Electronic Journal of Applied Statistical Analysis
EJASA, Electron. J. App. Stat. Anal.
http://siba-ese.unisalento.it/index.php/ejasa/index
e-ISSN: 2070-5948
DOI: 10.1285/i20705948v14n2p389

Selection of features and prediction of wine quality using artificial neural networks

By Athanasiadis, Ioannides

Published: 20 November 2021

Electronic Journal of Applied Statistical Analysis
Vol. 14, Issue 02, November 2021, 389-416
DOI: 10.1285/i20705948v14n2p389

Selection of features and prediction of wine quality using artificial neural networks

Ioannis Athanasiadis*and Dimitrios Ioannides

University of Macedonia, Thessaloniki, Greece
Department of Economics

Published: 20 November 2021

Assessing wine taste quality is a key factor for successful sales in the wine industry, where the aim is to meet consumer's demands. This is obtained by the analytic method of wine vinification. Evaluation of the wine's quality as determined by wine experts, according to sensor properties, plays an important role. This study introduces an alternative method for wine quality prediction by using machine learning techniques, such as linear regression and neural networks. The purpose is to make the evaluation process less expensive and less time-consuming at various stages of the production process. Our data analysis is based on a real wine dataset provided by an established winery in Greece. First of all, we study the correlation of quality with physicochemical features of wine considered at the winery. Statistical analysis indicates a significant interaction between the variables involved. We handle this problem by using Artificial Neural Networks(ANN). It is shown that ANN improves the accuracy of predicting wine quality based on the features selected.

keywords: Linear regression, neural networks, physicochemical properties, prediction, statistical methods, wines.

*Corresponding author: athang@uom.edu.gr

ISSN: 2070-5948
http://siba-ese.unisalento.it/index.php/ejasa/index

1. Introduction

The reputation of quality a wine label has as a product and its impact on price cost can influence a winery's business profitability. There is a huge volume of literature concerning the factors affecting wine quality. These include physicochemical features, geographical origins, and climate parameters. Using such features of the wine data mining techniques can help predict its quality. Such techniques are mainly focused on three areas:

- classification according to geographical origin,
- classification and prediction according to age and
- classification and prediction according to the quality

To start with, concerning geographical origins, Frank and Kowalski (1984) attempted to predict quality using the Partial Least Squares (PLS) method and demonstrated that chemical data provide sufficient information to predict geographical origin, specific sensory parameters and the overall quality of a wine label. Furthermore, they showed that sensory data do not provide sufficient information to distinguish wines from different areas, such as France and the USA. Their study was a primitive attempt to predict the geographic origin, using a small number of samples and, therefore, there were many question marks as to the validity of their findings. One of the first attempt to use novel experimental techniques and statistical methods was made by Arvanitoyannis et al. (1999) with the use of principal component analysis (PCA), discriminant analysis (DA), canonical analysis (CA) and cluster analysis (CLA), as well Anova, Linear Discriminant Analysis (LDA). Their work mainly intended to classify and doscover clusters of wines predicting geographic origin but, similar to the work of Frank and Kowalski (1984), Arvanitoyannis et al. (1999) had a mediocre percentage of valid predictions and small samples. Later, Kallithraka et al. (2001) used mainly Principal Component Analysis and multivariate analysis to classify 33 Greek wine labels according to their geographical region and they concluded that anthocyanins and sensory analysis in red wines resulted in an effective distinction between wines from north and south Greek regions,whereas minerals(ions) and phenols did not lead to any valid clustering. Of course, their results are subject to limitations of very small samples similar to the previous ones. Furthermore, Thiene et al. (2013), in Italy, showed that not only does the region play a significant role in wine quality but also the winery and its vinification methods even in the same geographical area and with the same kind of grapes. Such an example is *Prosecco* wine. The study of Thiene et al. was based on a questionnaire filled in by wine clients trying to classify the wines but used relatively obscure and not well-defined statistical methods.

Secondly, for purposes of prediction and classification according to age, Yu et al. (2008) introduced more sophisticated data mining algorithms and methods such as Least Squares Support Vector Machines (LSSVM) combined with near-infrared (NIR) spectra for predicting and discerning rice wine age. They proved that LS-SVM gave better results than discriminant analysis (DA). Interesting though their approach was, their sample was small (147 wine labels) resulting in poor prediction; however, it was promising as a correlation method. Controlling ageing time of red wine was the main target of the study by Astray et al. (2019), who employed prediction models such as Artificial Neural Networks (ANN), Support Vector Machines (SVM) and Random Forest (RF). From the three methods, the RF model has demonstrated to have a significantly powerful capacity to predict wine age, but their work was based on a small sample number (<60) and their prediction metrics were based only on Root Mean Squared Error (RMSE) and Average Absolute Percentage Deviation (AAPD). Nevertheless, there searchers used commercial software for their results, namely, RapidMiner® and EasyNN®. Abbal et al. (2016) proposed a decision-support system for vine-growers to assess the quality of the vineyard to be planted. This was based on a Bayesian network obtaining a predictive performance of around 75%. Their efforts are interesting in that they use this type of network model but, in some way, they depend on human experts who were asked to fill the probability tables of vineyard qualities.

Finally, for the third main area of classification or/and prediction of wine quality, Ashenfelter (2008) modeled the quality and prices of Bordeaux red wine labels observing the weather during the growth period of the grapes and some other variables, such as historical reputation or quantity produced. It was mainly a clustering effort depending on the weather and age rather than a purely well-known regression model. Beltrán et al. (2008) based their work on the information contained in wine aroma/bouquet chromatograms measured with a fast GC analyzer (commercial instrument). Using principal component analysis (PCA) and discrete wavelet transform (DWT) as feature selection techniques, the authors used linear discriminant analysis (LDA) as well novel data mining methods, such as the radial basis function Neural Networks (RBFNN), and support vector machines (SVM) for classifying the wine labels. Their sample was 100 wine labels from three different Chilean wine types. Ten (10) wavelet observations were made for each wine type.The wine was randomly divided into a 90% training group and a 10% testing set of the sample. They concluded that the best results were obtained using wavelet decomposition as a feature extraction method and using SVM with a radial base function (RBF) type of kernel as a classification technique. Their weak point, besides the small sample number is that their percentages in the testing phase range between 37% and 90% and they depended on a commercial wavelet chromatography machine. Cortez et al. (2009) introduced in a big dataset of wine labels (4898 white and 1599 red ones) and three powerful and interesting classification methods, namely, Multiple Regression (MR), Neural Networks (NN) and Support Vector Machines (SVM), in their effort to predict a quality score based on wine physicochemical properties. They concluded that the superiority of SVM over NN is probably due to differences in the training phase. The SVM algorithm guarantees optimum fit, while NN training may fall into a local minimum. This was one of the most influential studies because of the big sample dataset in two types of wines (white and red); however, there were some issues with tolerance and its impact on results, the single prediction metric (i.e., mean absolute deviation, MAD) and the absence of layers in the neural networks.

Research into many other variables and parameters that influence wine quality,such as soil, plants, meteorology, agronomic variables, vineyard parameters, grape characteristics, and enological practices, was the purpose of the study by Abbal et al. (2019).

They constructed a model validated for giving predictive scores of quality and compared with actual scores given by the International Press for 49 modeled wines. It was an interesting study with a wide range of variables, yet with a small sample and with few prediction metrics, such as correlation and marginal error. The emerging phenomenon of "electronic datasets" increases with the big web market data and, consequently, affects any company that wants to maintain its engagement in this field and seeking methods to mine useful information from these big datasets, as shown by Athanasiadis and Ioannides (2015). In this paper the authors, tried to predict the quality score of wines using multiple linear regression and logistic regression and proving that the latter had much better results. On the other hand, these authors did not use many prediction metrics.

This progress of data analysis produced many important tools that even a small company can use with safety today to investigate results for its own benefit. Wine quality needs to be classified and certified. Within this context wine qualification and certification prevents illegal adulteration of wine and ensures quality for the wine market Cortez et al. (2009).

Physicochemical and sensory tests are prerequisites for such a certified wine label, the first given by certain chemical properties indices and the latter mainly provided by human tasting. Taste is the least understood sense in humans and relationships between physicochemical results and sensory analysis are difficult and still not fully understood. (Legin et al., 2003 and Smith and Margolskee, 2006). Authors have tried to introduce an 'electronic' tongue so as to classify results using PCA and predict using Neural Networks with commercial software (NeuroSolutions®& Unscrambler®). The sample used by Legin et al. (2003) was small (56 winelabels) and reported one prediction metric, i.e., mean percentage error.

Pagliuca and Scarpato (2011), studying olive oil quality, stated that chemical, physical and microbiological characteristics are important, but these have only relative value if the product does not hold appropriate characteristics perceptible to the senses. Sensory analysis is the examination of a consumer product through the evaluation of the attributes perceptible by the human sensory organs, perceiving color, odor, taste, touch, texture, and sound. The use of human senses in judging food quality is part of our daily actions of consumption. Sensory science is used to understand consumer preferences, and to predict edible qualities using measuring instruments. Sensory evaluation defined as the "systematic study of human response to physicochemical properties" of consumer products can be used in very different contexts, ranging from production line to research laboratory and for a wide spectrum of applications. Consequently, Brentari and Zuccolotto (2011) divided the characteristics of their wine dataset into three categories, for the purposes of their work: Label variables, mainly based on geographical origin, Chemical variables, and Sensory variables. This listing is based on average evaluations by a panel of experienced judges concerning the most important sensory characteristics of wine, namely visual, olfactory, gustatory, and intense aromatic persistence.

Against this background, some years later, Gustafson et al. (2016) tried to correlate willingness to pay (WTP) with hedonic valuation among wine consumers, using labels produced from twelve different regions of the USA and seven varieties, with the price of another six alternative labels. The sample consisted of about 280 wines and 250 consumers. They almost proved that wine appellation has little to do with one's willingness to pay but it does relate to other wine quality attributes.

In our study we attempt to predict white wine quality based on physicochemical data from a winery in Northern Greece. The central role for this study is played by its typicality(formality). In general, the term 'formality' in a 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 reduce the wine quality, simply because a wine critic, taster or judge thinks that a wine is not typical and provides a low rating.

Thus, the formality can give information about the place of wine production, the time of grape harvest, the label category, and the way of vinification. It is further examined by absence of defects, taste and aromatic balance of each wine label, formality in terms of the expression of a variety (for single-variety wines) or the region for PDO (protected designation of origin) or PGI (protected geographical indication) labels.Of course, physicochemical measurements of wine characteristics are provided. The protocol of this vinification is described in the Master's Thesis of 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 the evaluation of quality will be provided in the next section. Therefore, the aim of our research is to evaluate quality for wine from ethe main vineyards of Northern Greece. In Section 2.1, we present and analyze primary white wine data and present an exploratory analysis of the wine labels. We started by building the full regression model of all explanatory variables in Section 2.2. The next step is to examine the use of full stepwise regression approach for model building and some other alternative approach, such as a forward, backward, and bidirectional stepwise selection method for deciding on the variables. A summary of these regression results is presented in 2.3. The residual analysis presented in 2.4 reveals that some assumptions for our linear models are not valid, consequently, we apply a nonlinear model as an alternative method to predict quality. Some significant interaction between certain explanatory variables is described in section 2.5. We tackle this problem using Neural Networks. Thus, in Section 3.1, a brief description of Artificial Neural Networks(ANN) is presented, while section 3.2 discusses the data setup. Evaluation of the model's prediction is given in 3.3, while a graphical visualization of the metrics applied is found in 3.4.

To predict the quality score, previous results are split into two parts: the training part accounts for 75% of our data and the testing part for 25%. Tables and graphics group critical results in sections 3.3-3.4. Finally, all conclusions and results are summarized in Section 4.

2. Materials and methods

2.1. Wine data

The data used in our research were provided by a winery in Northern Greece that enjoys a big market share, not only in Greece but also in Europe. This dataset includes 2312 records of wine labels with almost all physicochemical properties, spanning a period of about 9 years of production (from 2004 to 2013). The white grape variety that was selected was the native Vitis vinifera L. var. ‘Opsimos Edessis’. For more details of this variety and the vinification process (Vlachou, 2011).

Our study mainly aims to provide methods for predicting wine quality score based on physicochemical results and sensory analysis. Sensory evaluation has evolved into a scientific discipline, which quantifies, analyzes, and interprets reactions to sensory wine characteristics perceived by sight, smell, taste, and touch (Noble, 1988). Finding a way in an environment of increasingly larger massive datasets generated by production in order to predict a quality score and guide business decisions, accordingly, offers a competitive advantage. Inputs include objective tests (e.g., pH values) and output is based on sensory data, which, in our case, concern the wine quality data ranging from 3 to 10. At this point it should be mentioned that these ratings are given after alcoholic fermentation (clearing, initial filtering) has been completed and chemical and organoleptic control of the wine labels by winery experts, who mainly taste the balance (correlation with wine chemical characteristics) and the aromatic state/bouquet of the wine (expression of variety, formality) has taken place. Based on this classification, blends (of the same or different varieties) are determined before finalizing the end product.

Thus, in our case, inputs(features) include 8 variables (or physicochemical features), namely: alcohol(alc), sugar(su), total acidity (ta), volatile acidity (va), pH (ph), free sulfur dioxide (fsd), color intensity (col), total sulfur dioxide (tsd), and the outputs of our sensor analysis, as described above. All these measurements are made from the winery’s chemical laboratory.

The main goal is to predict wine quality based on a range of its physicochemical properties. This study focuses on a white wine set but the same procedure could be applied for any other type of wine. The purpose is dual: firstly,to optimally reduce the number of independent variables using multiple linear regression, and, secondly, to find the most efficient prediction method using neural networks. This combination aims at achieving a faster and more precise way for selection and prediction. Implementation of all methods is performed through the R language.

Table 1: Physicochemical properties of white wines

Attribute (units)	min	max	mean
quality	3	10	6.568
alcohol (% vol.)	9	13.65	11.71
pH	2.7	3.98	3.38
total acidity (g(tartaric acid)/dm3)	3.37	8.93	5.01
volatile acidity (g(acetic acid)/dm3)	0.100	1.430	0.306
sugar (g/dm3)	0.50	36	5.14
color intensity	0.030	3.080	0.077
free sulfur dioxide (mg/dm3)	6	65	38.5
total sulfur dioxide (mg/dm3)	26	248	138.7

Each entry denotes a given test (analytical and sensory) and the final database was exported into a single sheet (.csv). A description of the characteristics of our data is presented in Table 1. The data set is available upon reasonable request. The set contains 2312 instances with 8 features for white wine and tastings of its quality. These features include alcohol, pH,total acidity, volatile acidity, sugar, color intensity, free sulfur dioxide and total sulfur dioxide. From Table 1, one easily recognizes that most wine labels contain at least 9% and, at most, 13.65% alcohol, and range from 2.7 to 3.98 on the pH scale. Ranges for other characteristics are also presented. For the dependent variable (quality) the values range from 3 to 10, with a mean value of 6.568.

The importance of quality for our study is shown separately in Figure 1 as a histogram.

In Figure 2 the distribution plots for various characteristics of our data is presented.

Almost all distribution plots in Figure 2 appear to be asymmetric, and of course the Quality in Figure 1.

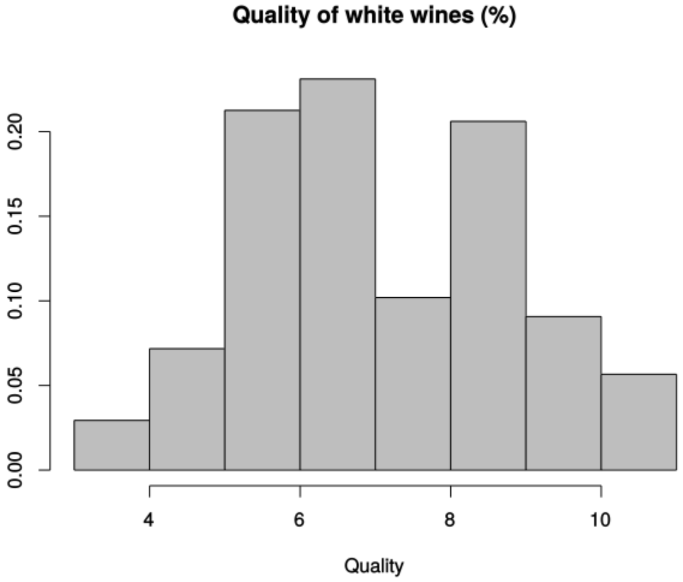

Figure 1: Quality histogram

Table 2: Wine Labels per Year

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

Table 3: Annual mean value per variable

means per year	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)/dm3)	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)/dm3)	0.43	0.32	0.28	0.27	0.34	0.35	0.32	0.26	0.28	0.32
sugar (g/dm3)	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/dm3)	38.6	39.11	39.12	37.89	39.29	39.02	36.92	39.53	39.02	37.61
total sulfur dioxide (mg/dm3)	156.6	142.29	132.2	136.3	142.32	143.61	142.32	138.83	133.12	129.58

To indicate an impression of how characteristics fluctuate from year to year, Table 2 presents the sample scale for each variable, while Table 3 presents annual mean values. Deviations from the mean values are, at most, one unit, which is compatible with the information the winery expert provided to us.

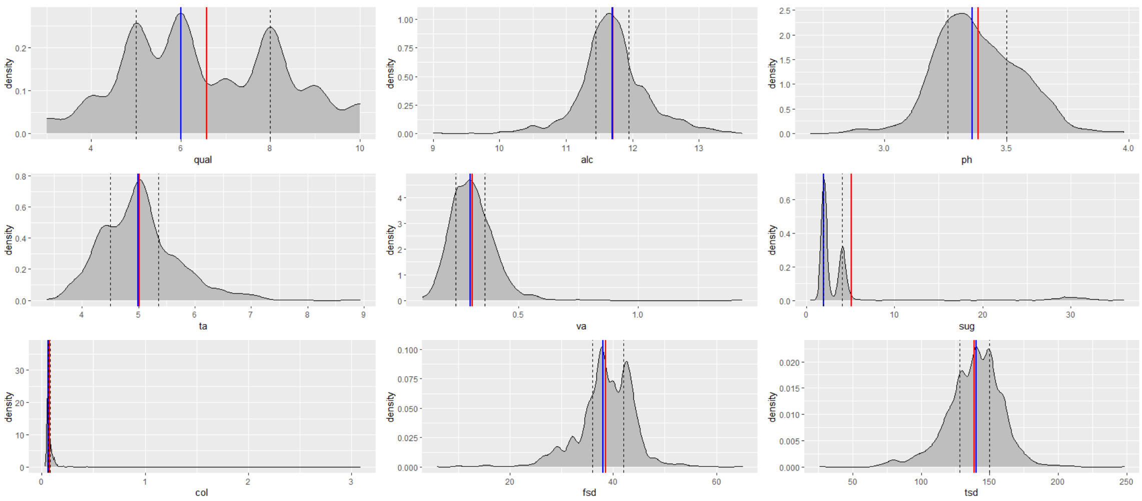

Figure 2: All variable distibutions(red:mean, blue: median, dotted: quartiles)

In Figure 2 red lines show the mean and blue lines the median value of distribution. Dotted lines indicate the 1st and 3rd quartiles.

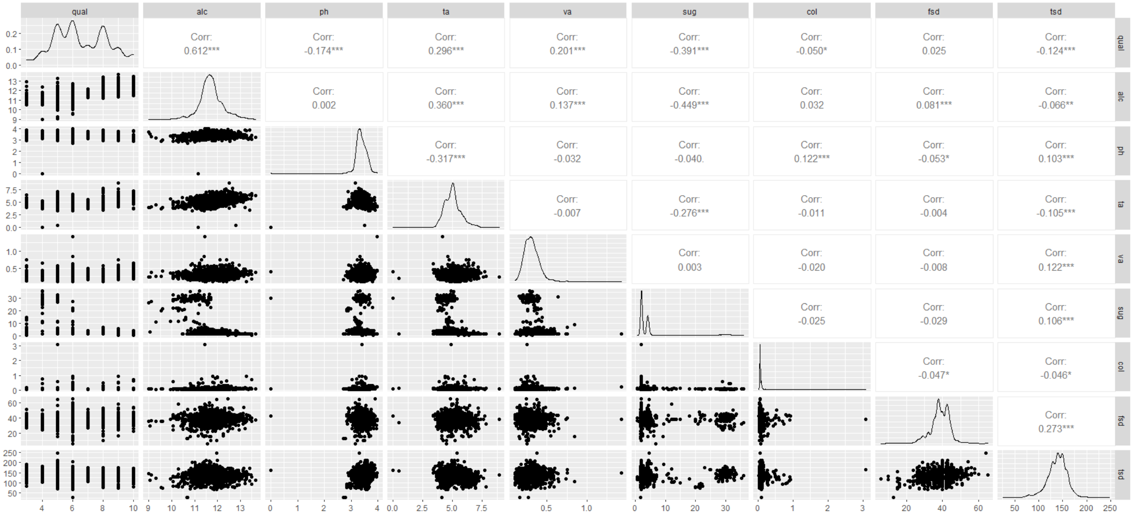

Figure 3: Correlation plot presenting all variables

Figure 3 is the scatter plot of characteristics along with their correlation and p-value.

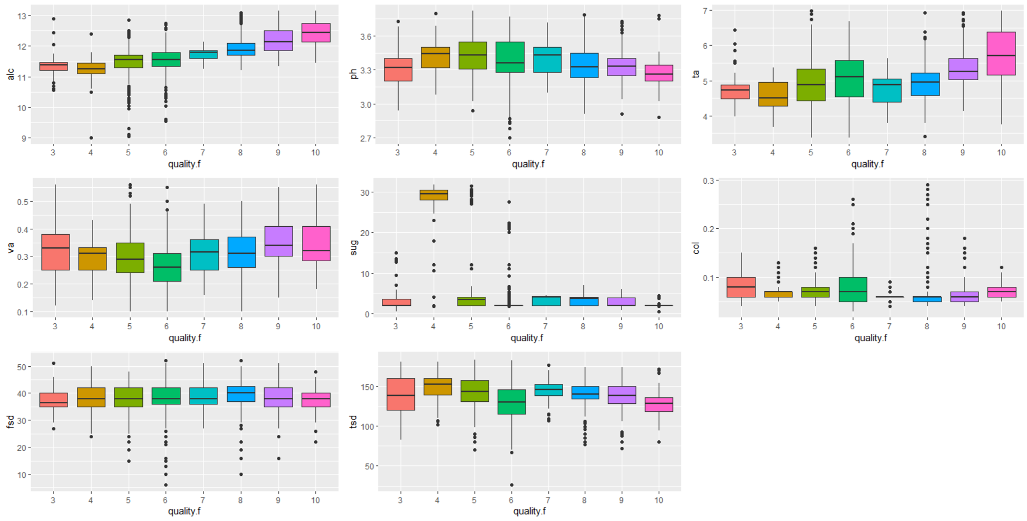

Figure 4: All variables vs. Quality

The box plots in Fig. 4 illustrate the impact of variables on quality. The first box plot shows that high quality corresponds to high alcohol; in similar way we can draw conclusions concerning other pairs.

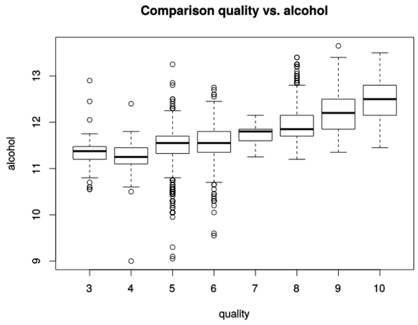

Figure 5: (Detail of Fig. 4) Quality vs. alcohol

To highlight the significance of quality versus alcohol its box plots are presented separately in Figure 5.

Table 4 indicates that the highest positive correlation is between Quality and Alcohol(0.61), followed by Alcohol and Total Acidity(0.36), and then follow negative values between Alcohol and Sugar(-0.45), Quality and Sugar(-0.39), and, finally, Total Acidity and pH (-0.32).

Table 4: The correlation coefficients

	qual	alc	ph	ta	va	sug	col	fsd	tsd
qual	1.00								
alc	0.61	1.00							
ph	-0.17	0.00	1.00						
ta	0.30	0.36	-0.32	1.00					
va	0.20	0.14	-0.03	-0.01	1.00				
sug	-0.39	-0.45	-0.04	-0.28	0.00	1.00			
col	-0.05	0.03	0.12	-0.01	-0.02	-0.03	1.00		
fsd	0.03	0.08	-0.05	0.00	-0.01	-0.03	-0.05	1.00	
tsd	-0.12	-0.07	0.10	-0.10	0.12	0.11	-0.05	0.27	1.00

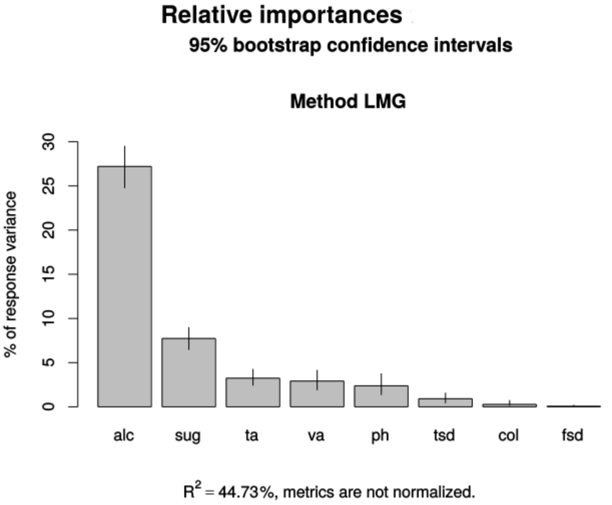

Figure 6: Relative importance for quality at 95% bootstrap confidence intervals (Lindeman, Merenda, Gold LMG method)

Another measure for the significance of each predictor (feature) is given by the LMG (Lindenman-Merenda-Gold) method , see Grömping et al. (2006) and Lindeman et al. (1980).

The Relative importance of the predictors(features) for quality) are depicted in (fig. 6), and their numerical values are :

alc (0.272), sug (0.077), ta (0.032), va (0.029), ph (0.024), tsd (0.009), col (0.003), fsd(0.001)

Here, LMG calculates the relative contribution of each predictor to the R^2 with considering the sequence of predictors appearing in the model.

The entire analysis above suggests that multivariate regression techniques should be applied for determining the relationship between wine quality and wine physicochemical characteristics.

2.2. Multiple Linear regression methods and selection of variables

2.2.1. Full selection method

In the full selection method, we examine all candidate variables for purposes of possible elimination. A description of the algorithm is given in Appendix A, according to James et al. (2013). As per Table 6 and using the Mallows $C(p)$ -method, there are two models close to 9, as Table 5 indicates, namely the 6-features [alc, ph, va, sug, col, tsd] model and the 7-feature model [alc, ph, ta, va, sug, col, tsd], that the one chosen is the model with the highest adjusted R^2. Figure 7 presents the selection procedure by R^2, $C(p)$, Adj.R^2; AIC is also given and leads to the same conclusions.

Table 5: All models

Model Index	Predictors
6	alc ph va sug col tsd
7	alc ph ta va sug col tsd
8	alc ph ta va sug col fsd tsd

Table 6: Subsets - Regression Summary

Model	Adj. R-Square	Pred. R-Square	R-Square	C(p)	AIC	SBIC	SBC	MSEP	FPE	HSP	APC
1	0.3762	0.3760	0.375	328.9534	8082.3817	1520.6866	8099.6193	4456.4211	1.9292	8e-04	0.6248
2	0.4144	0.4138	0.4127	169.8491	7938.6313	1377.0537	7961.6148	4185.9696	1.8129	8e-04	0.5872
3	0.4321	0.4313	0.4301	96.9343	7869.5941	1308.1232	7898.3234	4061.0691	1.7596	8e-04	0.5699
4	0.4470	0.4461	0.4445	35.6914	7809.8773	1248.5954	7844.3525	3955.8114	1.7147	7e-04	0.5554
5	0.4512	0.4501	0.4483	19.9141	7794.2282	1233.0155	7834.4493	3927.4320	1.7031	7e-04	0.5516
6	0.4544	0.4530	0.451	8.5075	7782.8286	1221.6901	7828.7955	3906.4304	1.6947	7e-04	0.5489
7	0.4550	0.4533	0.451	8.1042	7782.4183	1221.3013	7834.1311	3904.0544	1.6944	7e-04	0.5488
8	0.4552	0.4534	0.4508	9.0000	7783.3100	1222.2086	7840.7687	3903.8787	1.6951	7e-04	0.5490

```
AIC: Akaike Information Criteria
SBIC: Sawa's Bayesian Information Criteria
SBC: Schwarz Bayesian Criteria
MSEP: Estimated error of prediction, assuming multivariate normality
FPE: Final Prediction Error
HSP: Hocking's Sp
APC: Amemiya Prediction Criteria
```

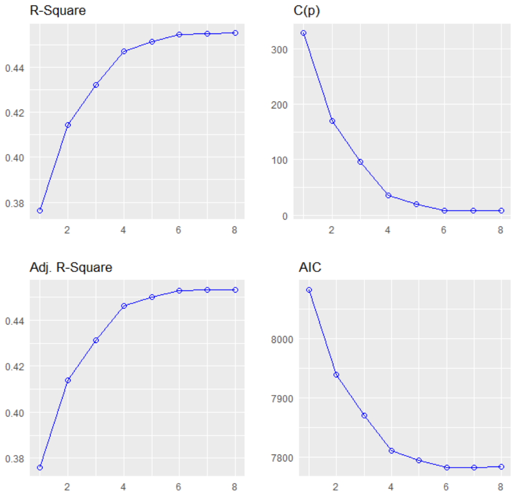

Figure 7: Full model selection

As seen in Fig. 7 and Table 6, all criteria, including Mallows $C(p)$ and Akaike Information Criteria, show that the most efficient number of selected variables is 6.

2.2.2. Forward selection method

Forward selection begins with a model containing no predictors, and then predictors are added to the model, one-at-a-time, until all predictors have been incorporated. More specifically, at each step, the variable providing the greatest additional improvement to the fit is added to the model. The forward stepwise selection procedure is presented in detail in Appendix A (James et al., 2013).

```
Selection Algorithm: forward
         alc ph  ta  va  sug col fsd tsd
1  ( 1 ) "*" " " " " " " " " " " " " " "
2  ( 1 ) "*" "*" " " " " " " " " " " " "
3  ( 1 ) "*" "*" " " " " "*" " " " " " "
4  ( 1 ) "*" "*" " " "*" "*" " " " " " "
5  ( 1 ) "*" "*" " " "*" "*" " " " " "*"
6  ( 1 ) "*" "*" " " "*" "*" "*" " " "*"
7  ( 1 ) "*" "*" " " "*" "*" "*" "*" "*"
8  ( 1 ) "*" "*" "*" "*" "*" "*" "*" "*"

(Intercept)  alc    ph      ta     va    sug    col   fsd    tsd
TRUE        TRUE  TRUE   FALSE   TRUE  TRUE   TRUE FALSE  TRUE

Coefficients:
(Intercept)    alc         ph        sug        va         tsd        col
-8.289395 1.740929  -1.527687 -0.036431 2.607340  -0.006456 -1.031260
```

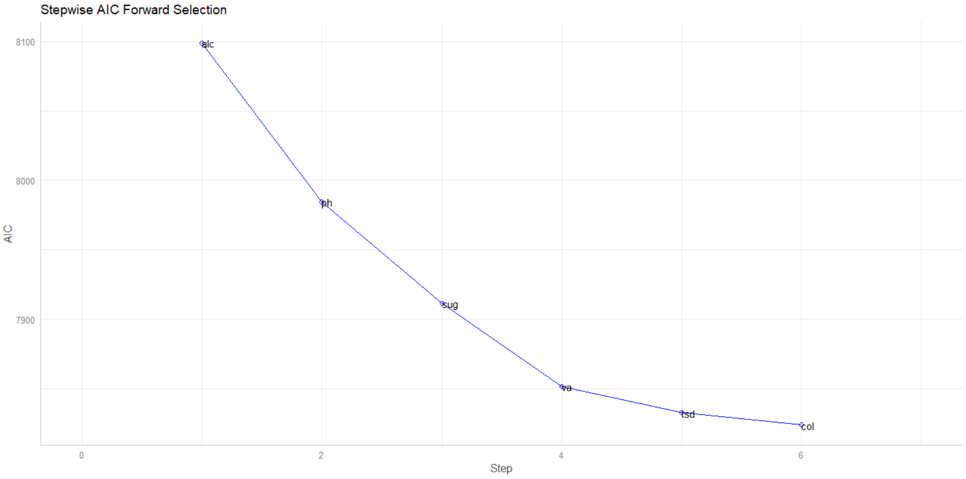

Figure 8: Stepwise AIC forward selection

Fig. 8 presents which variables should remain.

2.2.3. Backward selection method

Similar to the forward one, backward selection provides an efficient alternative for selecting the most efficient subset. However, unlike forward selection, the backward one begins with the full least squares model containing all p predictors, and then iteractively removes the least useful predictor, one-at-a-time. Details are given in Appendix A.

```
Selection Algorithm: backward
         alc ph  ta  va  sug col fsd tsd
1  ( 1 ) "*" " " " " " " " " " " " " " "
2  ( 1 ) "*" "*" " " " " " " " " " " " "
3  ( 1 ) "*" "*" " " " " "*" " " " " " "
4  ( 1 ) "*" "*" " " "*" "*" " " " " " "
5  ( 1 ) "*" "*" " " "*" "*" " " " " "*"
6  ( 1 ) "*" "*" " " "*" "*" "*" " " "*"
7  ( 1 ) "*" "*" " " "*" "*" "*" "*" "*"
8  ( 1 ) "*" "*" "*" "*" "*" "*" "*" "*"

(Intercept)  alc      ph     ta     va    sug     col    fsd    tsd
TRUE        TRUE    TRUE  FALSE   TRUE   TRUE    TRUE FALSE   TRUE

Coefficients:
(Intercept)       alc         ph        va         sug        col        tsd
-8.289395    1.740929 -1.527687 2.607340 -0.036431 -1.031260 -0.006456
```

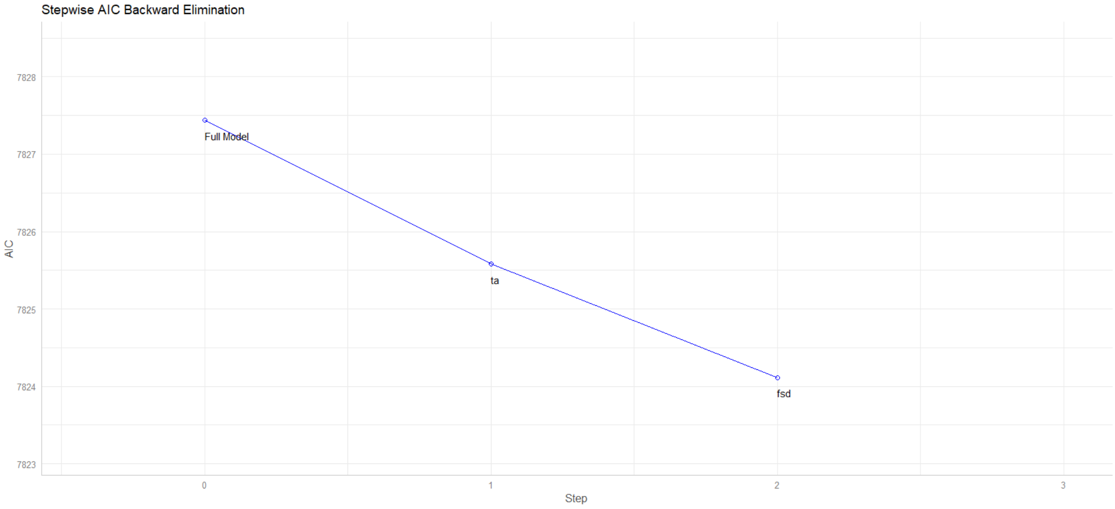

Figure 9: Stepwise AIC backward selection

Fig. 9 shows which variables should be eliminated.

2.2.4. Stepwise (or bidirectional) selection method

Bidirectional method is a combination of the previous two, testing at each step which variables should be included or excluded.

```
Selection Algorithm: exhaustive
alc ph  ta  va  sug col fsd tsd
1  ( 1 ) "*" " " " " " " " " " " " " " "
2  ( 1 ) "*" "*" " " " " " " " " " " " "
3  ( 1 ) "*" "*" " " " " "*" " " " " " "
4  ( 1 ) "*" "*" " " "*" "*" " " " " " "
5  ( 1 ) "*" "*" " " "*" "*" " " " " "*"
6  ( 1 ) "*" "*" " " "*" "*" "*" " " "*"
7  ( 1 ) "*" "*" " " "*" "*" "*" "*" "*"
8  ( 1 ) "*" "*" "*" "*" "*" "*" "*" "*"

(Intercept)  alc    ph    ta    va    sug    col   fsd     tsd
TRUE        TRUE   TRUE FALSE  TRUE    TRUE    TRUE  FALSE    TRUE

Coefficients:
(Intercept)     alc        ph       sug        va       tsd       col
-8.289395  1.740929 -1.527687 -0.036431  2.607340 -0.006456 -1.031260
```

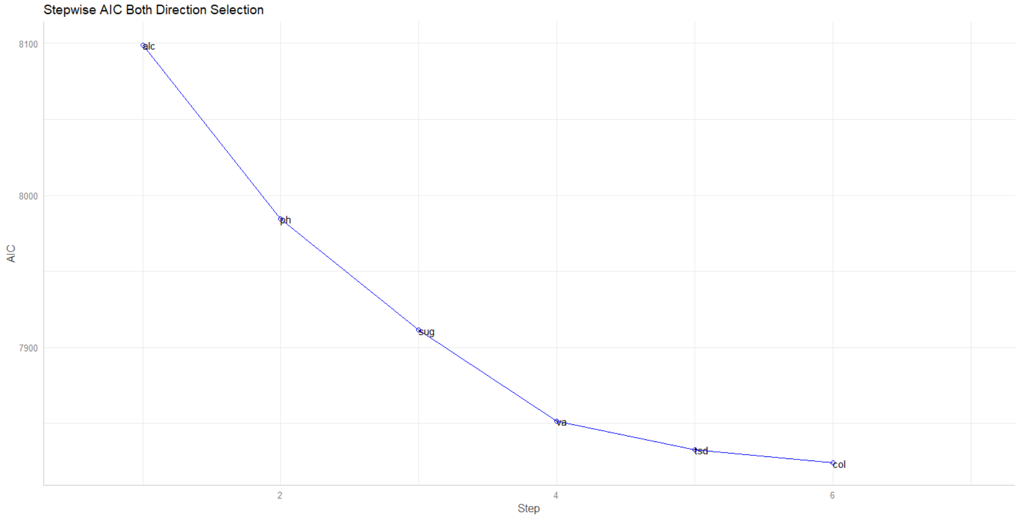

Figure 10: Stepwise AIC bidirectional selection

Fig. 10 shows that these 6 variables should remain based on AIC criteria.

2.3. Estimated Multiple Linear Regression equation

From the analysis above it can be concluded that the optimal Linear regression model for wine quality versus eight predictors, namely, alcohol (alc), sugar(su), total acidity (ta), volatile acidity (va), pH (ph), free sulfur dioxide (fsd), color intensity (col), total sulfur dioxide (tsd),should leave out “total acidity” and “free sulfur dioxide”. The output of this regression is the following(75% training, 25% testing set):

```
Call:
```

```
lm(formula = f_train, data = mydataset_train)
Residuals:
Min      1Q  Median      3Q     Max
-4.3621 -0.8211  0.0338  0.9429  4.0468
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -8.064823   1.030369  -7.827 8.63e-15 ***
alc          1.817574   0.068046  26.711  < 2e-16 ***
ph          -1.774729   0.183188  -9.688  < 2e-16 ***
va           2.478807   0.347115   7.141 1.36e-12 ***
sug         -0.035290   0.004553  -7.752 1.54e-14 ***
col         -2.061575   0.531934  -3.876  0.00011 ***
tsd         -0.007684   0.001582  -4.857 1.30e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 1.285 on 1728 degrees of freedom
Multiple R-squared:  0.4674,Adjusted R-squared:  0.4656
F-statistic: 252.8 on 6 and 1728 DF,  p-value: < 2.2e-16

After predicting in testing set:
    rmse     mape
1.347311 0.188078
```

2.4. Analysis of Residuals

Check if Residuals have **normal** distribution.

```
Shapiro-Wilk normality test

data:  regressor$residuals
W = 0.99346, p-value = 6.03e-07
```

As we can see from the above test residuals do not have normal distribution.
Testing **homoscedasticity** using the **Breusch-Pagan** test:

```
studentized Breusch-Pagan test
data:  regressor
BP = 57.634, df = 6, p-value = 1.36e-10
```

There is **no homoscedasticity** because resulting p-values are much lower than 0.05. Consequently, there is no homogeneous distribution of residuals for the quality dependent variable.

In addition,here is the plot of Residuals versus Fitted values

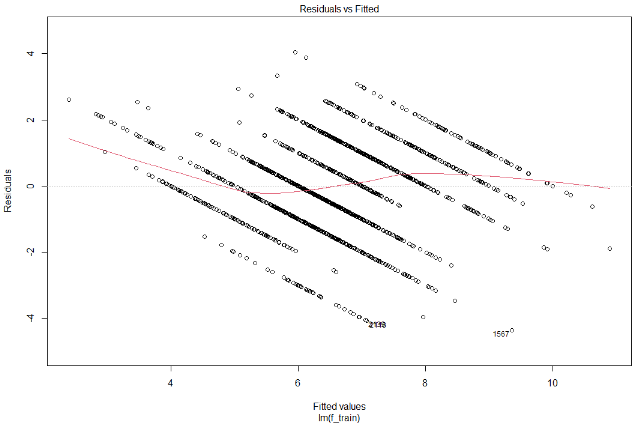

Figure 11: Residuals vs. Fitted values

Collinearity

After investigating whether there is collinearity between variables and discovering none is over the value 5, it is concluded there is **no collinearity** between variables (v.i.f. values: variance inflation factors in parenthesis):

1. alcohol (1.421),2. Sugar(1.301), 3. Total acidity (1.345), 4. Volatile acidity (1.052), 5. pH (1.184), 6. Free sulfur dioxide (1.108), 7. Color intensity (1.020), 8.total sulfur dioxide (1.141).

Durbin - Watson Test for autocorrelation

```
durbinWatsonTest(regressor)
lag Autocorrelation D-W Statistic p-value
1       0.2840115      1.430299       0
Alternative hypothesis: rho != 0
```

Therefore, there is **autocorrelation** because p-value is lower than 0.05.

Our principal goal is to identify which of these remaining variables have a significant impact on wine quality. Fig. 12 below presents the best/most efficient number of variables according to four criteria: Residual Sum Squares (RSS), Adjusted R^2 (Adj.R^2), Mallows $C(p)$ and Bayesian Information Criteria (BIC).

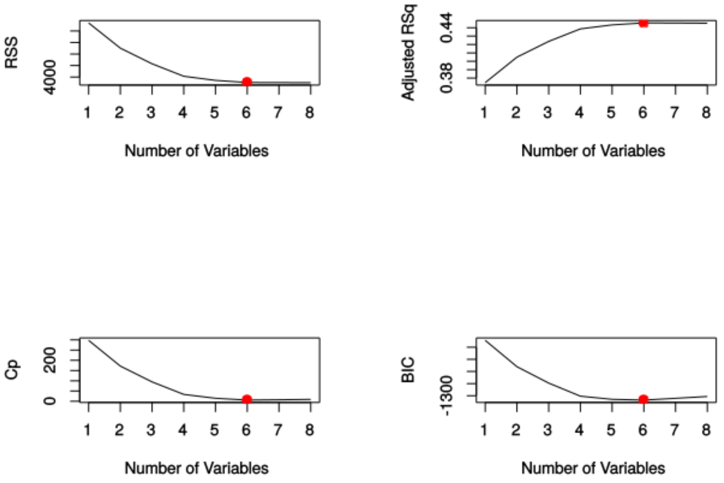

Figure 12: The best/most efficient number of variables using linear regression with RSS, Adj.R^2, Mallows C(p) and BIC criteria

By using k-folders cross-validation (k=10), the mean of errors is calculated, as follows: 1.933323(1), 1.843186(2), 1.783837(3), 1.739202(4), 1.723362(5), 1.721544(6), 1.724151(7), 1.725392(8), to make the prediction for every folder.

The minimum value is the best number of selected variables (1.721544 with 6 vars.). Increased quality is associated with lower pH (ph), sugar(sug), color intensity(col) and total sulfur dioxide(tsd) values. It can be easily concluded that quality is associated with higher values of alcohol(alc), total acidity(ta), volatile acidity(va) and free sulfur dioxide(fsd). All plots above indicate that the six most significant variables for wine quality are: alcohol (alc), sugar(su), volatile acidity (va), pH (ph), color intensity (col), total sulfur dioxide (tsd)).

Since the normality and homoscedasticity of the residuals are not confirmed we proceed to analyze the dataset using a non linear model, namely that of artificial neural networks, in section 3.

2.5. Best combination of interactions in multiple linear regression

In their work, Safi and White (2017) tried to predict stock prices using methods such as Neural Networks, with three different forecasting criteria as primary evaluation metrics: Mean Absolute Error (MAE), Root Mean Squared Error (RMSE),and Mean Absolute Percentage Error (MAPE). We attempted to achieve these metrics mainly using the training and testing statistical test, and dividing the dataset into training and testing, using 75% and 25% of observations, respectively. As mentioned above in the correlation matrix, it was evident that some properties with strong relationships to each other might have a cumulative effect,through interaction,on multiple linear regression. Below is the best combination among all others, presenting the best results, after training with the training set (75%) and testing the prediction with the testing set (25%).

```
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:

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

As seen from the results above there is significant interaction that maximizes prediction metrics through multiple linear regression. There are interactions between Alcohol and Sugar and between Alcohol and Volatile Acidity, while total sulfur dioxide, total acidity and free sulfur dioxide are not entailed at all.

To overcome all these problems with linear methods Artificial Neural Networks are going to be used.

3. Artificial Neural Networks Model

3.1. The artificial neural networks (ANN) approach

Neural networks use neurons as building blocks to construct complex models of data. Although there are numerous variants of neural networks, each can be defined in terms of the following characteristics:

- An activation function, which transforms a neuron's net input signal into a single output signal to be broadcasted further in the network
- A network topology (or architecture), which describes the number of neurons in the model as well as the number of layers and manner in which they are connected
- The training algorithm that specifies how connection weights are set to inhibit or excite neurons in proportion to the input signal (Lantz, 2013).

In this work, the neural network implemented consists of one hidden layer at the first stage, and, secondly, of another hidden layer. We experiment with increasing of dimensions (the number of variables) and with increasing the number of neurons, recording the results at every stage.

3.2. Processing and analyzing data

ANN works best when input data are scaled to a narrow range around zero; here values can be seen ranging anywhere from zero to over two hundred. In this study, the following normalization method has been used:

$$x = a + \frac{x - x_{min}}{x_{max} - x_{min}}\left(b - a\right) \tag{1}$$

where $a = 0.1$ and b = 1
using as activation function the Sigmoid function:

$$y = f\left(x\right) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \ldots + \beta_6 x_6)}} \tag{2}$$

with $x = (x_1,...,x_6)$ and $\beta = (\beta_0,...,\beta_6)$

After the normalization of our data, the dataset was randomly divided into two parts: the training set, for training with our ANN model and the testing set, for validating our predicted values of quality from the training ANN model, using four validation metrics to be analyzed later. The two datasets are divided into 75%(1735 obs.) - 25%. (577 observations), respectively.

As mentioned above, current use of ANN permits the use of several neurons per hidden layer and, also, several hidden layers. Our experiment entails finding the best combination of variables, neurons, and hidden layers, to achieve the best prediction. We start with 2 variables, as ordered above, and at every passage we predict in the testing set using the model of ANN created based on the training set; results are validated with the following metrics:

correlation coefficient (cor), mean absolute percentage error (MAPE), root-mean squared error (RMSE) and Coefficient of Determination R^2.

3.3. Predicting and evaluating model performance

In the previous chapter the use of prediction metrics with multiple linear regression reached the following results (with the 6 best variables):
r = 0.6863, mape = 0.1716, rmse = 1.2797, R^2 = 0.4710

Table 7 below summarizes all results predicting wine quality produced with a different number of features and neurons. The process starts with the 2 most important features (alcohol, sugar) and 1 neuron.

By increasing the features -neurons combinations the best results are found in lines 25 and 36. We realized that the best results are reached using the 6 variables already obtained through the selection methods of multiple linear regression. When all variables selected were used, results were worse.

As seen the cor, RMSE and R^2 metrics are the best so far, including the results from linear methods. The entire source code of the program employed in this section can be found in Appendix B.

Table 7: All the results of metrics with ANN

nr.	dim	neur.	cor (r)	mape	rmse	R^2
1	2	1	0.6541	0.3652	0.1710	0.4275
2	2	2	0.6483	0.3502	0.1725	0.4173
3	2	3	0.6462	0.3503	0.1730	0.4141
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
24	6	4	0.7649	0.3017	0.1456	0.5851
25	**6**	**5**	**0.7672**	**0.2915**	**0.1450**	**0.5884**
26	7	1	0.6896	0.3476	0.1637	0.4751
.	.	.	.	.	.	.
.	.	.	.	.	.	.
35	8	5	0.7499	0.3045	0.1497	0.5614
36	**6**	**6-1**	**0.7756**	**0.2891**	**0.1426**	**0.6016**
37	6	6-2	0.7559	0.2906	0.1483	0.5691

3.4. Visualizing the results

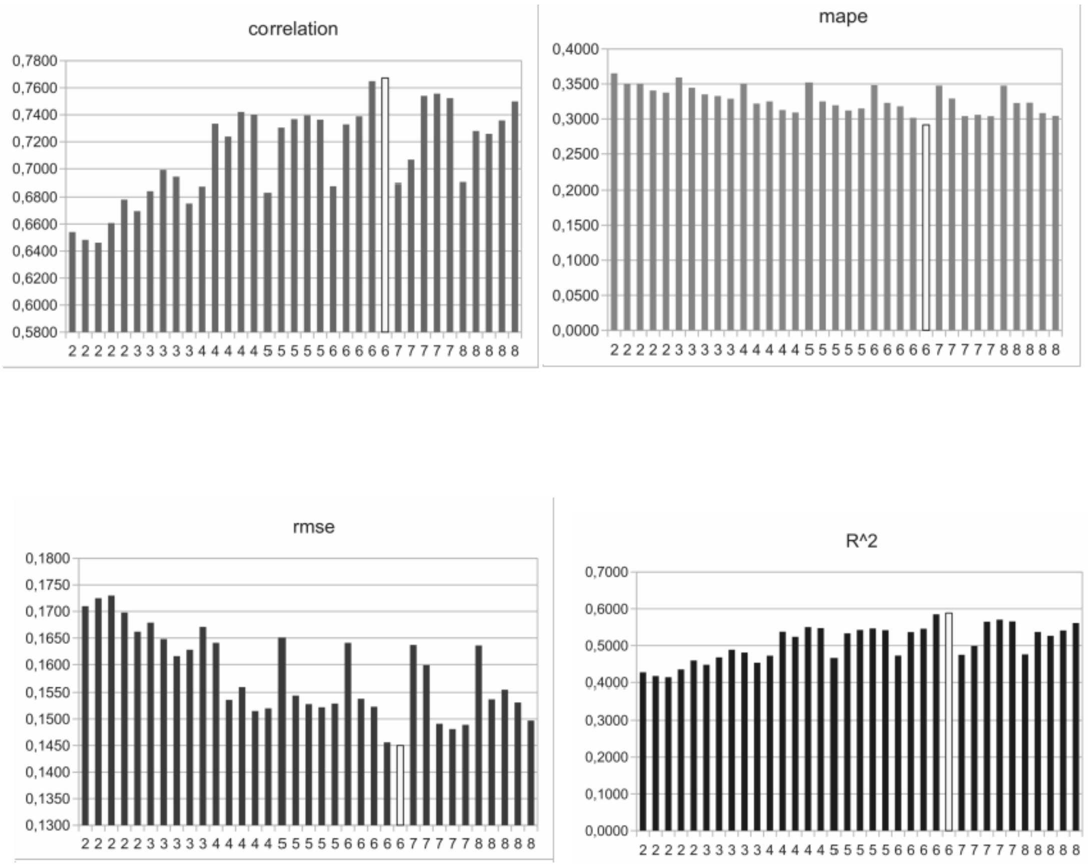

Figure 13: Barplots of metrics(best results in white) with ANN

The plots in Fig. 13 present a clockwise correlation, MAPE, RMSE and R^2 evaluation metrics. As shown in all these metrics, best results are achieved with 6 neurons; in the model, these 6 remaining variables were taken using the selection methods already applied.

4. Conclusion

Having a really big dataset with 2312 wine labels, and their physicochemical properties (or variables), what we tried to achieve in this paper is to predict results examining two main machine learning models using the white wine dataset, namely, the "Multiple Linear Regression" and the "Artificial Neural Networks".

All our efforts used two main datasets for statistical training and testing: the training set with 75% and the testing set with the remaining 25%. First of all, we ran the multiple linear regression model to select what variables from the selection models (backward, forward, and bidirectional) should be kept for further examination. Moreover, we tried to identify any possible interaction in the multiple linear regression and two interactions giving better results were noted.

Then, in the resulting dataset without the variables excluded, the prediction was validated using artificial neural networks (ANN). The ANN is quite heavily dependent on the number of neurons per layer used to construct the topology of the network. All the possible combinations using 2 to 6 variables and 1 to 5 neurons were checked and a hidden layer was added. In ANN research it was found that the best result is given with 6 neurons/layers and 1 neuron in a hidden layer, ensuring the most effective results in four metrics: correlation, mean absolute percentage error (MAPE), root mean squared error (rmse) and R^2.

This conclusion showed that elimination and selection of variables using the methods above is a crucial step because ANN demonstrated that the best/most effective results are obtained when the exact number of 6 variables (not less, nor more), with 5 neurons (n-1) and with only one hidden layer with 6 and 1 neuron, and no more are applied.

Some limitations of this study are as follows. First of all, normalization was used to achieve correct results with ANN and a different, less conventional formula was used to avoid dividing by zero. Furthermore, with ANN the number of variables as neurons in the hidden layer was not exceeded. The ANN model showed in Table 7 that there are some irregularities in the sequence of results. In the part of linear regression, it is crucial to examine the collinearity, normality, and homoscedasticity between variables and, especially, between the dependent variable and the independent variables, checking, of course, all assumptions, including residuals.

This study could be used by any winery having these physicochemical properties available, which would lower the price of blind wine tasting for a large amount of wine produced. On the other hand, the study also showed that some variables are not crucial for the final quality score. Furthermore, research could be undertaken using other activation functions in ANN, other types of NN, and for validation purposes, k-folder cross-validation could be used. Finally, it would be interesting to test all these methods with other wine datasets or other types of datasets and compare the results.

A. APPENDIX: The algorithms for selecting features

Backward :
1. select a significance level to stay in the model (e.g. $\alpha = 0.05$)
2. fit the full model with all predictors
3. consider the predictor with the highest p-value. If p $> \alpha$, go to step 4, otherwise go to END
4. remove the predictors
5. fit the model without this variable and go to step 3
END: your model is ready
Forward :
1. select a significance level to stay in the model (e.g. $\alpha = 0.05$)
2. fit all simple regression models y xn. Select the one with the lowest p-value
3. keep this variable and fit all possible models with one extra predictor added to the one(s) already have
4. consider the predictor with the lowest p-value. If p $< \alpha$, go to step 3, otherwise go to END.
END: keep the previous model.
Bidirectional :
1. select a significance level to enter and to stay in the model (e.g. $\alpha = 0.05$)
2. perform the next step of forward selection (new variables must have p $< \alpha$)
3. perform all steps of backward elimination (old variables must have p $< \alpha$). Go to step 2
4. no new variables can enter and no old variables can exit
END: your model is ready.
Full :
1. select a criterion of the goodness of fit (e.g. AIC, BIC, etc)
2. construct all possible regression models: (2N – 1) total combinations (10 variables means 1023 models)
3. select the one with the best criterion
END: your model is ready

B. APPENDIX: Code used in R (section neural networks)

Prediction with neural networks:

```
# using the package ”neuralnet”
library(”neuralnet”)
# n neurons for each layer
# training the nn model, we use all variables to prove extensively the selection of the 6 variables
myformula <- ”quality ∼ alcohol + sugar + pH + volatile.acidity + total.sulfur.dioxide + color.intensity + total.acidity + free.sulfur.dioxide”
nn_model_1 <- neuralnet(myformula, data=norm_train_set, hidden=n, linear.output=T)
# we generate predictions on the testing dataset
model_results_1 <- compute(nn_model_1, norm_test_set[,2:9])
```

For the prediction with the different metrics we use the following commands:

```
# using the package “MLmetrics”
library(MLmetrics)
cor_1 <- cor(predicted_quality, norm_test_set$quality)
mape_1 <- MAPE(predicted, actual)
rmse_1 <- RMSE(predicted, actual)
rsq2_1 <- R2_Score(predicted, actual)
```

References

Abbal, P., Sablayrolles, J.-M., Matzner-Lober, É., Boursiquot, J.-M., Baudrit, C., and 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.

Abbal, P., Sablayrolles, J.-M., Matzner-Lober, E., and Carbonneau, A. (2019). A model for predicting wine quality in a rhône valley vineyard. *Agronomy Journal*, 111(2):545–554.

Arvanitoyannis, I., Katsota, M., Psarra, E., Soufleros, E., and 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.

Ashenfelter, O. (2008). Predicting the quality and prices of bordeaux wine. *The Economic Journal*, 118(529):F174–F184.

Astray, G., Mejuto, J. C., Martínez-Martínez, V., Nevares, I., Alamo-Sanza, M., and Simal-Gandara, J. (2019). Prediction models to control aging time in red wine. *Molecules*, 24(5):826.

Athanasiadis, I. and 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.

Beltrán, N. H., Duarte-Mermoud, M. A., Vicencio, V. A. S., Salah, S. A., and 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.

Brentari, E. and Zuccolotto, P. (2011). The impact of chemical and sensorial characteristics on the market price of italian red wines. *Electronic Journal of Applied Statistical Analysis*, 4(2):265–276.

Cortez, P., Cerdeira, A., Almeida, F., Matos, T., and Reis, J. (2009). Modeling wine preferences by data mining from physicochemical properties. *Decision Support Systems*, 47(4):547–553.

Frank, I. and Kowalski, B. R. (1984). Prediction of wine quality and geographic origin from chemical measurements by parital least-squares regression modeling. *Analytica Chimica Acta*, 162:241–251.

Grömping, U. et al. (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., and Sumner, D. A. (2016). Consumer sorting and hedonic valuation of wine attributes: exploiting data from a field experiment. *Agricultural economics*, 47(1):91–103.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). *An introduction to statistical learning*, volume 112. Springer.

Kallithraka, S., Arvanitoyannis, I., Kefalas, P., El-Zajouli, A., Soufleros, E., and 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.

Lantz, B. (2013). *Machine learning with R.* Packt publishing ltd.

Legin, A., Rudnitskaya, A., Lvova, L., Vlasov, Y., Di Natale, C., and 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.

Lindeman, R., Merenda, P., and Gold, R. (1980). Introduction to bivariate and multivariate analysis. Technical report, Scott, Foresman.

Noble, A. (1988). Analysis of wine sensory properties. In *Wine Analysis*, pages 9–28. Springer.

Pagliuca, M. M. and Scarpato, D. (2011). Food quality, consumer perception and preferences: An analysis on olive oil. *Electronic Journal of Applied Statistical Analysis*, 4(2):215–226.

Safi, S. K. and 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. V. and Margolskee, R. F. (2006). Making sense of taste. *SciAm*, 16(3):84–92.

Thiene, M., Scarpa, R., Galletto, L., and Boatto, V. (2013). Sparkling wine choice from supermarket shelves: the impact of certification of origin and production practices. *Agricultural Economics*, 44(4-5):523–536.

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., and 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.

DATA AND CODE AVAILABILITY :

The datasets analyzed during the current study are not publicly available due to the policy of the winery's private data but 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.