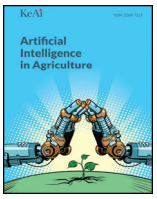
Artificial Intelligence in Agriculture 5 (2021) 157–162

Contents lists available at ScienceDirect

Artificial Intelligence in Agriculture

journal homepage: http://www.keaipublishing.com/en/journals/artificial-intelligence-in-agriculture/



Predictive modeling for wine authenticity using a machine learning approach

Nattane Luíza da Costa [a,b,*], Leonardo A. Valentin [c], Inar Alves Castro [c], Rommel Melgaço Barbosa [a]

[a] Instituto de Informática, Universidade Federal de Goiás, Goiânia-Go, Brazil
[b] Informatics Nucleo, Goiano Federal Institute of Education, Science and Technology, Campus Urutaí, Rodovia Geraldo Silva Nascimento, km 2.5, Zona Rural, Bloco de Informática, Urutaí, GO, Brazil
[c] LADAF, Department of Food and Experimental Nutrition, Faculty of Pharmaceutical Sciences, University of São Paulo, Av. Lineu Prestes, 580, B14, 05508-900 São Paulo, Brazil

ARTICLE INFO

Article history:
Received 3 November 2020
Received in revised form 1 July 2021
Accepted 8 July 2021
Available online 10 July 2021

Key-words:
Wine classification
Support vector machines
Feature selection
South American wines
Machine learning

ABSTRACT

The purpose of this paper is to classify wines from 4 different countries in South America. Each class of wines is formed by samples considered by experts as representatives of the following commercial categories: "Argentinean Malbec (AM)", "Brazilian Merlot (BM)", "Uruguayan Tannat (UT)" and "Chilean Carménère (CC)". The 83 samples collected were analyzed according to their composition of volatiles, semi-volatiles and phenolic compounds. We built a decision system for classification based on support vector machines (SVM), along with Correlation-based Feature selection (CFS), and Random Forest Importance (RFI), which measures the relative importance of the input variables. First, we use CFS to select a subset of variables among 190 chemical compounds. Thirteen chemicals were selected as correlated to the category and uncorrelated with each other. Afterwards, these chemical compounds were organized according to the importance ranking given by the RFI and classified with SVM. The study clearly indicated that SVM in combination with feature selection methods was able to identify the most important chemicals to classify the wine samples. Among the compounds identified in the wine samples, the variable subset defined by the feature selection methods, which were catechin, gallic, octanoic acid, myricetin, caffeic, isobutanol, resveratrol, kaempferol, and ORAC, were able to achieve an accuracy of 93.97% in classifying the commercial categories.

1. Introduction

The combination of instrumental analyses with machine learning have allowed the successful classification of various products and foods according to geographical origin, varieties, and the type of farming system (conventional or organic) (Jiménez-Carvelo et al., 2019; Versari et al., 2014). The main reason to discriminate food products is quality control and authentication, in order to avoid unfair competition that can create a destabilized market and disrupt the regional economy and even the national economy (Callao and Ruisánchez, 2018; Cordella et al., 2002).

Particularly, the wine classification according to variety and region is an important issue because it is an easily adulterated product in terms of dilution of wines with water, the addition of alcohol, mislabeling, and blending with, or replacement by, wine of a lesser quality of denomination. Analytical techniques such as High Performance Liquid Chromatogrphy (HPLC), coupled mass spectrometry detection (MS), Fourier Transform Infrared (FTIR), among others, are able to describe wines in chemicals compounds that can reveal useful information about the identity of each wine (Versari et al., 2014; Villano et al., 2017).

The world wine production reached in 2018 a volume of 292.3 million of hectoliters, being more than 30 million hectoliters only in Argentina, Chile, Brazil and Uruguay (Wine, 2019), representing an important sector of economy of these countries. South American wines have been the object of study of several research in order to classify the geographical origin and varietal. These studies have been performed mainly by the use of Principal Component Analysis (PCA) and Linear Discrimination Analysis (LDA) (Belmiro et al., 2017; Llobodanin et al., 2014; Pisano et al., 2015; Versari et al., 2014). The use of linear methods is enough to obtain satisfactory results and the methods are easy for chemists to understand. However, there are several physical-chemical parameters that characterize samples, resulting in complex data with some nonlinearity. Classical linear methods such as discriminant analysis cannot model this nonlinearity. This way, nonlinear methods are required (Jiménez-Carvelo et al., 2019; Li et al., 2009).

Support Vector Machines (SVM) is one of the most popular and sophisticated machine learning techniques. This technique separates the samples by an optimal hyperplane that maximizes the distances

* Corresponding author at: Instituto de Informática, Universidade Federal de Goiás, Goiânia-Go, Brazil.
E-mail addresses: nattane.luiza@ifgoiano.edu.br (N.L. da Costa), leonardo.valentin@usp.br (L.A. Valentin), inar@usp.br (I.A. Castro), rommel@inf.ufg.br (R.M. Barbosa).

https://doi.org/10.1016/j.aiia.2021.07.001

between classes. When there are no linear decision boundaries, the original dataset (x) is converted into a new space f(x) which there is a linear decision boundary that separates the samples into their classes. SVM has been successfully used in many different applications, such as: food science (Araújo et al., 2019; Richter et al., 2019; Soares et al., 2018; Turra et al., 2017), medicine (Froz et al., 2017; Vogado et al., 2018), forensic science (Maione et al., 2018), among others.

In previous studies, some *Vitis Vinífera* wines from South America were analyzed with machine learning using feature selection and support vector machines based on their antioxidant activity, phenolic substances, anthocyanins and color (Costa et al., 2018, 2019; da Costa et al., 2016). The use of feature selection allowed to identify the most important chemical descriptors that characterize the geographical origin of wines. The study of Soares (Soares et al., 2018) also classified South America wines based on elemental concentrations, machine learning and feature selection. The authors proposed a new filter feature selection method and achieved seven elements that classified the wines in 99.9% of accuracy.

Information like a reduced set of chemical descriptors that can classify food samples obtained from feature selection methods can reduce costs related to the chemical analysis of uncorrelated compounds, reduce the computational time required to analyze the data, can improve accuracy and provide useful information to prevent and to identify food fraud.

In this study, samples of South American wines were analyzed based on the phenolic, semi-volatile and volatile compounds. Few studies had used feature selection methods to classify wine samples based on phenolic and volatile compounds. The study of (Rebolo et al., 2000) used five chemical compounds selected based on Stepwise Bayesian Analysis algorithm to classify Galician red wines along with LDA, k-nearest neighbors and soft independent class modeling. Mihena et al. (Mihnea et al., 2015) classified different pre-fermentative maceration techniques of Mencía wines with forward stepwise linear discriminant analysis. This technique was also used to select the most important chemical parameters to classify Chinese rice wines (Shen et al., 2012), Andalusian sweet wines (Márquez et al., 2008), and the variety of Chinese red wines (Zhang et al., 2010).

The objective of this study is to characterize the main phenolic, semi-volatile and volatile compounds which are able to discriminate the four "Commercial Categories" of South American wines, named Argentinean Malbec (AM)", "Brazilian Merlot (BM)", "Uruguaian Tannat (UT)" and "Chilean Carménère (CC)", based on Support vector machines (SVM) combined with feature selection methods named Correlation Based Feature Selection and Random Forest Importance. To our knowledge, this is the first study to combine SVM and two feature selection methods to classify wine samples based on phenolic, semi-volatile and volatile compounds.

2. Material and methods

2.1. Material

A group of wine experts from the "Brazilian Association of Sommeliers (ABS-SP, São Paulo, Brazil)" was invited to list examples of wines that were the best representatives of four categories: "Argentinean Malbec (AM)", "Brazilian Merlot (BM)", "UruguaianTannat (UT)" and "Chilean Carménère (CC)". The ABS-SP group is formed by experts with elevated experience in the wine degustation, market and technology (www.abs-sp.com.br). The following question was placed to the experts: "If you had to show to someone what is an authentic, for example, Argentinean Malbec, which wine would you recommend?" To answer this question, the experts elaborated a list containing the brand and vintage of each category. From this list, 83 wines (2008–2015) were purchased from the local stores covering a price range from 12 to 135 USD for 750 mL bottle. The list of wines applied in this study included 14 UT, 26 MA, 33 CC and 10 BM. The bottle was opened, the wine was immediately aliquoted and kept at -80 °C until the analysis. The reagents 2,2-diphenyl-1-picrylhydrazil (DPPH), fluorescein, 2,2′-Azobis (2-methylpropionamidine) dihydrochloride (AAPH), epicatechin, caffeic acid, gallic acid, ferulic acid, rutin, trans-resveratrol, myricetin, quercetin, kaempferol were purchased from Sigma-Aldrich (Sigma Chemical Co, St. Louis, United States). The anthocyanins cyanidin-3-glucoside, delphinidin-3-glucoside, peonidin-3-glucoside, petunidin-3-glucoside and malvidin-3-glucoside were obtained from Polyphenols Laboratories AS (Polyphenols SA, Sandnes, Norway). All other chemical and solvents were analytical grade.

2.2. Methods

The semi-volatile compounds were analyzed using Quadrupole Time-of-Flight (QTOF) Liquid Chromatography–Mass Spectrometry according to 5991-3335EN Application Note (Agilent Technologies, Inc., Palo Alto, USA). Volatiles and phenolic compounds were analyzed according to Cabredo-Pinillos et al. (2004) and Jaitz et al. (2010), respectively. The antioxidant activity of the samples was analyzed by three methods. ORAC was carried out according to Huang et al. (2005). DPPH analysis was based on method reported by Brand-Williams et al. (1995), while FRAP was performed according to Benzie and Strain (1996). The distribution of the compounds found is as follows: 21 phenolic components, 29 volatiles and 140 semi-volatiles. More information about the HPLC chromatograms is given by Valentin et al. (2020).

2.2.1. Support vector machines

Support Vector Machines (Cortes and Vapnik, 1995) is one of the best known and most used classification algorithms which aim to find a hyperplane with maximum margin to separate the classes of data. This hyperplane is a special type of linear model. In the linear case a database with two class of data is easily separated by a hyperplane in two groups, which one represents one of the classes and the other group represents the other class. The samples that are closest to the maximum-margin hyperplane are called support vectors. The non-linear cases begin to be solved in 1995 when Cortes & Vapnik introduced the use of the soft-margin, which uses slack variables to penalize samples that characterize the non-linearity of data. When there are no linear decision boundaries the original dataset (x) is converted into a new space f(x) by the kernel trick. In the new space f(x) there is a linear decision boundary that separates the samples into their classes.

2.2.2. Feature selection

The feature selection (variable elimination) helps in understanding data, reducing the effect of high dimensionality, reducing computational time and improving the classifier performance (Chandrashekar and Sahin, 2014). We use Correlation-based feature selection (CFS) to select a subset of the most important variable and Random Forest Importance (RFI) to order the variables selected. The CFS selected a feature subset that contains features highly correlated with the class, yet uncorrelated with each other (Hall, 1999). The Random Forest Importance (RFI) uses the random forest classifier to measure the importance of input variables (Breiman, 2001). This technique is a filter feature selector, which uses variable ranking as the principle criteria for variable selection by ordering.

2.2.3. Performance analysis

The classifications occurred by the use of 10-fold cross validation to evaluating the performance of SVM. This technique splits the dataset (in our case, 83 samples) into 10 folds and performs 10 classifications. Each classification is accomplished in the following way: nine folds are used to train the model and the tenth fold is used to test the model, generating a precision rate (accuracy). This process is performed until all ten folds are used for testing. At the end of the process, the performance of the classification model is given as the average of the accuracy

obtained in each partition. All machine learning tasks and graphs were performed using the R software (R Core Team, 2017).

3. Results and discussion

Fig. 1 shows the steps carried out in this study. After select the wines, they were analyzed according to their composition of phenolic, volatiles and semi-volatiles compounds. Next, we applied the machine learning techniques to investigate the distinction of the commercial categories.

About 600 peaks were isolated to volatile compounds, semi-volatile compounds by negative mode and semi-volatile compounds by positive mode compounds. From them, just the peaks that it was possible to identify the formula were selected (169 peaks). The dataset used in this study has 83 samples described in 169 volatile and semi-volatile compounds peaks besides 21 variables relative to antioxidant activity and phenolic compounds associated to one of the classes (AM, BM, CC, UT). The number of samples of each class was distributed as follows: 26 Argentinean Malbec samples, 10 Brazilian Merlot samples, 33 Chilean Carménère samples, and 14 Uruguayan Tannat samples. The class labels (the chemical categories) were selected by a panel of wine experts on base on their sensory properties. This aspect is essential because researchers, usually, do not have this information, and proportion of market share can be more related to price conditions than wines authenticity.

3.1. Classification and analysis of all dataset

The classification algorithm SVM was applied considering the original 190 variables first (169 volatile and semi-volatile compounds peaks and 21 variables relative to antioxidant activity and phenolic compounds). Results achieved in this step are considered as "references" to which any others can be compared. The SVM model obtained 39.76% of accuracy by the use of all 190 variables. All the samples were classified as "Chilean Carménère". The accuracy value obtained in this model is explained by the number of CC samples. There are 33 CC samples in the database and all samples were classified as CC. Only the 33 CC samples were correctly classified, which represent 39.76% of the total samples.

An exploratory analysis by means of principal component analysis (PCA) was carried out to investigate this result. This technique is based on a linear transformation of data into different orthogonal coordinates using maximum variance and minimum correlation. By reducing the number of features from 190 original variables to two principal components, the information is enough to allow an examination of the samples according to the commercial category in a 2D-plot (Fig. 2). There is no natural separation among the classes. Three clusters can be identified, however, each one is formed with samples of all classes.

The first conclusion is that all phenolic and volatile compounds cannot discriminate among the commercial categories. The low classification

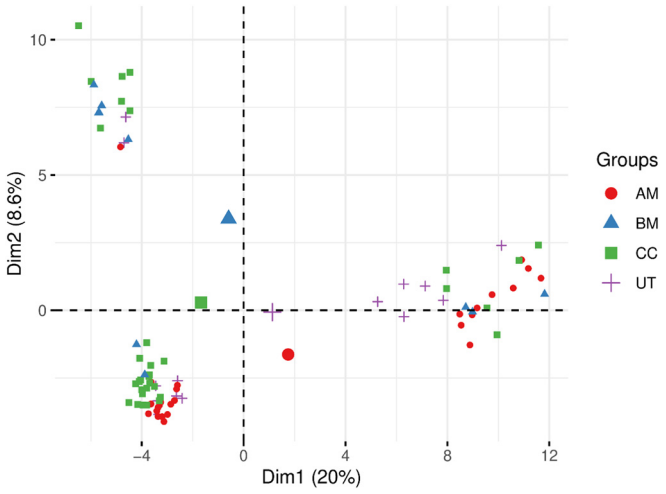


Fig. 2. PCA score plot of all dataset.

rate can be explained considering the fact that is possible to exists noise or irrelevant variables in our dataset. Therefore, we applied feature selection methods to select a subset of variables from the input which can efficiently describe the input data and discriminate among the classes.

3.2. Feature selection

CFS were applied to all dataset and returned 13 variables. They are: ORAC, octanoic acid, isobutanol, 2-phenyletanol, isorhamnetin, ethyl lactate, sv(+) 16 - not identified semi-volatile, gallic, catechin, caffeic, myricetin, resveratrol and kaempferol. Table 1 shows the mean and the standard deviation of the chemical's concentration for each commercial category along with the results of Kruskal Wallis test. The variables ORAC, 2-phenyletanol and gallic present a very close concentration values for the commercial categories AM, BM, and CC, and a higher concentration for the UT (p value <0.05). The UT concentrations for these variables presented a significant difference. The variables octanoic acid and isobutanol showed highest values for the BM and CC categories (p value <0.05) and AM and UT were not significant different from each other. Ethyl lactate and resveratrol presented a highest concentration values for the BM wines (p value <0.05). There is no pattern among the statistical differences in the variable's concentrations among the commercial categories. An advanced analysis is need to differentiate the categories based on this chemical composition.

A PCA analysis was carried out on the selected variables. Fig. 3 shows the score plot of the first two principal components. A natural

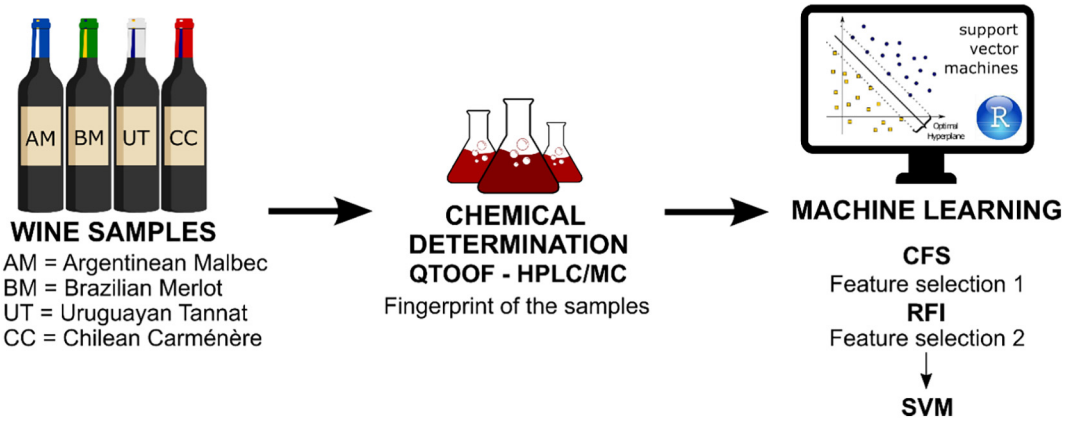


Fig. 1. Methodology applied on South America wines.

Table 1
Descriptive statistical analysis of South American commercial Categories wine data.

Variable	AM	BM	UT	CC
ORAC (mMol TE)	21.31 ± 1.13^{ut}	21.82 ± 0.99^{ut}	30.18 ± 1.53^{am,cc,bm}	21.33 ± 0.66^{ut}
octanoic acid (area)	27,478,983 ± 5 517 205$^{cc,\ bm}$	48,950,606 ± 4 335 284$^{ut,\ am,\ cc}$	15,443,778 ± 3 057 059$^{cc,\ bm}$	63,179,169 ± 2 182 888$^{ut,\ am,\ bm}$
Isobutanol (area)	7,046,394 ± 2 077 253$^{cc,\ bm}$	24,145,070 ± 5 148 480$^{ut,\ am}$	3,803,071 ± 523 621$^{cc,\ bm}$	26,713,251 ± 2 124 151$^{ut,\ am}$
2-phenyletanol (area)	749,658 ± 234 101^{ut}	407,573 ± 73 070^{ut}	2,953,941 ± 687 182$^{am,\ cc,\ bm}$	624,054 ± 105 318^{ut}
Isorhamnetin (area)	929,142 ± 290 665^{cc}	1,021,545 ± 602,395	279,789 ± 81,894	410,724 ± 173,173 am
Ethyl lactate (area)	5,590,711 ± 1 407 323$^{ut,\ cc,\ bm}$	20,089,649 ± 3 469 183$^{ut,\ am,\ cc}$	11,677,131 ± 3 897 358$^{am,\ bm}$	10,422,179 ± 1 898 642$^{am,\ bm}$
sv(+) 16 - not identified semi-volatile (area)	10,158,094 ± 2,425,254	123,298,219 ± 47,604,613	7,899,224 ± 2,743,240	17,979,593 ± 7,955,919
Gallic (mg/L)	4.80 ± 0.018^{ut,cc}	3.62 ± 0.015^{ut,cc}	6.61 ± 0.025^{am,cc,bm}	3.62 ± 0.015^{ut,am,bm}
Catechin (mg/L)	2.15 ± 0.006$^{cc,\ bm}$	1.59 ± 0.006$^{ut,\ am,\ cc}$	2.04 ± 0.006^{cc,bm}	1.04 ± 0.005$^{ut,\ am,\ bm}$
Caffeic (mg/L)	0.37 ± 0.0007 $^{ut,\ bm}$	0.58 ± 0.001$^{am,\ cc}$	0.84 ± 0.002$^{am,\ cc}$	0.36 ± 0.003$^{ut,\ bm}$
Myrecetin (mg/L)	1.80 ± 0.009$^{ut,\ cc,\ bm}$	1.12 ± 0.004$^{am,\ cc}$	1.2 ± 0.008^{am}	0.93 ± 0.006$^{am,\ bm}$
Resveratrol (mg/L)	0.17 ± 0.001$^{cc,\ bm}$	0.34 ± 0.001$^{ut,\ am,\ cc}$	0.12 ± 0.001^{bm}	0.08 ± 0.003$^{am,\ bm}$
Kaempferol (mg/L)	0.03 ± 0.002^{cc}	0.08 ± 0.002^{cc}	0.02 ± 0.004^{cc}	0.17 ± 0.007$^{ut,\ am,\ bm}$

^{am} : Statistical difference from AM group (p value <0.05).
^{bm} : Statistical difference from BM group (p value <0.05).
^{cc} : Statistical difference from CC group (p value <0.05).
^{ut} : Statistical difference from UT group (p value <0.05).

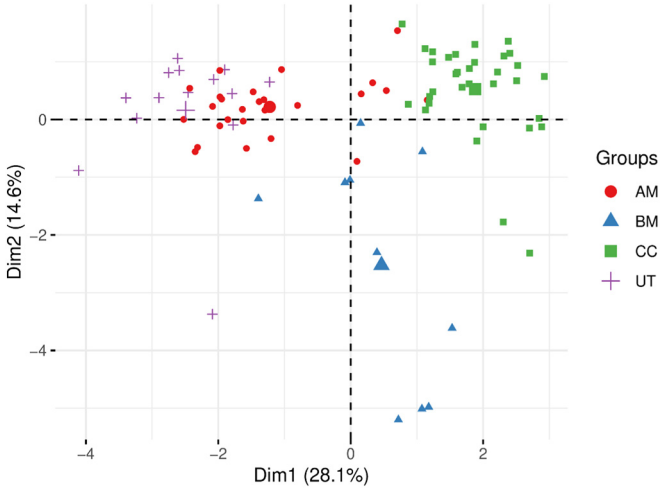


Fig. 3. PCA score plot of the variables selected by the CFS.

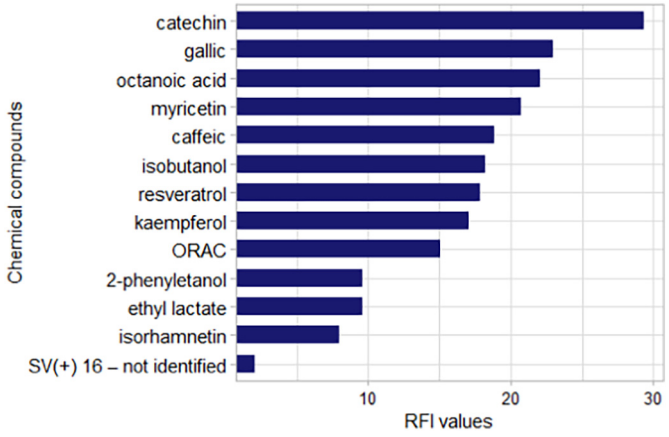


Fig. 4. Importance ranking of chemical compounds selected by CFS according to RFI.

separation between BM and CC can be observed. The samples of AM and UT are separated from BM and CC, but there is an overlap between these classes. The variables selected by CFS showed more discriminative power than all phenolic and volatile compounds when analyzing the distance among the samples.

The variables selected by CFS were submitted to a filter method of feature selection, the RFI. This method ranks the features according to the classifications performed by the Random Forest classifier by the use of one variable at time as predictor. The feature selection performed by the CFS aimed to reduce the number of variables and the RFI goal was ordering the features wherein the variables were selected and applied to the predictor according to the importance of each one. Fig. 4 shows the importance of each variable according to RFI.

3.3. Classification results

Based on the information of the RFI, we performed 13 classification models as follows. The first classification model used only the most important variable (catechin), the second classification model used the two most important variables (catechin and gallic), and so forth, until the thirteenth classification model, which used the thirteen variables selected by CFS. Table 2 shows the results obtained from these thirteen classification models. It is presented the model ID, the variables used as input to SVM and the accuracy values for each classification model. The higher result is the one that used nine variables, catechin, gallic, octanoic acid, myricetin, caffeic, isobutanol, resveratrol, kaempferol and ORAC, with 93.97% of accuracy. Just one variable results in the lower accuracy, 68.49%. These value increases when variables are added. The peak of accuracy occurs with the #09 classification model. This value decreases and stabilizes between a range of values when more variables are added.

Table 3 shows the confusion matrix of the classification model #09. All samples of CC class were correctly classified. One sample of BM class, two samples of AM class, and two samples of UT class were misclassified. These results suggest that the CC class is the purest among them, in which the nine chemical components of the subset #09 is able to correctly classify all the samples. The other data classes have few samples that were misclassified, indicating that there is a little similarity among the others commercial categories. However, the classification model achieved a good classification rate of 93.97%.

4. Discussion

Phenolic and volatile profiles are the most involved compounds in the sensory identity of wine (Villano et al., 2017). Thus, these compounds can be used on discriminatory analysis to classify wines.

Table 2
Results from classification with SVM associated to feature selection.

Model	Variables	SVM (%)
#01	Catechin	68.49
#02	catechin, gallic	80.16
#03	catechin, gallic, octanoic acid	81.59
#04	catechin, gallic, octanoic acid, myrecetin	85.69
#05	catechin, gallic, octanoic acid, myrecetin, caffeic	83.15
#06	catechin, gallic, octanoic acid, myrecetin, caffeic, isobutanol	84.40
#07	catechin, gallic, octanoic acid, myrecetin, caffeic, isobutanol, resveratrol	86.63
#08	catechin, gallic, octanoic acid, myrecetin, caffeic, isobutanol, resveratrol, kaempferol	88.99
#09	**catechin, gallic, octanoic acid, myrecetin, caffeic, isobutanol, resveratrol, kaempferol, orac**	**93.97**
#10	catechin, gallic, octanoic acid, myrecetin, caffeic, isobutanol, resveratrol, kaempferol, orac, 2-phenyletanol	90.52
#11	catechin, gallic, octanoic acid, myrecetin, caffeic, isobutanol, resveratrol, kaempferol, orac, 2-phenyletanol, ethyl lactate	90.52
#12	catechin, gallic, octanoic acid, myrecetin, caffeic, isobutanol, resveratrol, kaempferol, orac, 2-phenyletanol, ethyl lactate, isorhamnetin	88.43
#13	catechin, gallic, octanoic acid, myrecetin, caffeic, isobutanol, resveratrol, kaempferol, orac, 2-phenyletanol, ethyl lactate, isorhamnetin, sv(+) 16 - not identified semi-volatile	89.54

Table 3
Confusion matrix of #09 classification model.

		Prediction				
		AM	BM	CC	UT	Correct classifications
Reference	AM	24	–	1	1	92.30%
	BM	–	9	1	–	90.00%
	CC	–	–	33	–	100%
	UT	2	–	–	12	85.71%
			Total Accuracy:			93.97%

However, as the results of this study demonstrated, there is some variables that do not contributed to the classification model. The use of all chemicals resulted on a low predictive performance and there is no clear separation on the PCA score plot. The selection of a feature subset played an important role to classify the commercial categories of South American wines.

It was found that a group of phenolic and volatile was the more discriminative compounds. The model #09 used seven phenolic and two volatile compounds. The ORAC was previously identified by the use of feature selection as one of the main chemicals that are able to classify Chilean and Brazilian Sauvignon Cabernet wines (da Costa et al., 2016), Uruguayan and Brazilian Tannat wines (Costa et al., 2018), and South American Merlot wines (Costa et al., 2019). Zhang et al. (2010) selected 11 volatiles compounds from Chinese red wines by the use of stepwise linear discriminant analysis. The authors found that esters and alcohols are responsible for varietal classification of Chinese Cabernet Sauvignon, Cabernet Gernischt and Merlot wines.

Márquez et al. (2008) found that the variables α-terpineol, 5-methyl-2-furaldehyde, isoamyl alcohols, ethyl decanoate, hexanal and p-cymene was the mainly descriptors to classify Andalusian sweet wines. The study of Mihnea et al. (2015) correctly classified all samples of Mencía wines by the use of 22 of 42 variables, mainly alcohols, esters and acids.

Although the use of variable selectors improves the performance of classification models and reduces the experimental costs, there is little research in the area of wine classification using these techniques. The future of food authentication depends on combining the information from different analytical techniques (Villano et al., 2017) and modeling algorithms capable to handle the non-linearly of data and complex problems, such as artificial neural networks, SVM, decision trees and random forest, that show a great potential and more advantages compared to conventional ones (Jiménez-Carvelo et al., 2019). Therefore, our results strongly contribute as an application of machine learning and feature selection methods to a wine database composed of four commercial categories described on phenolic and volatile compounds.

5. Conclusion

This work demonstrated the use of SVM and two feature selection methods to classify four commercial categories of South American red wines ("Argentinean Malbec (AM)", "Brazilian Merlot (BM)", "Uruguayan Tannat (UT)" and "Chilean Carménère (CC)"). Among the 190 variables that described volatiles, semi-volatiles, antioxidant activity and phenolic compounds, 13 were selected by CFS. From this variable subset nine chemicals were identified by RFI with a classification capacity greater than the set of 13 variables. The SVM was able to classify the samples according to their respective classes in an accuracy of 93.97%. The compounds selected were catechin, gallic, octanoic acid, myricetin, caffeic, isobutanol, resveratrol, kaempferol, and ORAC. That not only reduced the experimental cost effectively, but also made the analysis of the results more comprehensible due the reduced number of variables. PCA was used to an exploratory analysis before and after feature selection. The first plot did not show any patters or groups. On the other hand, the PCA score plot based on the selected variables shows some trends and separation among the classes.

One aspect that differentiates this research from the other studies that conducted wine classification is that we use feature selection techniques to identify a subset of chemical compounds that were able to classify the data from multiple sources. In addition, our methodology can be applied in other products and wines to identify the most important chemicals which characterize the classes, reduce data dimensionality, and improve the classification rate.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Araújo, E.M., de Lima, M.D., Barbosa, R., Alleoni, L.R.F., 2019. Using machine learning and multi-element analysis to evaluate the authenticity of organic and conventional vegetables. Food Anal. Methods, 1–13 https://doi.org/10.1007/s12161-019-01597-2.

Belmiro, T.M.C., Pereira, C.F., Paim, A.P.S., 2017. Red wines from South America: content of phenolic compounds and chemometric distinction by origin. Microchem. J. 133, 114–120.

Benzie, I.F.F., Strain, J.J., 1996. The ferric reducing ability of plasma (FRAP) as a measure of "antioxidant power": the FRAP assay. Anal. Biochem. 239, 70–76.

Brand-Williams, W., Cuvelier, M.-E., Berset, C., 1995. Use of a free radical method to evaluate antioxidant activity. LWT-Food Sci. Technol. 28, 25–30.

Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32. https://doi.org/10.1023/A:1010933404324.

Cabredo-Pinillos, S., Cedrón-Fernández, T., Parra-Manzanares, A., Sáenz-Barrio, C., 2004. Determination of volatile compounds in wine by automated solid-phase microextraction and gas chromatography. Chromatographia 59, 733–738.

Callao, M.P., Ruisánchez, I., 2018. An overview of multivariate qualitative methods for food fraud detection. Food Control 86, 283–293. https://doi.org/10.1016/j.foodcont.2017.11.034.

Chandrashekar, G., Sahin, F., 2014. A survey on feature selection methods. Comput. Electr. Eng. 40, 16–28. https://doi.org/10.1016/j.compeleceng.2013.11.024.

Cordella, C., Moussa, I., Martel, A.C., Sbirrazzouli, N., Lizzani-Cuvelier, L., 2002. Recent developments in food characterization and adulteration detection: technique-oriented perspectives. J. Agric. Food Chem. 50, 1751–1764.

Cortes, C., Vapnik, V., 1995. Support-vector networks. Mach. Learn. 20, 273–297.

Costa, N., Llobodanin, L., Castro, I., Barbosa, R., 2018. Geographical classification of Tannat wines based on support vector machines and feature selection. Beverages 4, 97. https://doi.org/10.3390/beverages4040097.

Costa, N.L., Llobodanin, L.A.G., Castro, I.A., Barbosa, R., 2019. Using support vector machines and neural networks to classify merlot wines from South America. Inf. Process. Agric. 6, 265–278. https://doi.org/10.1016/j.inpa.2018.10.003.

da Costa, N.L., Castro, I.A., Barbosa, R., 2016. Classification of cabernet sauvignon from two different countries in South America by chemical compounds and support vector machines. Appl. Artif. Intell. 30, 679–689. https://doi.org/10.1080/08839514.2016.1214416.

Froz, B.R., de Carvalho Filho, A.O., Silva, A.C., de Paiva, A.C., Nunes, R.A., Gattass, M., 2017. Lung nodule classification using artificial crawlers, directional texture and support vector machine. Expert Syst. Appl. 69, 176–188.

Hall, M.A., 1999. Correlation-Based Feature Selection for Machine Learning. University of Waikato Hamilton.

Huang, D., Ou, B., Prior, R.L., 2005. The chemistry behind antioxidant capacity assays. J. Agric. Food Chem. 53, 1841–1856.

Jaitz, L., Siegl, K., Eder, R., Rak, G., Abranko, L., Koellensperger, G., Hann, S., 2010. LC–MS/MS analysis of phenols for classification of red wine according to geographic origin, grape variety and vintage. Food Chem. 122, 366–372.

Jiménez-Carvelo, A.M., González-Casado, A., Bagur-González, M.G., Cuadros-Rodríguez, L., 2019. Alternative data mining/machine learning methods for the analytical evaluation of food quality and authenticity – a review. Food Res. Int. 122, 25–39. https://doi.org/10.1016/j.foodres.2019.03.063.

Li, H., Liang, Y., Xu, Q., 2009. Support vector machines and its applications in chemistry. Chemom. Intell. Lab. Syst. 95, 188–198.

Llobodanin, L.G., Barroso, L.P., Castro, I.A., 2014. Prediction of the functionality of young South American red wines based on chemical parameters. Aust. J. Grape Wine Res. 20, 15–24.

Maione, C., de Oliveira Souza, V.C., Togni, L.R., da Costa, J.L., Campiglia, A.D., Barbosa, F., Barbosa, R.M., 2018. Establishing chemical profiling for ecstasy tablets based on trace element levels and support vector machine. Neural Comput. & Applic. 30, 947–955.

Márquez, R., Castro, R., Natera, R., García-Barroso, C., 2008. Characterisation of the volatile fraction of Andalusian sweet wines. Eur. Food Res. Technol. 226, 1479.

Mihnea, M., González-SanJosé, M.L., Ortega-Heras, M., Pérez-Magariño, S., 2015. A comparative study of the volatile content of Mencía wines obtained using different pre-fermentative maceration techniques. LWT-Food Sci. Technol. 64, 32–41.

Pisano, P.L., Silva, M.F., Olivieri, A.C., 2015. Anthocyanins as markers for the classification of Argentinean wines according to botanical and geographical origin. Chemometric modeling of liquid chromatography-mass spectrometry data. Food Chem. 175, 174–180.

R Core Team, 2017. R: A Language and Environment for Statistical Computing.

Rebolo, S., Peña, R.M., Latorre, M.J., García, S., Botana, A.M., Herrero, C., 2000. Characterisation of Galician (NW Spain) Ribeira sacra wines using pattern recognition analysis. Anal. Chim. Acta 417, 211–220.

Richter, B., Rurik, M., Gurk, S., Kohlbacher, O., Fischer, M., 2019. Food monitoring: screening of the geographical origin of white asparagus using FT-NIR and machine learning. Food Control 104, 318–325.

Shen, F., Li, F., Liu, D., Xu, H., Ying, Y., Li, B., 2012. Ageing status characterization of Chinese rice wines using chemical descriptors combined with multivariate data analysis. Food Control 25, 458–463.

Soares, F., Anzanello, M.J., Fogliatto, F.S., Marcelo, M.C.A., Ferrão, M.F., Manfroi, V., Pozebon, D., 2018. Element selection and concentration analysis for classifying South America wine samples according to the country of origin. Comput. Electron. Agric. 150, 33–40. https://doi.org/10.1016/j.compag.2018.03.027.

Turra, C., de Lima, M.D., Fernandes, E.A.D.N., Bacchi, M.A., Barbosa, F., Barbosa, R., 2017. Multielement determination in orange juice by ICP-MS associated with data mining for the classification of organic samples. Inf. Process. Agric. 4, 199–205.

Valentin, L., Barroso, L.P., Barbosa, R.M., de Paulo, G.A., Castro, I.A., 2020. Chemical typicality of South American red wines classified according to their volatile and phenolic compounds using multivariate analysis. Food Chem. 302, 125340. https://doi.org/10.1016/j.foodchem.2019.125340.

Versari, A., Laurie, V.F., Ricci, A., Laghi, L., Parpinello, G.P., 2014. Progress in authentication, typification and traceability of grapes and wines by chemometric approaches. Food Res. Int. 60, 2–18. https://doi.org/10.1016/j.foodres.2014.02.007.

Villano, C., Lisanti, M.T., Gambuti, A., Vecchio, R., Moio, L., Frusciante, L., Aversano, R., Carputo, D., 2017. Wine varietal authentication based on phenolics, volatiles and DNA markers: state of the art, perspectives and drawbacks. Food Control 80, 1–10.

Vogado, L.H.S., Veras, R.M.S., Araujo, F.H.D., Silva, R.R.V., Aires, K.R.T., 2018. Leukemia diagnosis in blood slides using transfer learning in CNNs and SVM for classification. Eng. Appl. Artif. Intell. 72, 415–422.

Wine, I.O., 2019. STATE OF THE VITIVINICULTURE WORLD MARKET [WWW Document]. URL. http://www.oiv.int/public/medias/6679/en-oiv-state-of-the-vitiviniculture-world-market-2019.pdf (accessed 3.30.20).

Zhang, J., Li, L., Gao, N., Wang, D., Gao, Q., Jiang, S., 2010. Feature extraction and selection from volatile compounds for analytical classification of Chinese red wines from different varieties. Anal. Chim. Acta 662, 137–142.