, mean percentage error.

Against this background, some years later, Gustafson et al.(2016) tried to combine consumers’ willingness to pay (WTP) with hedonic evaluation of wine, produced from twelve different regions of the USA and seven varieties, and comparing with the price of six alternatives wine labels. The sample consisted of about 280 wine labels and 250 consumers. The authors almost proved that wine appellation has little to do with consumers’ willingness to pay but other wine quality attributes seem to be significant. Furthermore, in their work regarding the New Jersey area of viniculture and wineries, Moscovici et al.(2017) identified, significant sustainability characteristics, such as land/territory protection, water, energy, food miles and wildlife impacting the vineyard environment. When Vlontzos et al.(2017) attempted to classify wine, they pointed out that there are two main classes according to EU legislation, i.e., Protected Designation of Origin (PDO) and Protected Geographical Indication (PGI) labels. Regarding the classification of aged wines, they report aroma/bouquet and taste as the most important wine quality properties.

This study aspires to predict a wine quality score based on physicochemical results and, in this sense, it might be of great help to wine producers in deciding on their wine prices and promotion. The LASSO method of dimensionality reduction is used to find whether there is significant elimination of variables so as to simplify prediction of wine quality score to guide business decisions accordingly, offering producers a competitive advantage.

The concept and the classification of the wine quality is a multi-faceted process. Characteristics considered are soil (composition, gradient, and orientation), micro-climatic conditions of the vineyard region (sunshine, rainfall, temperature range, humidity wind, etc.), adaptation and expression of each variety in the region, methods of wine making, grape ripening, ageing, etc. In this particular classification, the typicality of wine is dominant. The following are examined: absence of defects, flavor and aromatic balance of each wine label, formality in the expression of the variety (for mono-variety wines) or the region in the case of PDO (Protected Designation of Origin) or PGI (Protected Geographical Indication) wine labels, as well as physicochemical (analytical) and color features. The classification process described above allows the analyst-taster to overcome, to a significant extent, the problem of distinguishing wines from different harvests; this follows consideration of all observations, from the stage of receiving the grapes to that of the final product and concerns the total number of wine samples of each harvest. Tasters, finally, adjust classification based on mono-variety or blend variables, to reach a final result.

The quality index emerges as a mean value of various important criteria. The most important one is the absence of defect, which classifies the wine at a low score without considering other factors. Subsequently, the tasting test of standardization counts for about 80%, while chemical characteristics count for 20%.These physicochemical characteristics of each wine are extracted from the winery laboratory. It was noted that the final quality score of a wine label of similar physicochemical properties, but different year of vinification/harvest, was only one point higher or lower except for years with catastrophic weather conditions.

In our study we predict white wine quality based on physicochemical data from a winery in Northern Greece. The central role for this study is played by wine typicality (formality). In general, the term formality in wine must indicate the origin of the grape variety, the category of the wine (dry, semi-dry, sweet, semi-sweet), its geographical origin, and the year of its production. Formality can raise or lower wine quality, simply because a wine critic, taster or judge thinks that a wine is not typical and gives a low rating.

In other words, formality can give information about the place of wine production, the time of grape harvest, and the wine category and way of vinification. Other variables examined are the absence of defects, the taste and aromatic balance of the wine, its formality in terms of expressing the variety (for single-variety wines) or the region for PDO (protected designation of origin) or PGI (protected geographical indication) wines. Of course, the physicochemical measurements of wine characteristics are provided. The protocol of this vinification is described in a Master's Thesis by Vlachou (2011),p.57.

In this controlled environment, expert “tasters or judges” can give a more objective evaluation of wine quality. Some additional details about our data and quality evaluation will be presented in the next section.

The focus of our study is on a white wine set, but the same procedure could easily be applied for different wine varieties. A second goal is to find the best prediction using the Random Forest technique. The purpose is to achieve a faster and simpler way of selecting variables and increasing the predictive capacity of the model selected in this manner. As Fathi et al.(2019) pointed out in their paper correlating weather and crop yield data, there are very promising data mining algorithms to use for Decision Tree and Random Forest predictions, which, in their case, was about the crop yield of olive trees.

The structure of the paper is organized as follows: Section 2 presents the wine data, the Preliminary Graphic Analysis, and the variable selection approach using the LASSO technique. Section 3 introduces the Random Forest prediction method which helps analyze the dataset and assign a quality score using 10-fold cross-validation (CV). Moreover, by dividing the set into training (70%) and testing (30%), prediction metrics are made available for the testing set. The section also includes all critical results and relevant tables and graphic figures. Section 4 summarizes and concludes by comparing the results of the 5 Machine Learning methods.

2. Data Description and Variable Selection with LASSO

2.1 Data Description

This study considers 2312 labels of white wine produced by a well-known winery of Northern Greece. Most of these wine labels are exported, mainly to European countries. The data were collected from April/2004 to October/2013 (almost 9 years) and tested so as to issue an official certification by the company quality department. Each entry denotes a given test (analytical and sensory) and the final database was exported into a single sheet (.csv). Data description is presented in Table 1 (in brackets the abbreviated name used in the dataset).

Figure 1. Histogram of Wine Quality

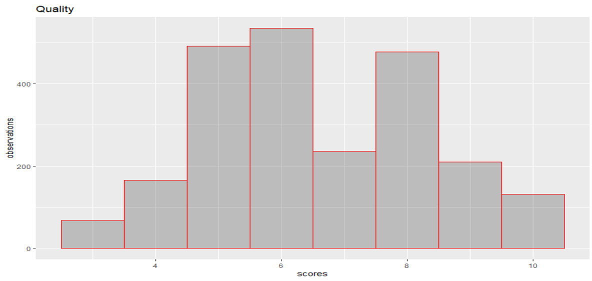


Table 1. Physicochemical properties data

	White wine		
Attribute (units)	min	max	mean
Quality [qual]	3	10	6.568
alcohol (% vol.) [alc]	9	13.65	11.71
PH [ph]	2.7	3.98	3.38
total acidity (g(tartaric acid)/dm^3) [ta]	3.37	8.93	5.01
volatile acidity (g(acetic acid)/dm^3) [va]	0.100	1.430	0.306
sugar (g/dm^3) [sug]	0.50	36	5.14
color intensity [col]	0.030	0.93	0.077
free sulfur dioxide (mg/dm^3) [fsd]	6	65	38.6
total sulfur dioxide (mg/dm^3) [tsd]	26	248	138.7

Figure 2. Chart of correlations and p-values

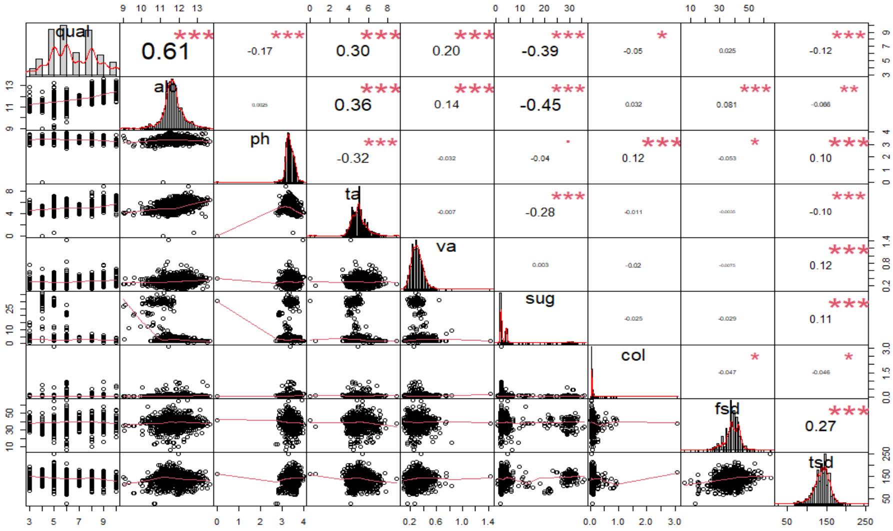


Table 2. Wine labels per Year

Year	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013
Wine labels	20	285	347	391	278	345	264	111	153	118

Table 3. Annual Mean Value per Variable

Annual Mean Value	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013
quality	7.2	7.15	6.87	6.72	6.43	6.25	6.48	6.31	6	6.04
alcohol (% vol.	11.79	11.9	11.88	11.73	11.72	11.55	11.56	11.65	11.7	11.47
pH	3.51	3.48	3.39	3.44	3.44	3.35	3.32	3.27	3.25	3.32
total acidity (g(tartaric acid)/dm^3)	4.66	4.95	4.98	4.99	4.8	5.14	5.24	5.22	5.15	4.72
volatile acidity (g(acetic acid)/dm^3)	0.43	0.32	0.28	0.27	0.34	0.35	0.32	0.26	0.28	0.32
sugar (g/dm^3)	7.55	5.22	4.99	5.82	5.55	4.64	4.62	4.11	4.54	6.07
color intensity	0.06	0.07	0.08	0.08	0.08	0.07	0.07	0.08	0.08	0.11
free sulfur dioxide (mg/dm^3)	38.6	39.11	39.12	37.89	39.29	39.02	36.92	39.53	39.02	37.61
total sulfur dioxide (mg/dm^3)	156.6	142.29	132.2	136.3	142.32	143.61	142.32	138.83	133.12	129.58

2.2 The *k*-fold CV

The *k*-fold CV protocol is based on splitting data into *k* mutually exclusive groups, termed 'folds'. The ordinary value of *k*, is *k* = 10, yielding the 10-fold CV. One of the folds is selected and put aside to play the role of the test set. The remaining 9 folds are combined into what is called the training set. Only the predicting variables of the training set were standardized. Then variable selection was performed using LASSO to feed the RF algorithm with the variables selected. The same predicting variables selected are used for the test set and scaled using the means and standard deviations of the same predicting variables of the training set. Then these scaled predicting variables of the test set are used to predict the values of the response variable (percentages of accuracy) of the test set.

Subsequently, another fold is selected to play the role of the test set and the remaining become the new training set and so on, till all folds have been tested. The process is repeated until all folds have played the role of the test set. In the end, all predictions are collected from each fold resulting in an $n \times M$ matrix, where n is the sample size and M the number of hyper-parameters, the total number of splits of variables, corresponding to M sets of predictions. The average predictive performance was computed for each hyper-parameter separately and the hyper-parameter with the highest predictive performance was chosen.

2.3 Internal evaluation in the dataset

What is illustrated now is the internal evaluation of RF in the datasets used exceeded our expectations. The 10-fold CV procedure (described in the next section) is implemented to tune the penalty parameter (λ) of LASSO. Then the LASSO penalization is performed using the chosen λ value to select the most important variables.

Furthermore, the predictive performance of RF is evaluated by contrasting each set of predictions (one set for each hyper-parameter) against true quality values using the Pearson Correlation Coefficient (PCC) (1) or *r,* and the Percentage of Variance Explained (PVE) or R^2 (2).

$$PCC = \frac{\sum_{i=1}^{n}(y_i-\overline{y})(\hat{y}_i-\overline{\hat{y}})}{\sqrt{\sum_{i=1}^{n}(y_i-\overline{y})^2}\sqrt{\sum_{i=1}^{n}(\hat{y}_i-\overline{\hat{y}})^2}} \tag{1}$$

$$PVE = 1 - \frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{\sum_{i=1}^{n}(y_i-\overline{y})^2} \tag{2}$$

Where $\hat{y}$ refers to the predicted value of the i-th observation and $\overline{y}$ denotes the mean value. In linear models the metrics above are the PCC (= *r*) (or correlation) and the PVE (= R^2) (or coefficient of determination).

Furthermore, the mean squared error is calculated as follows:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{3}$$

Of course, the Root Mean Squared Error (RMSE) used in this work is the squared root of the measure already mentioned.

On the other hand, the mean absolute error is calculated as follows:

$$\text{MAE} = \frac{\sum_{i=1}^{n}|\hat{y}_i-y_i|}{n} \tag{4}$$

The metrics presented above can be compared against a benchmark value. Consequently, we use in our evaluation metrics, i.e., the RMSE, and the mean absolute percentage error,

$$\text{MAPE} = \frac{\sum_{i=1}^{n}\left|\frac{y_i-\hat{y}_i}{y_i}\right|}{n} \tag{5}$$

For the rest of this work we will employ only the RMSE and MAPE.

For both PCC and PVE, maximum value is 1, indicating excellent predictive performance, whereas minimum value is equal to 0 reflecting completely random predictions. Higher PCC values indicate a higher number of correct model-based predicted rankings, whereas PVE higher values indicate that, on average, errors of, model-based predictions are fewer than errors of random, model-free predictions.

The PVE metric is used for model assessment and perhaps the only safe conclusion drawn from Table 1is that non-linear models perform better than the linear model of LASSO. RF always produced PVE values above 0.8 (or 80%) indicating an excellent fit. Comparison between these PVE values and the R^2 values reported in previous papers we are delighted, not only because we outperformed their fit, but also because our PVE values are remarkably high. The cost of this high PVE is interpretability. RF does not produce a coefficient for each predictor variable that could reflect the (marginal) effect of the variable on wine data.

Accuracy is a measure of how many positive predictions were achieved, i.e., it is a measure of exactness. A higher value of precision means fewer false positives, while a lower precision value means more false positives. The best precision is 1.0, whereas the worst is 0.0. For example, when we have 6 correct predictions out of 7 observations, then:

Accuracy = TP/ (TP+FP) =6/ (6+1) =0.857.

Thus, the precision value is approximately 85% since there are fewer false-positive values in the dataset.

In this paper we use another accuracy approach named *min_max_accuracy,* as shown below*:*

$$\min\_\max\_accuracy = mean\left[\frac{min\,(y_i,\hat{y}_i)}{max\,(y_i,\hat{y}_i)}\right] \quad (6)$$

(Where $\hat{y}$ are the values predicted)

The need for such metric was the presence of continuous rather than discrete values as predicting ones, resulting in a false accuracy metric.

Let us present an example: If we have predicted a value of 6.8 and the actual value is 7, then 6.8/7 = 0.97, which shows percentage prediction of the approach. This should be impossible with an accurate measure because accuracy in such a case equals zero (0).

We define the following formula as Kappa measure:

Kappa = (observed accuracy – expected accuracy) / (1 – expected accuracy)

2.4 Multiple linear regression

Using the linear multi-regression model and using all methods (backward, forward, and stepwise) for checking all variables only six of the 8 variables remain: *alc, ph, sug, va, tsd* and *col.*

As seen from Table 4below, having Quality as a discrete variable, there are mediocre results not conducive to prediction when using multi-linear regression. We get $R^2 = 0.46$ and r = 0.63.

The regression line is:

```
qual = -8.4112 + 1.7459 * alc - 1.5201 * ph +
3.0058*va - 0.0408 * sug - 0.8147 * col - 0.0071 * tsd
```

With these 6 remaining variables we get the following metrics for the variable of Quality, by dividing the dataset into a 70% training and a 30% testing set:

Table 4. Evaluation metrics with multiple-linear regression

rmse	mape	cor	min_max accuracy
1.3891492	0.1948515	0.6318983	0.8454939

Prompted by these poor results, this paper attempts to explore non-linear machine learning methods, starting with methods like Decision Trees and Random Forest.

2.4.1 Best combinations of interactions in multiple linear regression

As already mentioned, concerning the correlation matrix, it was evident that some properties with strong relationships to each other may have a cumulative effect due to their interaction in the multiple linear regression. Below the best combination among all the others is presented, as having the best results after training in the training set the model (75%) and testing the prediction in the testing set (25%). The best interactions resulted from Alcohol vs. Sugar and Alcohol vs. Volatile Acidity.

Safi et al.(2017), attempting to predict stock prices using methods such as Neural Networks, used three different forecasting criteria as primarily evaluation metrics, namely, Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and Mean Absolute Percentage Error (MAPE). We will attempt to achieve these metrics, mainly using the training and testing statistical test, by dividing dataset observations into 75% for the training set and 25% for the testing set.

Multiple linear regression with interactions in the training set:

```
Call:
lm(formula = qual ~ ph +col +alc * sug +alc * va,
data=mydataset_train)

Residuals:
Min      1Q  Median      3Q     Max
-4.8646 -0.7137 -0.0126  0.8736  4.3055

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)  -4.899643   2.757518  -1.777 0.075772 .
ph           -1.648693   0.180390  -9.140  < 2e-16 ***
col          -2.008732   0.521272  -3.854 0.000121 ***
alc           1.438474   0.223309   6.442 1.53e-10 ***
sug           0.886569   0.098185   9.030  < 2e-16 ***
va          -27.141167   7.606742  -3.568 0.000369 ***
alc:sug      -0.083401   0.008843  -9.431  < 2e-16 ***
alc:va        2.500986   0.644772   3.879 0.000109 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.253 on 1727 degrees of freedom
Multiple R-squared:  0.4935,  Adjusted R-squared:  0.4915
F-statistic: 240.4 on 7 and 1727 DF,  p-value: < 2.2e-16
```

After testing prediction:

rmse	mape
1.2797	0.1716

The predicted equation is:

```
y = -4.89 + 1.43*alc - 1.64*ph -2.00*col
+0.88*sug -27.14*va -0.08*alc*sug +2.50*alc*va
```

2.5 Variable selection with Least Absolute Shrinkage and Selection Operator (LASSO)

Least Absolute Shrinkage and Selection Operator (LASSO) (Tibshirani, 1996) is a regression analysis method that performs variable selection and regularization of regression coefficients. The objective of this method is to improve prediction accuracy and interpretability of regression models by selecting a subset of the variables provided that exhibits the strongest effects on the response variable. LASSO can improve prediction error by shrinking large regression coefficients to reduce overfitting, while it can also perform variable selection, discarding variables responsible for large variance to make the model more easily interpretable.

LASSO minimizes the following penalized sum of squares:

$$\sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{n}(\beta_j x_{ij})\right)^2 + \lambda \sum_{j=1}^{p}|\beta_j| \qquad (7)$$

Where y_i is the *i– th* response value, x_{ij} denotes the *i– th* value of the *j – th* predictor variable, *n* denotes the sample size and *p* is the number of predictor variables.

Fine-tuning the penalty parameter λ is essential for the performance of LASSO since it determines the scale of regularization, the intensity of shrinkage and, ultimately, the number of variables selected for use in the final model. This is achieved through a cross-validation procedure, where the λ value yielding the lowest estimated prediction error is preferred.

Correlations between the wine dataset and performance measures showed statistically significant differences; however, not all correlations remain significant when all predictor variables are included into a regression model. The LASSO algorithm facilitates detection of the most important performance measures.

2.6 Random Forests algorithm

The RF algorithm is a fast and flexible data mining approach, well-suited for high-dimensional data. The algorithm is built on creating many classification or regression trees. According to Breiman(2001), RF randomly draws a subset of variables and a bootstrap sample and uses only this subset of features to grow a single tree. This process of randomly selecting variables and bootstrap samples is repeated numerous times and results are aggregated. By creating many trees at random (500 or 1000, for instance), one ends up with a random forest.

As stated in the introduction, the relationship between wine data and their performance concerning quality is not expected to be linear; hence, the RF algorithm will allow us to capture the non-linear components of this relationship.

3. Applying results from LASSO and Ensemble methods

3.1 LASSO approach

As seen in Tables 5,6,7 below, always using min lambda, there is no significant reduction of features (only a single “*fsd*”); consequently, the prediction metrics are not high, leading to almost the same results as the multiple linear regression seen in Table 2.

All LASSO models were made using the CV (cross-validation method).

The regression line according to LASSO is:

```
qual = -8.6260 + 1.7171 * alc - 1.4014 * ph + 0.0057 *
ta + 2.6851 * va - 0.0378 * sug - 0.5378 * col -
0.0057 * tsd
```

Table 5. Lambda values

min lambda	0.028714	(log=-3.55035)
best lambda	0.115921	(log=-2.15484)

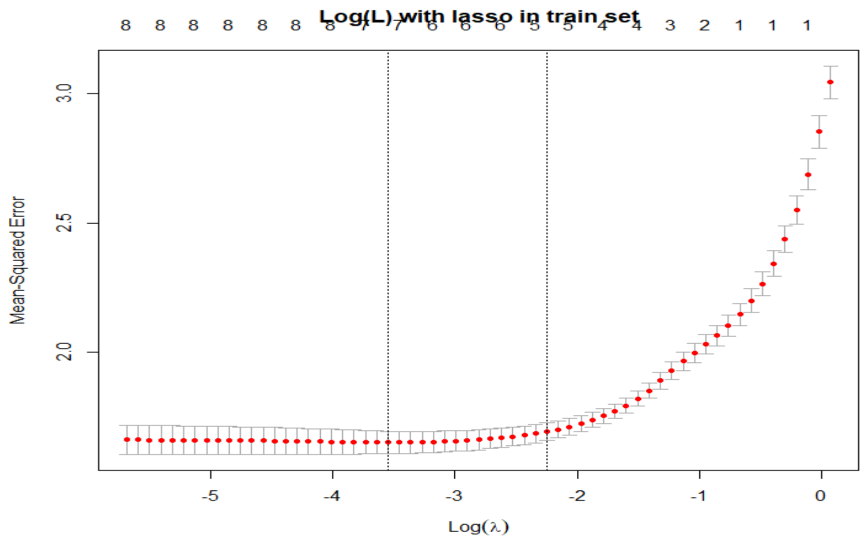
Table 6. Coefficients in the training set with min lambda

(Intercept)	alc	ph	ta	va	sug	col	fsd	tsd
-8.6259	1.7170	-1.4013	0.0057	2.6851	-0.0378	-16.846	0.0000	0.0057

Table 7. Evaluation metrics of LASSO

rmse	mape	Accuracy	Cor	min_max accuracy
1.385907	0.176776	0.316546	0.632268	0.845225

Figure 3. Lambda values for LASSO



3.2 Tree methods

3.2.1 Decision Trees

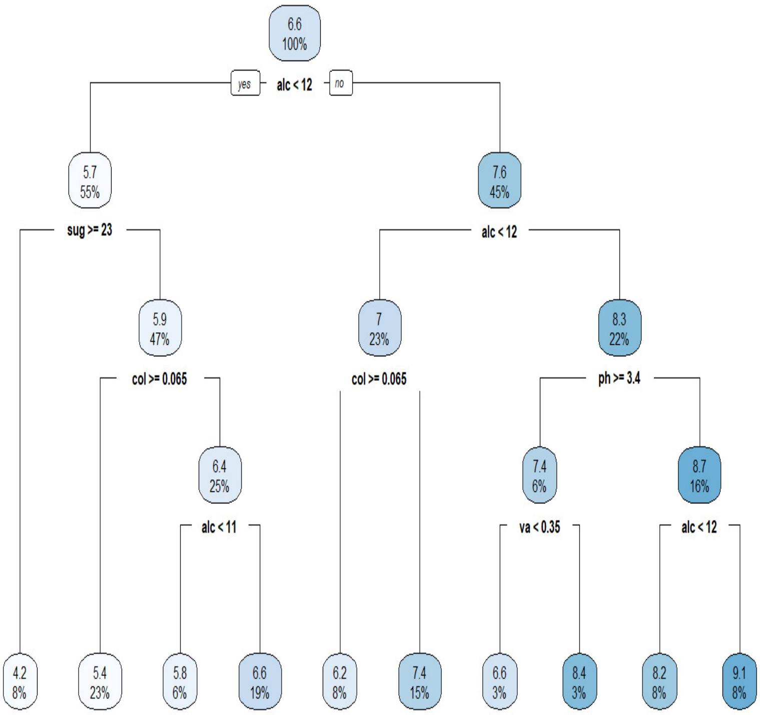
Attwal & Dhiman(2020) in their work for the crop yield, proved that Decision Trees have very promising results in the prediction processing.

As seen from Table 6 below, the method of Decision Trees yields a moderate MAPE, with low Accuracy. Moreover, viewing Figure 4, the final percentage varies from 3% to 23% max, showing no path with a higher percentage. In conclusion, there are no highly acceptable values.

Table 8. Prediction metrics of Decision Trees

rmse	mape	Accuracy	Cor	min_max accuracy
1.300802	0.159868	0.348201	0.693270	0.870657

Figure 4. Plot of decision trees



3.2.2 Random Forest

In Table 9 below, the "randomForest" package of R is used with a number of 500 (ntree) trees and mtry = 2, i.e.,'2' (the 'mtry') is the number of variables randomly sampled as candidates at each split or node. The method was applied to the training set and then the prediction method was applied to the testing set always using the CV tuning.

Table 9. Prediction metrics results with RF without 10-fold CV tuning

rmse	mape	Accuracy	cor	min_max accuracy
0.992779	0.087584	0.588489	0.832497	0.920161

In all metrics below, in Table 10, the “rf” method of the “caret” package of R with 10-fold cross-validation is used without tuning the grid to the training set.

Table 10. Results of RF, without tuning, using the “caret” package

mtry	Accuracy	Kappa
2	0.7776548	0.7297892
5	0.7678084	0.7180348
8	0.7690161	0.7690161

The best final value used for the model was mtry = 2.

In the metrics below, in Table 11, the “rf” method of the “caret” package of R is used with 10-fold cross-validation, with ntree=300, node size=14 and tuning the grid.

Table 11. Results of RF with tuning and ntree=300, node size=14

mtry	Accuracy	Kappa
5	0.7426821	0.6864764
6	0.7482077	0.6932486
7	0.7370839	0.6795875

The best final value used for the model was mtry = 6.

As seen from tables 10 and 11 above, the first model presents better Accuracy and Kappa measure. In Table 12 below, prediction metrics with RF(2) emerge from the first choice of RF. Results in prediction metrics with the second choice of RF presented the worst results in every measurement.

Table 12. All the results from Dec.Trees and RF methods

	Decision Trees	Random Forest (1)(without 10-fold CV)	Random Forest (2) (with 10-fold CV)
MAPE	0.159868	0.087584	0.064988
Accuracy	0.348201	0.588489	0.797122
Correlation	0.693270	0.832497	0.828797
min_max_accuracy	0.870657	0.920161	0.946254

4. Comparing Five ML models using 10-fold CV

Examining various Machine Learning methods, using the same dataset and the same method of 10-fold cross-validation allows us to choose which one could give better results between:

- Linear Discriminant Analysis (LDA)
- Classification and Regression Trees (CART).
- k-Nearest Neighbors (kNN).
- Support Vector Machines (SVM) with a linear kernel.
- Random Forest (RF)

Number of resamples: 10

As seen from the tables below, the Random Forest method achieves the best result out of all other ML methods.

Table 13. Accuracy values for every ML method

	Min	1st q.	Median	Mean	3rd q.	Max	NA's
LDA	0.42944	0.48844	0.51844	0.50950	0.52795	0.58490	0
CART	0.34969	0.37858	0.40431	0.40131	0.42097	0.45061	0
k-NN	0.39506	0.40941	0.44756	0.44192	0.46869	0.48447	0
SVM	0.56790	0.58779	0.61226	0.61986	0.64951	0.69565	0
RF	0.69938	0.76637	0.78703	0.78462	0.80797	0.85093	0

Table 14. Kappa values for every ML method

	Min	1st q.	Median	Mean	3rd q.	Max	NA's
LDA	0.29502	0.36939	0.40778	0.39578	0.41871	0.49043	0
CART	0.20071	0.23870	0.25163	0.25321	0.27418	0.31165	0
k-NN	0.26017	0.28283	0.32845	0.32028	0.35275	0.37103	0
SVM	0.46987	0.49758	0.52276	0.53305	0.56649	0.62894	0
RF	0.63456	0.71560	0.74102	0.73840	0.76733	0.81943	0

Table 15. All final evaluation metrics with RF using 10-fold-CV

Mape	Accuracy	Kappa	Cor	min_max accuracy
0.064988	0.797122	0.7543	0.828797	0.946254

Figure 5. Comparison of the 5 models

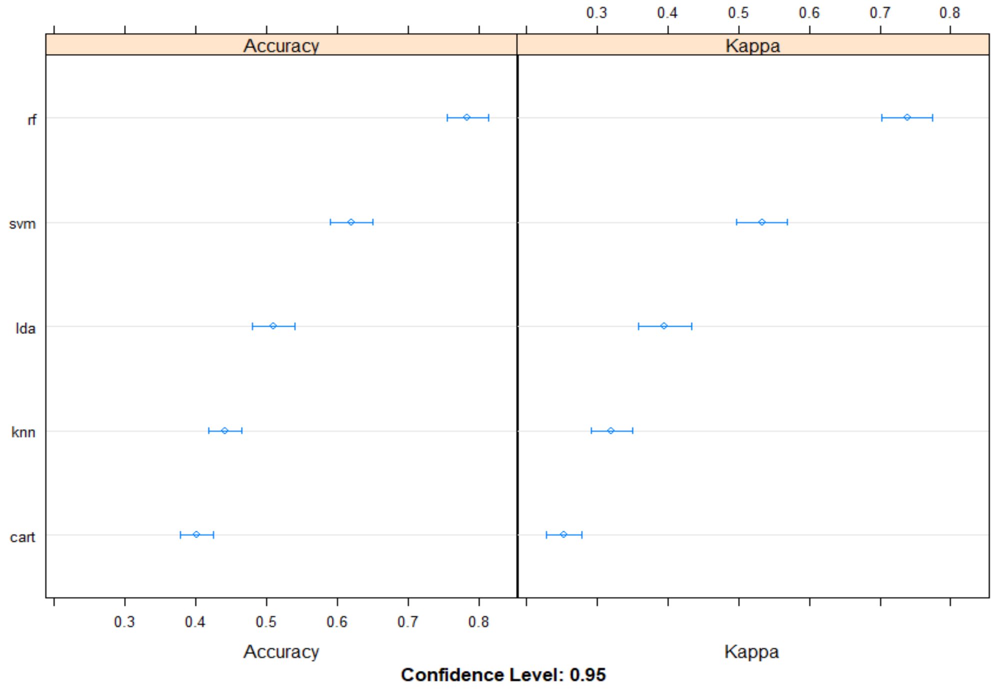
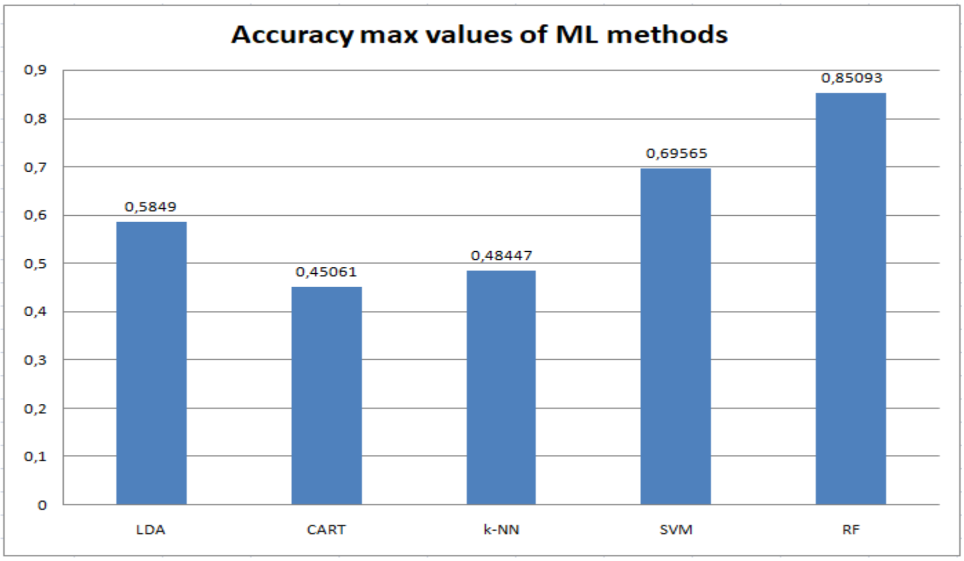


Figure 6.max Accuracy values for all ML methods



5. Results

In the first part of this experiment, we tried to predict wine quality using linear methods, such as multiple linear regression, with the assistance of the LASSO method. Given that quality is a straight discrete variable, measured from 0 to 10, this method gave poor results. We couldn't improve it as the multivariate linear regression case(see 2.4.1) because we believe this problem appears from the inability of LASSO method to consider interaction terms . Recently there are some interesting research papers dealing with the problem, Bien et al. (2013). We hope for resolving it in the future with the use of them. At the moment our best results are obtained by the methods that are described below.

Subsequently, we tried non-linear methods and started with decision trees and their successor, namely Random Forest. Using decision trees we had poor results even worse than those of linear methods. The method of the Random Forest gave us very promising results.

To test and check the validity of these results we introduced not only the 10-fold cross-validation method but, using the same method and the same dataset, another four ML methods, namely, Linear Discriminant Analysis (LDA), Classification and Regression Trees (CART), k-Nearest Neighbors (kNN) and Support Vector Machines (SVM).

The resulting comparison concludes that Random Forest is the Machine Learning method with the best prediction results, achieving almost 95% with large precedence.

6. Discussion

The theoretical implications of this paper are presented below:

The Lasso method used in this experiment showed that in cases where there are only few independent variables, there are no compelling reasons for its employment.On the contrary, Machine Learning methods, such as Random Forest, have a high predictive content as compared to the other four ML methods. Furthermore, in cases where a dependent variable is discrete and having more than two values, Machine Learning methods prove to be a very good (perhaps the best available) choice as this can is confirmed in Tables 13, 14.

Practical implications are as follows:

In problems of supervised learning, like those faced in this study, with numerical or quantitative variables and a discrete independent variable, a superset of decision trees, like the Random Forest method, is capable of achieving very promising prediction results, surpassing similar studies that use important ML methods, such as SVM or Neural Networks (Cortez et al., 2008).

Evaluation metrics often play a decisive role in the prediction process; therefore, using many meaningful metrics in this context, provides a wide range of results. In other words, the key role is not played only by Accuracy, which is a rigid metric, but also by other metrics, such as the Mean Absolute Percentage Error (MAPE) and the one introduced in this paper, namely, *min_max_accuracy*. Instead of using the tolerance index (Cortez et al., 2008), it is much more effective to use the two metrics above.

7. Conclusions

Assessment of wine quality has been a field of research for many decades. For many wineries around the world the main method for assessing wine quality still is human wine tasters. In the last decades, having the physicochemical properties of the wine in digital format allowed the use of Machine Learning methods in an effort to rate wine quality solely based on its physicochemical properties.

In this paper, two main ML methods were introduced: the LASSO, which belongs to the linear regression family, and the non-linear method of Random Forest, which is the rich descendant of the method of decision trees. Moreover, using the very precise 10-fold cross-validation to test the results, a large comparison with another four well-known ML methods is possible. Results affirmed that Random Forest was the best predicting method.

Our results suggest that a mixed model combining non-linear regressions, such as the appropriate (as this can be judged by the R^2) polynomial regression, and the use of Random Forest could substantially increase the percentage of right predictions coming up to almost 95%! Our results were derived from the study of wine data, but they can certainly find much wider applications. In effect, every dataset with a discrete dependent variable and low correlations between independent variables is a candidate for applying our mixed model suggested. From the above, it follows that the LASSO method could be of great help in improving the selection of appropriate independent variables from our dataset, and decreases their number accordingly, as a crucial step towards predicting and classifying the entire dataset. However LASSO's predictability is not as efficiency as the non linear methods ,at least in our belief, because its inability to deal with interactions terms. We stay it as an open problem and we hope to come on it in the near future

APPENDIX

Acronyms explained

CART	Classification and Regression Trees
CV	cross-validation
DT	Decision Trees
DWT	Discrete Wavelet Transform
EU	European Union
kNN	k-Nearest Neighbors
LASSO	Least Absolute Shrinkage and Selection Operator
LDA	Linear Discriminant Analysis
MAD	Mean Absolute Deviation
MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage Error
ML	Machine Learning
MR	Multiple Regression
MSE	Mean Squared Error
NN	Neural Networks
PCA	Principal Component Analysis
PCC	Pearson Correlation Coefficient
PDO	Protected Designation of Origin
PGI	Protected Geographical Indication
PVE	Percentage of Variance Explained
RBFNN	Radial Basis Function Neural Networks
RF	Random Forest
RMSE	Root Mean Squared Error
SVM	Support Vector Machines
WTP	Willingness To Pay

References

Abbal, P., Sablayrolles, J. M., Matzner-Lober, E., &Carbonneau, A. (2018). A Model for Predicting Wine Quality in a Rhône Valley Vineyard. *Agronomy Journal.* https://doi.org/10.2134/agronj2018.04.0269.

Abbal, P., Sablayrolles, J. M., Matzner-Lober, É.,Boursiquot, J. M., Baudrit, C., &Carbonneau, A. (2016). A decision support system for vine growers based on a Bayesian network. *Journal of agricultural, biological, and environmental statistics*, *21*(1), 131-151. https://doi.org/10.1007/s13253-015-0233-2.

Arvanitoyannis, I. S., Katsota, M. N., Psarra, E. P., Soufleros, E. H., & Kallithraka, S. (1999). Application of quality control methods for assessing wine authenticity: Use of multivariate analysis (chemometrics). *Trends in Food Science & Technology*, 10(10), 321-336.https://doi.org/10.1016/S0924-2244(99)00053-9.

Ashenfelter, O. (2008). Predicting the quality and prices of Bordeaux wine. *The Economic Journal*, 118(529), F174-F184. https://doi.org/10.1111/j.1468-0297.2008.02148.x

Astray, G., Mejuto, J. C., Martínez-Martínez, V., Nevares, I., Alamo-Sanza, M., &Simal-Gandara, J. (2019). Prediction Models to Control Aging Time in Red Wine. *Molecules*, 24(5), 826. https://doi.org/10.3390/molecules24050826

Athanasiadis I., Ioannides D., (2015). A Statistical Analysis of Big Web Market Data Structure Using a Big Dataset of Wines. *Procedia Economics and Finance*, *33*, 256-268. https://doi.org/10.1016/S2212-5671(15)01710-4

Attwal, K.P.S. and Dhiman, A.S. (2020). Investigation and comparative analysis of data mining techniques for the prediction of crop yield. *Int. J. Sustainable Agricultural Management and Informatics*, Vol. 6, No. 1, pp.43–74.

Beltrán, N. H., Duarte-Mermoud, M. A., Vicencio, V. A. S., Salah, S. A., & Bustos, M. A. (2008). Chilean wine classification using volatile organic compounds data obtained with a fast GC analyzer. *IEEE Transactions on Instrumentation and Measurement*, 57(11), 2421-2436.https://doi.org/10.1109/TIM.2008.925015

Bien, J., Taylor, J., & Tibshirani, R. (2013). A lasso for hierarchical interactions. *Annals of statistics*, 41(3), 1111.

Breiman, L. (2001). Random Forests. *Machine Learning 45*(1), 5-32

Browne, M.W. (2000). Cross-Validation Methods. *Journal of Mathematical Psychology*, Volume 44, Issue 1, 2000, Pages 108-132, ISSN 0022-2496, https://doi.org/10.1006/jmps.1999.1279

Cortez Paulo, António Cerdeira, Fernando Almeida, Telmo Matos, José Reis, (2009). Modeling wine preferences by data mining from physicochemical properties. *Journal Decision Support Systems*, Volume 47, Issue 4, November 2009, Pages 547–553. https://doi.org/10.1016/j.dss.2009.05.016

Fathi, M.T. and Ezziyyani, M. (2019). How can data mining help us to predict the influence ofclimate change on Mediterranean agriculture?. *Int. J. Sustainable AgriculturalManagement and Informatics*, Vol. 5, Nos. 2/3, pp.168–180.

Frank, I. E., & Kowalski, B. R. (1984). Prediction of wine quality and geographic origin from chemical measurements by partial least-squares regression modeling. *Analytica Chimica Acta*, 162, 241-251. https://doi.org/10.1016/S0003-2670(00)84245-2

Grömping, U. (2006). Relative importance for linear regression in R: the package relaimpo. *Journal of statistical software*, *17*(1), 1-27.

Gustafson, C. R., Lybbert, T. J., Sumner, D. A. (2016). Consumer sorting and hedonic valuation of wine attributes: exploiting data from a field experiment. *Agricultural economics*, *47*(1), 91-103. https://doi.org/10.1111/agec.12212

Guyon I. and Elisseeff A., (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3(7–8):1157–1182.

Hair F.J. et al., (2014). Multivariate Data Analysis. (7th ed.), *Pearson Ed.*

Kallithraka, S., Arvanitoyannis,I.S.,Kefalas, P.,El-Zajouli, A., Soufleros, E., & Psarra, E. (2001). Instrumental and sensory analysis of Greek wines; implementation of principal component analysis (PCA) for classification according to geographical origin. *Food Chemistry*,73(4),501-514.https://doi.org/10.1016/S0308-8146(00)00327-7

Lantz Brett, 2013. Machine Learning with R, (2nd ed.), *Packt Publishing*

Legin, A., Rudnitskaya, A., Lvova, L., Vlasov, Y., Di Natale, C., &D'amico, A. (2003). Evaluation of Italian wine by the electronic tongue: recognition, quantitative analysis and correlation with human sensory perception. *Analytica Chimica Acta*, *484*(1), 33-44. https://doi.org/10.1016/S0003-2670(03)00301-5

Lindeman RH, Merenda PF, Gold RZ (1980). Introduction to Bivariate and Multivariate Analysis. Scott, Foresman, Glenview, IL.

Lindsey C., Sheather S., (2010). Variable selection in linear regression. *The Stata Journal, 10, nr.4, pp.650-669.* https://doi.org/10.1177%2F1536867X1101000407

Mendenhall W.,Sincich T., (2012). A Second Course in Statistics. *Regression Analysis,* (7thed.),Prentice Hall ed.

Moscovici, D. and Gottlieb, P.D. (2017). Finding a state of sustainable wine: implications forsustainable viticulture and oenology in New Jersey. USA, *Int. J. Sustainable Agricultural Management and Informatics*, Vol. 3, No. 3, pp.196–214.

Safi, S. K., & White, A. (2017). Short and long-term forecasting using artificial neural networks for stock prices in Palestine: a comparative study. *Electronic Journal of Applied Statistical Analysis*, 10(1).

Smith D. and Margolskee R., (2006). Making sense of taste. *Scientific American*, Special issue, 16(3):84–92. https://doi.org/10.1038/scientificamerican0906-84sp

Thiene, Mara, Riccardo Scarpa, Luigi Galletto, and Vasco Boatto. (2013). Sparkling wine choice from supermarket shelves: the impact of certification of origin and production practices. *Agricultural Economics* 44, no. 4-5, 523-536. https://doi.org/10.1111/agec.12036

Vlontzos, G. and Pardalos, P.M. (2017). Data mining and optimisation issues in the foodindustry. *Int. J. Sustainable Agricultural Management and Informatics*, Vol. 3, No. 1, pp.44–64.

Vlachou, M. (2011). Technical interventions during vinification of the opsimos edessis wine variety. *Master's thesis, Chemical Department, Aristotle University of Thessaloniki, Greece,* http://ikee.lib.auth.gr/record/129139/files/GRI-2012-8598.pdf. [Online; accessed 01-August-2020].

Yu, H., Lin, H., Xu, H., Ying, Y., Li, B., & Pan, X. (2008). Prediction of enological parameters and discrimination of rice wine age using least-squares support vector machines and near-infrared spectroscopy. *Journal of agricultural and food chemistry*, 56(2), 307-313.https://doi.org/10.1021/jf0725575

DATA AND CODE AVAILABILITY:

The datasets generated and/or analyzed in the current study are not publicly available; they are confidential and the property of the winery firm but they are available from the corresponding author on reasonable request.

The source code used during the current study is available from the corresponding author on reasonable request.