Peer-reviewed version available at *Molecules* **2019**, *24*, 826; doi:10.3390/molecules24050826

Prediction models to control aging time in red wine

Gonzalo Astray[a], Juan-Carlos Mejuto[a], Víctor Martínez-Martínez[b], Ignacio Nevares[b], Maria Alamo-Sanza[c,*], Jesús Simal-Gandara[d,*]

[a] Department of Physical Chemistry, Faculty of Food Science and Technology, University of Vigo, Ourense Campus, 32004 Ourense, Spain.

[b] Department of Agricultural and Forestry Engineering, UVaMOX-University of Valladolid, Valladolid, Spain.

[c] Department of Analytical Chemistry, UVaMOX-University of Valladolid, Valladolid, Spain.

[d] Nutrition and Bromatology Group, Department of Analytical and Food Chemistry, Faculty of Food Science and Technology, University of Vigo, Ourense Campus, 32004 Ourense, Spain.

Gonzalo Astray [a] (gastray@uvigo.es)

Juan-Carlos Mejuto [a] (xmejuto@uvigo.es)

Víctor Martínez-Martínez [b] (victor.martinez.martinez@uva.es)

Ignacio Nevares [b] (ignacio.nevares@uva.es)

Maria Alamo-Sanza [c,*] (maria.alamo.sanza@uva.es)

Jesus Simal-Gandara [d,*] (jsimal@uvigo.es)

* Corresponding authors:

Maria Alamo-Sanza (maria.alamo.sanza@uva.es)

Jesús Simal-Gándara (jsimal@uvigo.es)

Abstract

A combination of physical-chemical analysis has been used to monitor the aging of red wines from D.O. Toro (Spain). The changes in the chemical composition of wines that occur along aging time can be permitted to discriminate wine samples collected after one, four, seven and ten months of aging. Different computational models were used to develop a good authenticity tool to certificate wines. In this research different models have developed: Artificial Neural Network models (ANNs), Support Vector Machine (SVM) and Random Forest (RF) models. The results obtained for the ANN model developed with sigmoidal function in the output neuron and the RF model permit to determine the aging time, with an average absolute percentage deviation below 1% and it can conclude that these two models have demonstrated its capacity as a valid tool to predict the wine age.

Keywords: Food authenticity; Toro appellation of origin; Prediction Models; Wine; Aging.

Peer-reviewed version available at Molecules 2019, 24, 826; doi:10.3390/molecules24050826

1. Introduction

In the last decade consumers are interested in foods identified with a place of origin (Luykx and van Ruth, 2008), and in their characteristics and quality (Saurina, 2010, Luykx and van Ruth, 2008). One of these products is wine that it a beverage obtained from the alcoholic fermentation of grapes (da Costa, Llobodanin, de Lima, Castro and Barbosa, 2018), and it is one of the most popular (Chen, Tawiah, Palmer and Erol, 2018), complex (Rapeanu, Vicol and Bichescu, 2009) and consumed alcoholic beverages around the world (Hu, Yin, Ma and Liu, 2018). In the European Union (EU) wines produced in specified regions are clearly identified and controlled (Riovanto, Cynkar, Berzaghi and Cozzolino, 2011).

In this sense, there are different quality schemes under a geographical indication according to specific characteristics: i) protected designation of origin (PDO), ii) protected geographical indication (PGI) and iii) geographical indication of spirit drinks and aromatized wines (GI) (European Commission, 2018). As is understandable, the use of these schemes impacts on market recognition and, even, in a higher sale price, due to this, improper use of these geographical indications can be injurious to producers and consumers (Luykx and van Ruth, 2008). South European countries (Spain among others) are involved in food authentication studies, for example in wines and foodstuffs registered as, among others, PDO (Danezis, Tsagkaris, Camin, Brusic and Georgiou, 2016).

Wine adulterations such as water dilution or mixed with cheaper wine, are a common practice even since ancient Greece and Rome (Moldes, Mejuto, Rial-Otero and Simal-Gandara, 2017). Nowadays, the quality and the commercial value are linked to elaboration procedures and geographical places (Moldes, Mejuto, Rial-Otero and Simal-Gandara, 2017), as for example, Tempranillo red wine from D.O (*Denominación de Origen*) Toro (Spain), where the wine authenticity is a key factor in terms of differentiation, which has a significant influence on the final sale price (Moldes, Mejuto, Rial-Otero and Simal-Gandara, 2017).

The wine´s quality and organoleptic properties can be influenced by oenological parameters such as: grape variety, winemaking process, and aging system, among others (Serrano-Lourido, Saurina, Hernández-Cassou and Checa, 2012). As these parameters are related to the wine´s quality/price, the possible to find a relationship between physicochemical parameters and a specific aging practice developed in an appellation of origin (D.O. Toro) can be interesting, especially if the wine´s characterization and its combination with chemometric treatment can be provided good results that also reduce the operative costs compared to other methods like expert panellists (Saurina, 2010). Due to this, different computational models can be used. In this research four

different models are presented; i) two Artificial Neural Networks models (ANNs), ii) one Support Vector Machine (SVM) model and iii) one Random Forest (RF) model.

ANNs are a computational technique developed in the same way that biological neural system (Akintunde, Ajala and Betiku, 2015, Haykin, 1999, Bishop, 1995, Gonzalez-Fernandez, Iglesias-Otero, Esteki, Moldes, Mejuto and Simal-Gandara, 2018). McCulloch and Pitts in their research (McCulloch and Pitts, 1943) introduced the concept of the artificial neuron (Dawson and Wilby, 2001). These interconnected units (artificial neurons or nodes) are able to model complex nonlinear relationships between independent variables (also called inputs) and dependent variables (outputs) (Bishop, 1995, Beck et al., 2013). ANNs model based on a multi-layer perceptron (MLP), one of the most popular ANN topology (Dawson and Wilby, 2001), were used. An MLP is a feed-forward ANN model that maps input data onto output data (RapidMiner GmbH, 2018). This kind of models has multiple layers of neurons (input, hidden and output) with each layer all connected to the next network layer (RapidMiner GmbH, 2018).

One of the most important advantages for ANN is that it can extract information from complex data matrix due its capability to learn the relationship between independent and dependent variables (Chiang and Chang, 2009). According to this advantage, ANNs are applied in many different research fields, such as:

i) Hydrology to model the water quality using different water quality variables (Gazzaz, Yusoff, Aris, Juahir and Ramli, 2012),

ii) in Biotechnology to optimize 1,3-propanediol production using microorganisms like *Lactobacillus brevis* N1E9.3.3 (Narisetty, Astray, Gullón, Castro, Parameswaran and Pandey, 2017) or to optimize oil extraction from *Bauhinia monandra* seed that it is a potential biofuel candidate (Akintunde, Ajala and Betiku, 2015),

iii) in Food technology to develop an authentication model to predict the cultivar, the production type and the harvest date for tomatoes (Hernández Suárez M., Astray Dopazo G., Larios López D. and Espinosa F., 2015) to authenticate extra virgin oil varieties (Bucci, Magrí, Magrí, Marini and Marini, 2002),

iv) in Chemistry to predict percolation temperature (Montoya, Moldes, Cid, Astray, Gálvez and Mejuto, 2015), to predict the solvent accessibility of proteins (Ahmad and Gromiha, 2003), or in other fields where the ANN has proved its capacity for medical, economic or agro-food science purposes (Gonzalez-Fernandez, Iglesias-Otero, Esteki, Moldes, Mejuto and Simal-Gandara, 2018).

SVM was first introduced by Boser *et al.* in 1992 (Capron, Massart and Smeyers-Verbeke, 2007, Boser, Guyon and Vapnik, 1992). Support vector machine is a powerful non-linear method to develop classification and regression models (RapidMiner GmbH, 2018, Ríos-Reina, Elcoroaristizabal, Ocaña-González, García-González, Amigo and Callejón, 2017). An SVM model used input data to constructs a hyperplane, or a group of hyperplanes, in a high-dimensional space (RapidMiner GmbH, 2018). These hyperplanes allow that the SVM model can be used for different purposes (RapidMiner GmbH, 2018). It´s main advantage, in comparison with other classification techniques, for example PLS-DA, is that SVM is flexible to model complex classification non-linear problems (Ríos-Reina, Elcoroaristizabal, Ocaña-González, García-González, Amigo and Callejón, 2017) due to this in many studies and applications, Support Vector Machine models can be applied, such as:

i) to determine air specific heat ratios at elevated pressures (Kamari, Mohammadi, Bahadori and Zendehboudi, 2014),
ii) to classify glaucoma, a progressive optic neuropathy disease (Chan, Lee, Sample, Goldbaum, Weinreb and Sejnowski, 2002),
iii) can be used to forecast electricity load based due to its importance in the regional power system strategy management (Pai and Hong, 2005) or,
iv) even, to real-time crash risk evaluation in the active traffic management (ATM) (Yu, R. and Abdel-Aty, 2013), among other fields.

Random forest is a learning method for classification, or regression (Alhaj and Maghari, 2017, Tian et al., 2017) that was proposed by Breiman in 2001 (Tian et al., 2017, Breiman, Leo, 2001). RF model consists in a classifier with different decision trees, where the final prediction is obtained by all the single classification trees (Tian et al., 2017,Breiman, L., Friedman, Olshen and Stone, 1984), that is, for a quantitative response, the prediction is the average of each individual tree predicted values (Vigneau, Courcoux, Symoneaux, Guérin and Villière, 2018). This method is the key that converts to Random Forest in a powerful prediction method (Vigneau, Courcoux, Symoneaux, Guérin and Villière, 2018). Random forests correct the problem of overfitting that presents the decision trees (Alhaj and Maghari, 2017) and have been used in multiple research fields, such as:

i) Medicine to estimate the survivability of cancer patients within four years or not (Alhaj and Maghari, 2017),

ii) in Food technology to develop a model focused on the volatile organic compounds responsible of the olfactory perception (Vigneau, Courcoux, Symoneaux, Guérin and Villière, 2018),

iii) in Ecology where RF is one of the most used statistical method used for example to classify invasive plants (Cutler et al., 2007), or to estimate high-density biomass for wetland vegetation (Mutanga, Adam and Cho, 2012), *inter alia*.

The objective of this paper is to develop different prediction models as a tool of wine authenticity that could predict the aging time (1-4-7-10 months) of red wines from D.O. Toro (Spain).

2. Materials and methods

A red wine, variety Tempranillo or Tinta Toro, was studied. A total of 15 batches were settled down including different aging systems, three chips system (1, 2 and 3) with three toast levels were studied. The traditional aging system was studied in 6 barrels of mild, medium and strong toast in a duplicated way. Both the barrels and the chips were made with oak French (allier, Q. *sessilis*). All the ageing 225-Liter tanks were used with small doses of oxygen with an equipment (OenoAZ3) simulating the micro-oxygenation produced through wood pores in the barrel. In addition, a control (without contact with wood) wine in stainless steel tank in an inert way was studied during the experiment. In this research, 58 samples reported by Apetrei *et al.* (2012) in their original research were used (Apetrei, Rodríguez-Méndez, Apetrei, Nevares, del Alamo and de Saja, 2012).

Independent variables were obtained by Apetrei *et al.* (2012) using conventional chemical analyses of the wines following international regulations of International Organisation of Vine and Wine (International Organisation of Vine and Wine, OIV, 1990). These parameters were; tartaric acid (T), glycerol (G), potassium (K), total polyphenol index (TPI), alcoholic grade (AD), dry extract (DE), total acidity (TA), volatile acidity (VA), total-SO_2 (T-SO_2), free-SO_2 (F-SO_2), reducing sugars (S), relative density (DEN) and pH.

Data from the original paper were split randomly in three groups, one group used to develop the model (called training group, 35 cases), another group formed by 11 cases (validation group) used to validate the model and a third group to query the selected model (querying group, 12 cases). In this paper, the predictive power of different models was determined as a function of the coefficient of determination (R^2), the root mean squared error (RMSE) and the average absolute percentage deviation (AAPD).

According to the main purpose of this research, it is possible to locate in bibliography artificial neural networks, support vector machines and random forest models focused on different fields on the wine`s world. It is possible to find research papers about neural models to verify the wine origin (Aires-De-Sousa, 1996), to classify Slovak white wines from different producers, varieties and production year (Kruzlicova, Mocak, Balla, Petka, Farkova and Havel, 2009) or to geographical classification (Šelih, Šala and Drgan, 2014,da Costa, Llobodanin, de Lima, Castro and Barbosa, 2018), among others. On the other hand, SVM has been used to classify Syrah wines according to their origin (Mendoza –Argentina- and Central Valley -Chile-) and them compared with neural networks (da Costa, Llobodanin, de Lima, Castro and Barbosa, 2018), to authenticate wines from South Africa, Hungary, Romania and Czech Republic with efficiency (Capron, Massart and Smeyers-Verbeke, 2007), to characterize and authenticate different Spanish PDO wine vinegars

(vinagre de Jerez, vinagre de Montilla-Moriles and vinagre del Condado de Huelva) (Ríos-Reina, Elcoroaristizabal, Ocaña-González, García-González, Amigo and Callejón, 2017), to predict enological parameters and discrimination of rice wine age (Yu, H., Lin, Xu, Ying, Li and Pan, 2008) or to predict wine's grade (Chen, Tawiah, Palmer and Erol, 2018),*inter alia*. Finally, random forest have been used to classify wines according to their production regions using trace elements (Tian et al., 2017), to model the impact of climate change on the wine regions from Hungary (Gaál, Moriondo and Bindi, 2012) or in different European wine regions (Moriondo et al., 2013) and to classify the cultivars on the basis of different chemical present in wine (Ahammed and Abedin, 2018), among others

The first developed model was an ANN model. To obtain the best ANN model is necessary to develop different ANN topologies with many configuration options selected by a trial and error procedure (Dawson and Wilby, 2001, Iglesias-Otero, Fernández-González, Rodríguez-Caride, Astray, Mejuto and Rodríguez-Rajo, 2015). The ANN model's topology is composed by different kind of layers: i) a first layer (called input layer) destined to introduce the experimental data in the network, ii) after this first layer there is/are another kind of layers (called hidden or intermediate layers) and finally, iii) a last layer (output layer) where the predicted value is generated (Figure 1).

FIGURE 1

During the ANN training phase, the value connection between neurons (called weights) is adjusted to achieve the minimum error between the experimental and the predicted output (Dai, Shi, Li, Ouyang and Huo, 2009). This process occurs in the hidden layers and output layer, and allows the neural network to learn based on training experimental cases. Trial and error approach was used to find the best neural model. Different topologies and training cycles were used to provide the best results according to statistics in the validation phase.

In this research, two types of ANN have been analysed. The first network,ANN_1, with backpropagation algorithm, sigmoidal function in its intermediate neurons and a linear function in the output neuron, and a second type, ANN_2, also with backpropagation algorithm and sigmoidal function in all intermediate and output neurons.

A disadvantage of neural models based on back-propagation algorithm, is that consume a huge computational time to optimize the different parameters which constitute the neural model (da Costa, Llobodanin, de Lima, Castro and Barbosa, 2018,Huang, Zhu and Siew, 2006). Due to this,

other techniques, SVM and RF have been studied due to these techniques require less computational cost and time of execution.

SVM is a powerful technique for classification and regression (RapidMiner GmbH, 2018), in our case, it was used to regression tasks using C-SVC and nu-SVC SVM types (RapidMiner GmbH, 2018). The SVM model finds an optimum separating hyperplane to maximize the borderline of the decision surface (da Costa, Llobodanin, de Lima, Castro and Barbosa, 2018). In this study, the LIBSVM learner by Chang and Lin (RapidMiner GmbH, 2018, Chang and Lin, 2018) was used. SVM model used the RBF kernel and the configuration of parameters, gamma and C, were studied according to the range proposed by the updated guide provide by Hsu *et al.* (2003).

In random forest regression model, three parameters were optimized: i) the number of trees (1 to 100 in twenty linear steps), ii) the least square criterion, iii) maximal depth (-1 to 10 in eleven linear steps), and iv) apply pre-pruning (true or false).

Neural models have been implemented in an AMD Ryzen 7 1800X Eight-Core Processor 3.60 GHz with 16 GB of RAM memory. ANN_1, SVM and RF models were developed using RapidMiner Studio Educational License and RapidMiner Studio Trial License from RapidMiner Inc. Neural models ANN_2 were developed using the EasyNN plus v14.0d software from Neural Planner Software Ltd. Data were fitted using Microsoft Excel from Microsoft Office Professional Plus 2013. Figures were drawn with Microsoft PowerPoint from Microsoft Office Professional Plus 2013 and Sigmaplot 13 from Systat Software Inc.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 January 2019 doi:10.20944/preprints201901.0110.v1

Peer-reviewed version available at Molecules 2019, 24, 826; doi:10.3390/molecules24050826

3. Results and discussion

Numerous ANN models (ANN_1 and ANN_2) were developed using trial and error method to find the best neural model topology. Over seven thousand neural network models with different topologies and training cycles were developed (varying the number of intermediate neurons between one and $2n+1$, where n is the number of input variables used). The best neural model was chosen based on its validation performance, and then, the best models were rechecked with the querying data group.

Table 1 shows the adjustments for the best ANN_1 model selected. It can be observed that neural model implemented with linear function in the output layer presents a good determination coefficient in all phases (between 0.998 for the training phase and 0.989 for querying phase). For the training phase, the error is below 10% (an acceptable error for this type of variable -aging time-). Similar behaviour is observed in the validation phase. In both phases, the root mean squared error in under 0.29 months. In querying phase, the ANN_1 model presents a good R^2 (0.989), nevertheless, a slight worsening is observed in the prediction in terms of RMSE (0.40 months) and AAPD (13.51%).

TABLE 1

Figure 2 shows the real value of aging time (orange) and the values predicted by the best ANN_1 model (dark blue) developed in this research. It can be observed in the validation cases that the ANN_1 model overestimates the real value (cases 1 and 2) while for cases 4, 6 and 8 the overestimation is very slight (between 1.28% and 4.46%). Cases 1 and 2 present a high error, in fact, the real value is 1 and the values predicted were 1.31 and 1.41, respectively. For the rest of the validation cases, the estimates are slightly lower than the real value (between -0.81% and -2.84%). For query cases, it can be seen how the linear ANN model presents overestimation of the aging time value in nine of the twelve cases reserved (especially in cases 1, 3 and 4). Once again the cases with real aging time of 1 month were the cases with bigger errors. Cases 1 and 3 present an individual percentage deviation of 68.70% and 37.25%. This is the reason for the increase of RMSE and the AAPD values in the querying phase. This behaviour is also observed in the training phase, where cases with one year of aging show greater errors (between -0.91% and 69.43%) than the rest of the cases. In view of these results, it can be concluded that the ANN_1 model presents a good general performance in all its phases, nevertheless, for low aging times, the model does not work at all well.

The next model implemented is the ANN model (ANN_2) with logistic function in its output neuron. As can be seen in Table 1, the adjustment parameters improve the fits of the ANN_1 model. It can be seen that for the training and the validation phase, the model presents coefficients of determination of one, improving the R^2 of the ANN_1 model. It is also clear that the ANN_2 model improves the adjustments in terms of RMSE and AAPD, going from an RMSE of 0.20 months to 0.04 months, for the validation phase of the ANN_1 and ANN_2 model, respectively. This good behaviour is also observed in the querying phase where the ANN_2 model presents a good determination coefficient, which corresponds with a low value of root mean squared error (0.03 months) and an average absolute percentage deviation below 0.85%.

In Figure 2, it can be seen the real value of aging time (orange) and the values predicted by ANN_2 model (brown). In validation cases, the ANN_2 model predicts with accuracy the real value of aging time. This behaviour is also observed for the query cases, it can be seen how the logistic ANN model presents a good prediction of the aging time value for all cases which makes the adjustments of this phase are good (0.03 months of RMSE and a 0.84% of APPD. Contrary to the previous ANN model, in this model no high errors are observed in any of the aging periods studied, in fact, errors remain between -1.63% (case 5 in querying phase) and 3.99% (case 2 in validation phase). With these results, it can be said that ANN_1 can predict with accuracy the aging time of red wines from D.O. Toro (Spain).

FIGURE 2

A new model based on support vector machine model was developed using library LIBSVM by Chang and Lin (RapidMiner GmbH, 2018, Chang and Lin, 2018). Gamma and C values were studied using trial and error method to find the best combination according to the range proposed by the updated guide provided by Hsu et al. (2003) (Hsu, Chang and Lin, 2003).

In Table 1, it can be seen the adjustments for the selected SVM model. It can be observed that the model presents a good determination coefficient in the training phase (0.995) with a low APPD, around 6.72% and with only an RMSE of 0.24 months. For the validation phase, it can be seen how the value of the determination coefficient falls slightly to 0.973 and the average absolute percentage deviation grows until to 12.86% that corresponds with a root mean squared error of 0.56 months. This high AAPD in the validation phase is due to the case number 2 in which the model predicts an aging value of 1.85 when the real value is 1 month, that is, the model predicts this case with 85.12% of individual percentage deviation (see Figure 2 top). This high error affects, significantly, the

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 January 2019 doi:10.20944/preprints201901.0110.v1

Peer-reviewed version available at Molecules 2019, 24, 826; doi:10.3390/molecules24050826

model´s AAPD value (12.86%, see Table 1). Other two cases, 9 and 11, present an error close to the one considered as acceptable (10%), -9.64 % and -10.73%, respectively.

The same behaviour can be seen in the querying phase. In this case, the R^2 increases to 0.988 and the RMSE decreases to 0.37 months. Nevertheless, the APPD increases up to 16.35%, this huge value is due, once again, the prediction for cases with one month aging time (see Figure 2 bottom). For cases 1 and 3, the individual percentage deviations are around 46.07% and an 80.85%, respectively, the individual percentage deviation for case 2 is -36.03%. These high values distort the value of the APPD in the querying phase. In view of these results, it can be concluded that the SVM model presents bad results for low aging times.

Finally, the last model implemented in this research is a model of random forest. According to the parameters exposed above, the best random forest model is an RF model with only one tree that provides the results shown in Table 1. It can be observed that the RF model presents an optimum determination coefficient which causes that the other of analysed parameters, RMSE and APPD, are zero (Table 1). The random forest model could find that the variables that dominate the determination of aging time are: the total-SO_2 (T-SO_2), the alcoholic grade (AD) and the free-SO_2 (F-SO_2). A random forest with only one tree, and with these three parameters, is enough to predict with total accuracy all cases of the training, validation and querying phase (Figure 2). These results show that the implemented RF model can predict with accuracy the aging time.

It seems clear that the adjustments obtained for the ANN_1 and SVM models are not good when the wines with one month of aging come into play. For the rest of aging times, both models work reasonably well. The results obtained for ANN_2 (developed with thirteen input variables) and RF model (that used three input variables) make these two models usable to guarantee red wine aging authenticity from D.O. Toro. These two models are able to predict, with accuracy, the aging time with, in worst case scenario (ANN_2), an average absolute percentage deviation below 1% which corresponds to a maximum error of 0.04 months (in terms of RMSE). This results improve the principal component analysis (PCA) model developed by Apetrei et al. (2012) using the oenological parameters where the analysis can describe a 61% (28% for the first principal component, of the information; 21% for the second and a 12% for the third) (Apetrei, Rodríguez-Méndez, Apetrei, Nevares, del Alamo and de Saja, 2012). The partial least squares-discriminant analysis (PLS-DA) using the physicochemical analyses only can explain a 5º9% of the variance in calibration and 77% in prediction presenting an RMSE up to 0.347 (Apetrei, Rodríguez-Méndez, Apetrei, Nevares, del Alamo and de Saja, 2012).

Regarding the RF model and to our understanding, a single tree in the random forest model seems to indicate that the wines of the Toro designation of origin, studied in this research, show particular characteristics that can be a key factor to predict aging time. In addition to this, it is expected that the inclusion of new experimental data from different wines could lead to the development of an RF model with more trees.

4. Conclusions

In this study, different models were developed to monitor red wines from D.O. Toro (Spain). The results obtained for ANN model developed with sigmoidal function in the output neuron and the random forest model, which used physical-chemical parameters, permit to determine the aging time, with an average absolute percentage deviation below 1%. In view of the results obtained by the models, ANN_1 and SVM, it would be advisable to continue with the analysis of the wines of the D.O. Toro and, even, to incorporate wines from the close appellations of origin.

5. Acknowledgement

Astray G. thanks to Xunta de Galicia, Consellería de Cultura, Educación e Ordenación Universitaria, for his postdoctoral grant B, POS-B/2016/001, K645 P.P.0000 421S 140.08. This work received financial support from Programa de Cooperación Interreg V-A España – Portugal (POCTEP) 2014-2020 (project Ref: 0377_IBERPHENOL_6_E).

6. References

Ahammed, B., & M. M. Abedin (2018). Predicting wine types with different classification techniques. *Model Assisted Statistics and Applications,* 13(1), 85-93.

Ahmad, S., & M. M. Gromiha (2003). Design and training of a neural network for predicting the solvent accessibility of proteins. *Journal of Computational Chemistry,* 24(11), 1313-1320.

Aires-De-Sousa, J. (1996). Verifying wine origin: A neural network approach. *American Journal of Enology and Viticulture,* 47(4), 410-414.

Akintunde, A. M., S. O. Ajala, & E. Betiku (2015). Optimization of Bauhinia monandra seed oil extraction via artificial neural network and response surface methodology: A potential biofuel candidate. *Industrial Crops and Products,* 67 387-394.

Alhaj, M. A. M., & Maghari, A. Y. A. (2017). Cancer survivability prediction using random forest and rule induction algorithms. In *ICIT 2017 - 8th International Conference on Information Technology, Proceedings* (388-391).

Apetrei, I. M., M. L. Rodríguez-Méndez, C. Apetrei, I. Nevares, M. del Alamo, & J. A. de Saja (2012). Monitoring of evolution during red wine aging in oak barrels and alternative method by means of an electronic panel test. *Food Research International,* 45(1), 244-249.

Beck, H. E., A. I. J. M. Van Dijk, D. G. Miralles, R. A. M. De Jeu, L. A. Bruijnzeel, T. R. McVicar, & J. Schellekens (2013). Global patterns in base flow index and recession based on streamflow observations from 3394 catchments. *Water Resources Research,* 49(12), 7843-7863.

Bishop, C. M. (1995). *Neural Networks for Pattern Recognition.* Oxford, U.K.: Clarendon Press.

Boser, B. E., Guyon, I. M., & Vapnik, V. N. (1992). A Training Algorithm for Optimal Margin Classifiers. In *Proceedings of the 5th Annual Workshop on Computational Learning Theory (COLT'92), 27-29 July 1992, 144-152.* Pittsburgh: .

Breiman, L., Friedman, J., Olshen, R., & Stone, C. (1984). *Classification and Regression Tree.* CRC Press.

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

Bucci, R., A. D. Magrí, A. L. Magrí, D. Marini, & F. Marini (2002). Chemical authentication of extra virgin olive oil varieties by supervised chemometric procedures. *Journal of Agricultural and Food Chemistry,* 50(3), 413-418.

Capron, X., D. L. Massart, & J. Smeyers-Verbeke (2007). Multivariate authentication of the geographical origin of wines: A kernel SVM approach. *European Food Research and Technology,* 225(3-4), 559-568.

Chan, K., T. -. Lee, P. A. Sample, M. H. Goldbaum, R. N. Weinreb, & T. J. Sejnowski (2002). Comparison of machine learning and traditional classifiers in glaucoma diagnosis. *IEEE Transactions on Biomedical Engineering,* 49(9), 963-974.

Chang, C. C., & C. J. Lin (2018). LIBSVM: A Library for support vector machines. *ACM Transactions on Intelligent Systems and Technology,* 2018(12/28),.

Chen, B., Tawiah, C., Palmer, J., & Erol, R. (2018). Multi-class wine grades predictions with hierarchical support vector machines. In *ICNC-FSKD 2017 - 13th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery* (111-115).

Chiang, Y., & F. Chang (2009). Integrating hydrometeorological information for rainfall-runoff modelling by artificial neural networks. *Hydrological Processes,* 23(11), 1650-1659.

Cutler, D. R., T. C. Edwards Jr., K. H. Beard, A. Cutler, K. T. Hess, J. Gibson, & J. J. Lawler (2007). Random forests for classification in ecology. *Ecology,* 88(11), 2783-2792.

da Costa, N. L., L. A. G. Llobodanin, M. D. de Lima, I. A. Castro, & R. Barbosa (2018). Geographical recognition of Syrah wines by combining feature selection with Extreme Learning Machine. *Measurement,* 120 92-99.

Dai, X., H. Shi, Y. Li, Z. Ouyang, & Z. Huo (2009). Artificial neural network models for estimating regional reference evapotranspiration based on climate factors. *Hydrological Processes,* 23(3), 442-450.

Danezis, G. P., A. S. Tsagkaris, F. Camin, V. Brusic, & C. A. Georgiou (2016). Food authentication: Techniques, trends & emerging approaches. *TrAC - Trends in Analytical Chemistry,* 85 123-132.

Dawson, C. W., & R. L. Wilby (2001). Hydrological modelling using artificial neural networks. *Progress in Physical Geography,* 25(1), 80-108.

European Commission (2018). Quality schemes explained. 2018(09/18),.

Gaál, M., M. Moriondo, & M. Bindi (2012). Modelling the impact of climate change on the Hungarian wine regions using Random Forest. *Applied Ecology and Environmental Research,* 10(2), 121-140.

Gazzaz, N. M., M. K. Yusoff, A. Z. Aris, H. Juahir, & M. F. Ramli (2012). Artificial neural network modeling of the water quality index for Kinta River (Malaysia) using water quality variables as predictors. *Marine pollution bulletin,* 64(11), 2409-2420.

Gonzalez-Fernandez, I., M. Iglesias-Otero, M. Esteki, O. A. Moldes, J. C. Mejuto, & J. Simal-Gandara (2018). A critical review on the use of artificial neural networks in olive oil production, characterization and authentication. *Critical reviews in food science and nutrition,* 1-14.

Haykin, S. (1999). *Neural networks, a comprehensive foundation.* Pearson Prentice-Hall.

Hernández Suárez M., Astray Dopazo G., Larios López D., & Espinosa F. (2015). Identification of relevant phytochemical constituents for characterization and authentication of tomatoes by general linear model linked to automatic interaction detection (GLM-AID) and artificial neural network models (ANNs). *PLoS ONE,* 10(6), e0128566.

Hsu, C. W., C. C. Chang, & C. J. Lin (2003). A Practical Guide to Support Vector Classification, https://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf. 1-16.

Hu, L., C. Yin, S. Ma, & Z. Liu (2018). Rapid detection of three quality parameters and classification of wine based on Vis-NIR spectroscopy with wavelength selection by ACO and CARS algorithms. *Spectrochimica Acta - Part A: Molecular and Biomolecular Spectroscopy,* 205 574-581.

Huang, G., Q. Zhu, & C. Siew (2006). Extreme learning machine: Theory and applications. *Neurocomputing,* 70(1), 489-501.

Iglesias-Otero, M. A., M. Fernández-González, D. Rodríguez-Caride, G. Astray, J. C. Mejuto, & F. J. Rodríguez-Rajo (2015). A model to forecast the risk periods of Plantago pollen allergy by using the ANN methodology. *Aerobiologia,* 31(2), 201-211.

International Organisation of Vine and Wine, OIV (1990). Recueil des méthodes internationales d'analyse des vins et des mouts.

Kamari, A., A. H. Mohammadi, A. Bahadori, & S. Zendehboudi (2014). Prediction of air specific heat ratios at elevated pressures using a novel modeling approach. *Chemical Engineering and Technology,* 37(12), 2047-2055.

Kruzlicova, D., J. Mocak, B. Balla, J. Petka, M. Farkova, & J. Havel (2009). Classification of Slovak white wines using artificial neural networks and discriminant techniques. *Food Chemistry,* 112(4), 1046-1052.

Luykx, D. M. A. M., & S. M. van Ruth (2008). An overview of analytical methods for determining the geographical origin of food products. *Food Chemistry,* 107(2), 897-911.

McCulloch, W. S., & W. Pitts (1943). A logical calculus of the ideas immanent in nervous activity. *The Bulletin of mathematical biophysics,* 5(4), 115-133.

Moldes, O. A., J. C. Mejuto, R. Rial-Otero, & J. Simal-Gandara (2017). A Critical Review on the Applications of Artificial Neural Networks in Winemaking Technology. *Critical reviews in food science and nutrition,* 57(13), 2896-2908.

Montoya, L. A., O. A. Moldes, A. Cid, C. Astray, J. F. Gálvez, & J. C. Mejuto (2015). Influence prediction of alkylamines upon electrical percolation of AOT-based microemulsions using artificial neural networks. *Tenside, Surfactants, Detergents,* 52(6), 473-476.

Moriondo, M., G. V. Jones, B. Bois, C. Dibari, R. Ferrise, G. Trombi, & M. Bindi (2013). Projected shifts of wine regions in response to climate change. *Climatic Change,* 119(3-4), 825-839.

Mutanga, O., E. Adam, & M. A. Cho (2012). High density biomass estimation for wetland vegetation using WorldView-2 imagery and random forest regression algorithm. *International Journal of Applied Earth Observation and Geoinformation,* 18 399-406.

Narisetty, V., G. Astray, B. Gullón, E. Castro, B. Parameswaran, & A. Pandey (2017). Improved 1,3-propanediol production with maintained physical conditions and optimized media composition: Validation with statistical and neural approach. *Biochemical engineering journal,* 126 109-117.

Pai, P., & W. Hong (2005). Forecasting regional electricity load based on recurrent support vector machines with genetic algorithms. *Electric Power Systems Research,* 74(3), 417-425.

Rapeanu, G., C. Vicol, & C. Bichescu (2009). Possibilities to asses the wines authenticity. *Innovative Romanian Food Biotechnology,* 5 1-9.

RapidMiner GmbH (2018). RapidMiner Documentation. 2018.

Ríos-Reina, R., S. Elcoroaristizabal, J. A. Ocaña-González, D. L. García-González, J. M. Amigo, & R. M. Callejón (2017). Characterization and authentication of Spanish PDO wine vinegars using multidimensional fluorescence and chemometrics. *Food Chemistry,* 230 108-116.

Riovanto, R., W. U. Cynkar, P. Berzaghi, & D. Cozzolino (2011). Discrimination between Shiraz wines from different Australian regions: The role of spectroscopy and chemometrics. *Journal of Agricultural and Food Chemistry,* 59(18), 10356-10360.

Saurina, J. (2010). Characterization of wines using compositional profiles and chemometrics. *TrAC - Trends in Analytical Chemistry,* 29(3), 234-245.

Šelih, V. S., M. Šala, & V. Drgan (2014). Multi-element analysis of wines by ICP-MS and ICP-OES and their classification according to geographical origin in Slovenia. *Food Chemistry,* 153 414-423.

Serrano-Lourido, D., J. Saurina, S. Hernández-Cassou, & A. Checa (2012). Classification and characterisation of Spanish red wines according to their appellation of origin based on chromatographic profiles and chemometric data analysis. *Food Chemistry,* 135(3), 1425-1431.

Tian, Y., C. Yan, T. Zhang, H. Tang, H. Li, J. Yu, J. Bernard, L. Chen, S. Martin, N. Delepine-Gilon, J. Bocková, P. Veis, Y. Chen, & J. Yu (2017). Classification of wines according to their production regions with the contained trace elements using laser-induced breakdown spectroscopy. *Spectrochimica Acta Part B: Atomic Spectroscopy,* 135 91-101.

Vigneau, E., P. Courcoux, R. Symoneaux, L. Guérin, & A. Villière (2018). Random forests: A machine learning methodology to highlight the volatile organic compounds involved in olfactory perception. *Food Quality and Preference,* 68 135-145.

Yu, H., H. Lin, H. Xu, Y. Ying, B. Li, & X. Pan (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.

Yu, R., & M. Abdel-Aty (2013). Utilizing support vector machine in real-time crash risk evaluation. *Accident Analysis & Prevention*, 51 252-259.

Figure captions

Figure 1. Example of neural network topology 13-5-1 with 13 neurons in the input layer, five neurons in the intermediate layer and one in the output layer.

Figure 2. Bar graph for validation (top) and querying (bottom) cases according to the real value of aging time (orange) and the values predicted by the artificial neural network with linear function in output neuron (ANN_1, dark blue), artificial neural network with sigmoidal function in output neuron (ANN_2, brown), support vector machine (SVM, olive) and random forest (RF, light blue).

Figure 1

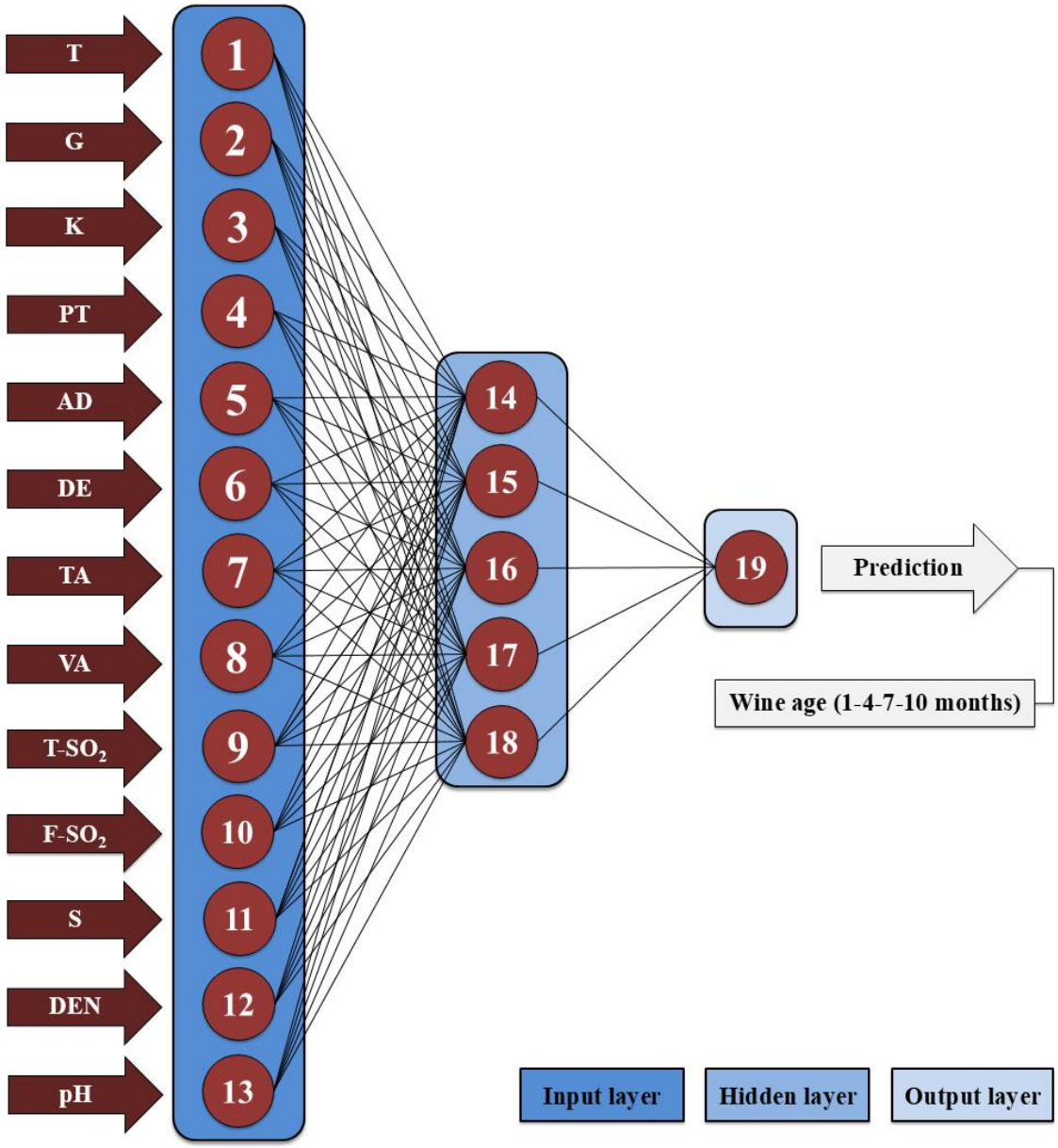

Figure 2

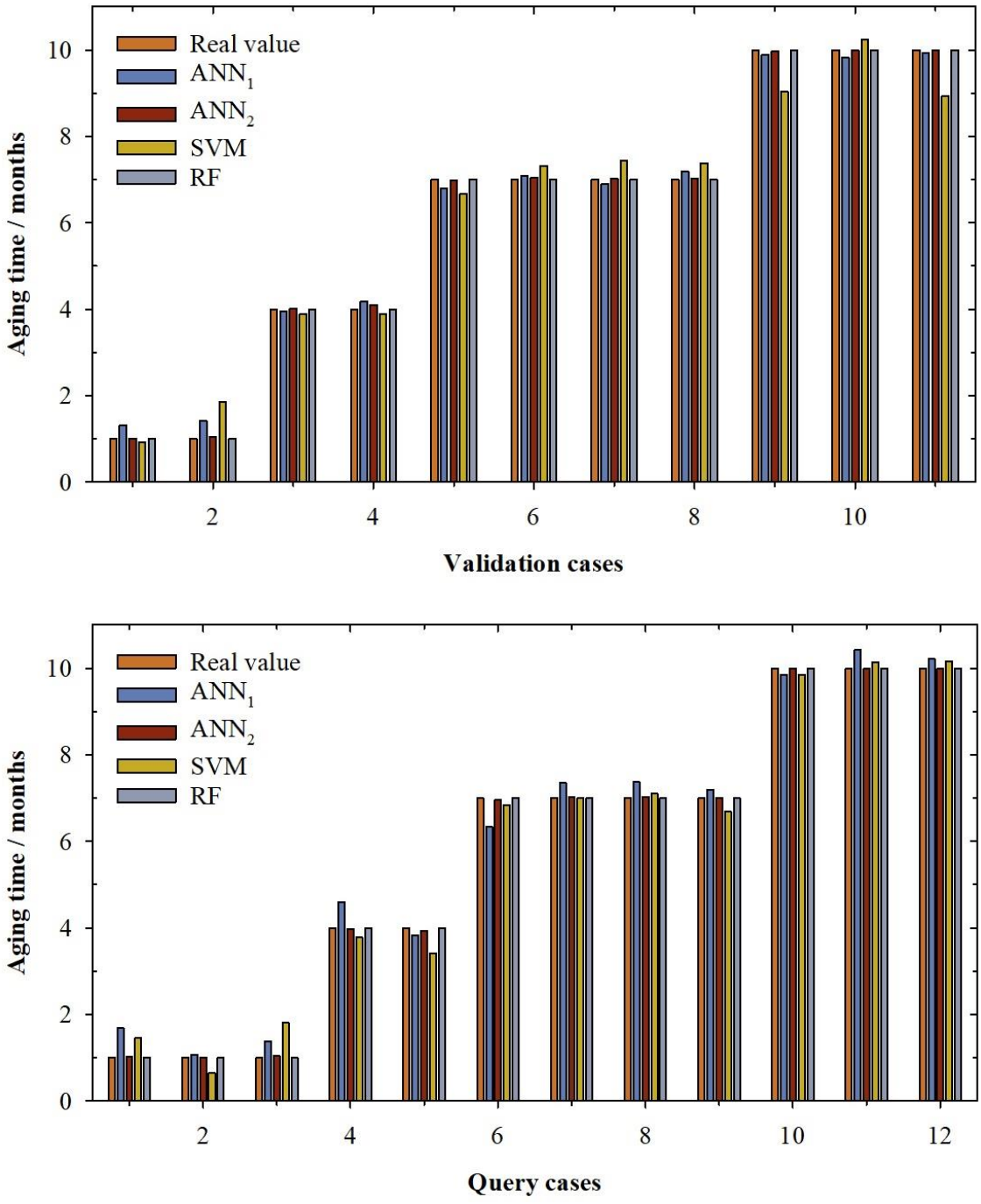

Table titles

Table 1. Coefficient of determination (R^2), root mean squared error (RMSE) and average absolute percentage deviation (AAPD) for training (T), validation (V) and querying (Q) phase, for each model present in this research, artificial neural network with linear function in output neuron (ANN_1), artificial neural network with sigmoidal function in output neuron (ANN_2), support vector machine (SVM) and random forest (RF).

Table 1

Table 1. Coefficient of determination (R^2), root mean squared error (RMSE) and average absolute percentage deviation (AAPD) for training (T), validation (V) and querying (Q) phase, for each model present in this research, artificial neural network with linear function in output neuron (ANN_1), artificial neural network with sigmoidal function in output neuron (ANN_2), support vector machine (SVM) and random forest (RF).

	Training			Validation			Querying		
Model	**R^2**	**RMSE**	**AAPD (%)**	**R^2**	**RMSE**	**AAPD (%)**	**R^2**	**RMSE**	**AAPD (%)**
ANN_1	0.994	0.28	8.07	0.998	0.20	8.20	0.989	0.40	13.51
ANN_2	1.000	0.02	0.42	1.000	0.04	0.87	1.000	0.03	0.84
SVM	0.995	0.24	6.72	0.973	0.56	12.86	0.988	0.37	16.35
RF	1.000	0.00	0.00	1.000	0.00	0.00	1.000	0.00	0.00