Contents lists available at Science-Gate

International Journal of Advanced and Applied Sciences

Journal homepage: http://www.science-gate.com/IJAAS.html

Application of supervised learning algorithm to determine the quality of slippers in WEKA

Jennilyn C. Mina *

College of Management and Business Technology, Nueva Ecija University of Science and Technology, Cabanatuan City, Philippines

ARTICLE INFO

Article history:
Received 17 January 2023
Received in revised form
15 June 2023
Accepted 6 July 2023

Keywords:
Decision tree
K-nearest neighbor
Linear regression
Multi-layer perceptron
Random forest
Support vector machine

ABSTRACT

This study is driven by the objective of evaluating the effectiveness of various regression algorithms in the prediction of slipper quality. The selected regression algorithms were implemented within the Waikato Environment for Knowledge Analysis. The assessment of their performance was conducted through the analysis of correlation coefficients, providing insights into their predictive capabilities. Notably, the Random Forest algorithm demonstrated the highest predictive power with an impressive correlation coefficient (r=0.76), surpassing other models in the analysis. Following Random Forest, the k-nearest neighbor algorithm achieved a substantial correlation coefficient of (r=0.65), followed by the Decision Tree (r=0.53), Linear regression (r=0.51), and the Multi-layer perceptron (r=0.51). In contrast, the Support Vector Machine showed a notably lower correlation coefficient (r=0.51), indicating its comparatively weaker predictive performance. Furthermore, this study uncovered two variables, "Easy to Wash" and "Water Resistance," which displayed significant correlations of (r=0.49) and (r=-0.35), respectively, in relation to the predictive performance of the regression model. However, no significant correlation was observed for other variables. In light of these findings, future research endeavors may explore alternative predictive models to further assess and compare their performance against the outcomes presented in this study, contributing to the ongoing enhancement of slipper quality prediction methodologies.

1. Introduction

Regression algorithms are widely employed for predictive modeling, aiming to forecast specific outcomes based on multiple contributing factors. Simple linear regression proves suitable when dealing with a single independent variable, whereas multiple linear regression becomes necessary when the independent variables encompass a multitude of inputs. It is worth noting that the field of big data analytics is experiencing rapid and transformative growth. In a related research endeavor, artificial intelligence techniques were harnessed to forecast optimal models, exemplified by a study focusing on wine quality prediction. This study harnessed a dataset comprising information on wine quality, leveraging the physicochemical attributes of white wine. The objective extended beyond mere prediction; it also sought to establish a tangible connection between the analysis outcomes and real-world business scenarios.

In today's social milieu, wines play a pivotal role in various occasions, ranging from weddings and social gatherings to birthday celebrations. This burgeoning demand for both white and red wines, along with diverse wine types, is attributed to their prevalence at these events and reflects a broader trend in consumer preferences (Buccola and VanderZanden, 1997). Physicochemical laboratory tests are routinely employed to delineate the attributes of white wine, encompassing measurements of pH, alcohol content, and density or value determination. While sensory assessments often depend on expert evaluations, it is imperative to underscore that within the realm of perception, taste emerges as one of the most intricately nuanced and challenging dimensions to apprehend (Smith and Margolskee, 2006).

As a consequence, the task of wine classification emerges as a formidable endeavor necessitating substantial investments to enhance classification

* Corresponding Author.
Email Address: jennilynmina011@gmail.com

https://doi.org/10.21833/ijaas.2023.08.012

Corresponding author's ORCID profile:
https://orcid.org/0000-0002-7835-6045

precision. Moreover, the intricate and challenging nature of comprehending the intricate associations between sensory analysis and physicochemical analysis further complicates this pursuit (Legin et al., 2003). The quality of wine cannot be determined using a simple process. This includes a variety of techniques, such as data collection from experts and physicochemical laboratory test findings for each of the wines. Financial evaluation of wine items, ensuring and guaranteeing wine quality, preventing wine corruption, and controlling refreshment preparation are some of the reasons (Khalafyan et al., 2021). The Waikato Environment for Knowledge Analysis (WEKA) is an application software developed at the University of Waikato in New Zealand to process data from agricultural fields. WEKA is a cutting-edge research and development laboratory for applying machine learning approaches to real-world data mining situations. The algorithms are applied directly to a dataset. WEKA is capable of a wide range of standard data mining tasks, including pre-processing, classifying, clustering, regression, visualization, and feature selection from given datasets (Frank et al., 2010).

The application's main premise is to use smart computer software to perform machine learning operations and generate useful data in the form of patterns and trends. WEKA is a free application released under the GNU General Public License. It offers a user-friendly graphical interface that enables quick setup and actions. WEKA demands that user data be in the form of a flat file or relation, which implies that each data item is represented by a defined number of characteristics, which are often alpha-numeric or numeric values of a specific kind. The WEKA application provides beginner users with a tool for discovering hidden information from database and file systems according to Frank et al. (2010), with easy settings and visual interfaces.

Despite the fact that some researchers have used machine learning techniques to assess slipper quality, there is still room for improvement. Support Vector Machine, Nave Bayes, and Random Forest are used to try to predict slipper quality. The training and testing sets' outcomes are compared, and the best of the three techniques is projected based on the training set's results. If the best features from different techniques are extracted and merged to improve accuracy and efficiency, better results can be attained. When examining the quality prediction using RStudio software, the Support Vector Machine performed best, with an accuracy of 67.25 percent, followed by the Random Forest, which has an accuracy of 65.83 percent, and the Nave Bayes method, which has an accuracy of 55.91 percent (Kumar et al., 2020).

The outcome of Sun et al. (1997) identified six geographic wine origins based on neural networks fed with 15 input factors. It used 170 data samples for their investigations in Germany. They have a 100% accuracy rate in their predictions. The result of the study utilized a neural network to classify Californian wine in a similar study. The level of grape ripeness and chemical analysis are used to classify wines (Vlassides et al., 2001).

Furthermore, Er and Atasoy (2016) used data mining techniques to predict various diseases such as tuberculosis, cancer, diabetes, and so on, with promising results, whereas Gutiérrez et al. (2017) used a wireless sensor network (WSN) and several techniques to forecast and predict heart disease from the Cleveland dataset's important attributes. On the other hand, Yu et al. (2008) compared many wine classification datasets. Chilean wine was classified using linear discriminant analysis, support vector machines, and neural networks (NN) (Beltrán et al., 2008). The study encompassed the analysis of three distinct Chilean wine varieties. As detailed by Cortez et al. (2009) study, a total of 147 bottles of rice wine were subjected to rigorous examination, leading to their categorization into three distinct wine classes. This investigation employed a dataset comprising nine distinct slipper properties as input variables, coupled with a single sensory data point serving as the output variable. All variables within this dataset are characterized by numerical values and were implemented within the WEKA application.

The input variables encompass a spectrum of characteristics, including comfort, warmth, temperature regulation, appropriate sole, arch support, ease of washing, water resistance, antibacterial fabric, and moisture-wicking. These attributes were derived from the results of a comprehensive survey.

Concurrently, the output variable, denoted as "quality," serves as the focal point for the analysis. Notably, the study extends its scope to encompass an exhaustive comparison of regression algorithms. This comparative assessment aims to elucidate the superior algorithms in terms of their performance metrics, adding a crucial dimension to the research.

2. Methodology

The data was taken from the online survey conducted by the researcher. The slipper dataset contains 400 occurrences with 9 inputs and 1 output variable. These 10 characteristics are taken into account when predicting the quality of the slipper. With the help of the study of Cortez et al. (2009), the analysis was also conducted. The input and output features of the quality of slipper data are summarized in Table 1.

2.1. Algorithms of the experiment

Multiple linear regression serves as the chosen method for predicting slipper quality based on the provided dataset. This technique entails the estimation of coefficients to establish the most fitting line or hyperplane in relation to the training dataset. Multiple linear regression is characterized by its simplicity and ease of comprehension, often yielding commendable results when the output variable within the data exhibits a linear relationship with the input variables.

Table 1: Characteristics of slipper

Attribute	Data	Range	Description
Comfort	Numeric	1–10	Input
Warmth	Numeric	1–8	Input
Temperature-regulation	Numeric	1–10	Input
Appropriate sole	Numeric	3–10	Input
Arch support	Numeric	4–8	Input
Easy to wash	Numeric	2–9	Input
Water resistance	Numeric	3–10	Input
Antibacterial fabric	Numeric	3–8	Input
Moisture-wicking	Numeric	2–8	Input
Quality	Numeric	1–10	Output

Additionally, the k-nearest neighbor algorithm, commonly referred to as kNN, was selected for slipper quality prediction in accordance with the dataset. K-nearest neighbor is a versatile algorithm suitable for both classification and regression tasks. In the prediction process, it retains the entire training dataset and identifies the k most analogous training patterns to derive predictions.

The decision tree methodology operates by constructing a tree structure to evaluate data instances, commencing from the root and progressing to the leaves (note that the tree's depiction is inverted). This process continues until a valid prediction can be generated. The development of a decision tree entails a greedy approach in selecting the optimal split points iteratively, iteratively expanding the tree until it reaches a predefined depth (Holmes et al., 1994).

Support Vector Machines were originally designed to solve binary classification problems, but the approach has now been extended to solve multi-class classification and regression problems. Support Vector Regression, or SVR for short, is an adaption of SVM for regression. Multi-layer perceptron method will solve the issue for both regression and classification. Artificial neural networks, or simply neural networks, are another name for them. Because there are so many configuration parameters that can only be tweaked efficiently by intuition and a lot of trial and error, neural networks are a difficult method to utilize for predictive modeling. It's an algorithm based on a model of biological neural networks in the brain, in which little processing units called neurons are grouped into layers capable of approximating any function if properly set.

Random Forest is an enhancement over bagged decision trees in that it interrupts the greedy splitting algorithm during tree building, allowing only a random portion of the input attributes to be used as split points. This small modification can have a significant impact on the similarity of the bagged trees and, as a result, the forecasts (McClendon and Meghanathan, 2015).

2.2. Evaluation measure of algorithm of the experiment

The mean-squared error is the most popular method for evaluating numeric predictions. To make it the same size as the projected value, the square root is sometimes used. Because it is the easiest measure to adjust conceptually, the mean-squared error is employed in various mathematical procedures, such as linear regression. The use of mean absolute error is another option. The total amount of individual errors is simply averaged, regardless of sign. Mean-squared error, but not absolute error, magnifies the impact of outliers (cases where the prediction error is larger than the others). All incorrect sizes are treated the same, regardless of their magnitude. The relative squared error is a whole different concept. The error is shown as a percentage of what would have happened if a simple predictor had been employed instead. The average of the actual values in the training data, denoted by a, is the simple predictor in question. As a result, the relative squared error divides the total squared error by the total squared error of the default predictor. The root relative squared error is, of course, discovered. To identify the statistical link between the a's and the p's, the correlation coefficient is used. The correlation coefficient spans from perfect positive (+1) to perfect negative (-1) results (McClendon and Meghanathan, 2015).

Below are the formulae for evaluating numerical predictions, where the test instances are denoted as p1, p2, ... pn, and the actual values are represented by a1, a2, .. an..

$$Mean\ square\ error\ = \frac{(p1-a1)^2+..+(pn-an)^2}{n} \tag{1}$$

$$Root\ mean\ square\ error\ = \sqrt{\frac{(p1-a1)^2+..(pn-an)^2}{n}} \tag{2}$$

$$Mean-\ absolute\ error\ = \frac{|p1-a1|+..+|pn-an|}{n} \tag{3}$$

$$Relative-\ absolute\ error\ = \frac{|p1-a1|+..+|pn-an|}{|a1-an|+..+|an-a|} \tag{4}$$

$$Correlation\ = \frac{SpA}{\sqrt{SpSA}} \tag{5}$$

where,

$$SpA\ = \frac{\sum(p1-p)(a1-a)}{n-1} \tag{6}$$

$$Sp\ = \frac{\sum(p1-p)^2}{n-1} \tag{7}$$

$$Sp\ = \frac{\sum(a1-a)^2}{n-1} \tag{8}$$

2.3. Experimental setup

The experiment was conducted in a computer terminal. The operating systems used were Microsoft Windows 10 Home Single Language and a WEKA application software was installed. Fig. 1 shows the procedure on WEKA operation and Fig. 2 shows the sample dataset in execution.

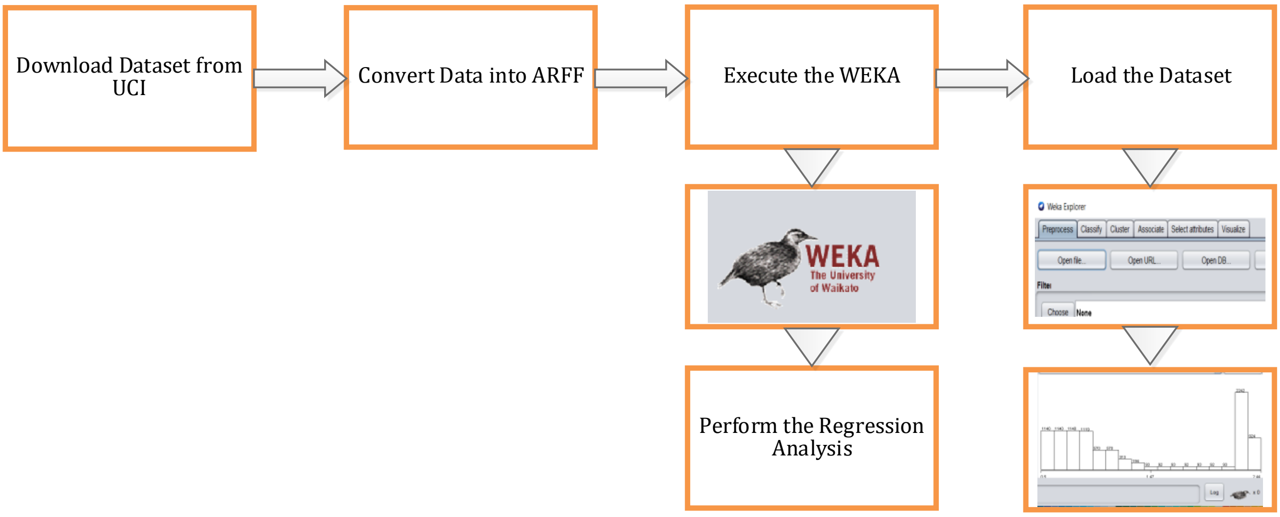
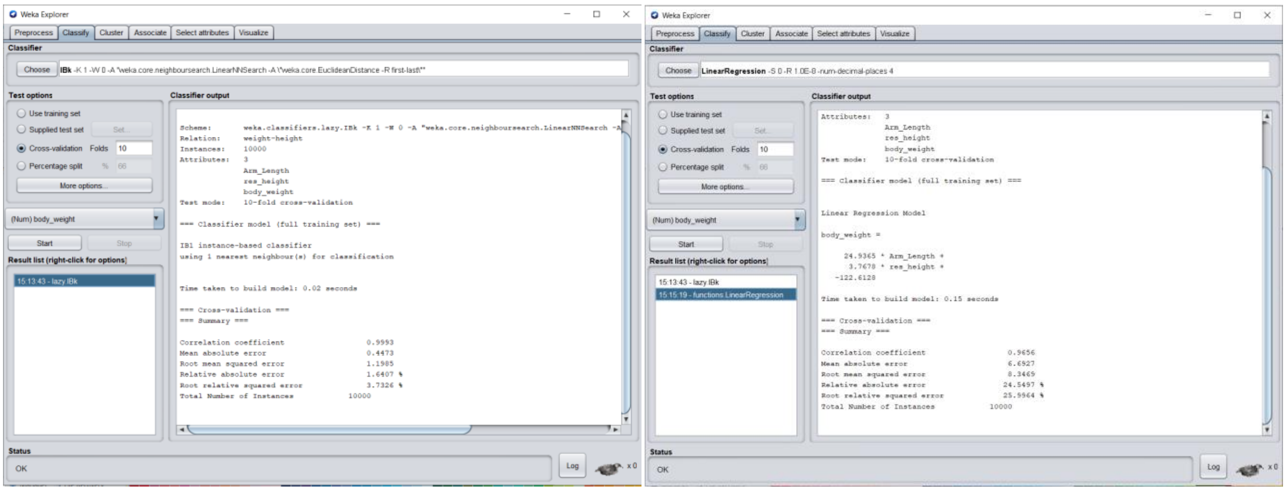


Fig. 1: Fragment of dataset



Classifier: IBk -K 1 -W 0 -A "weka.core.neighboursearch.LinearNNSearch -A \"weka.core.EuclideanDistance -R first-last\""

○ Use training set
○ Supplied test set
● Cross-validation Folds 10
○ Percentage split % 66

(Num) body_weight

Result list: 15:13:43 - lazy.IBk

```
Scheme:       weka.classifiers.lazy.IBk -K 1 -W 0 -A "weka.core.neighboursearch.LinearNNSearch -A
Relation:     weight-height
Instances:    10000
Attributes:   3
              Arm_Length
              res_height
              body_weight
Test mode:    10-fold cross-validation

=== Classifier model (full training set) ===

IB1 instance-based classifier
using 1 nearest neighbour(s) for classification


Time taken to build model: 0.02 seconds

=== Cross-validation ===
=== Summary ===

Correlation coefficient                  0.9993
Mean absolute error                      0.4473
Root mean squared error                  1.1985
Relative absolute error                  1.6407 %
Root relative squared error              3.7326 %
Total Number of Instances            10000
```

Status: OK

Classifier: LinearRegression -S 0 -R 1.0E-8 -num-decimal-places 4

○ Use training set
○ Supplied test set
● Cross-validation Folds 10
○ Percentage split % 66

(Num) body_weight

Result list: 15:13:43 - lazy.IBk
15:15:19 - functions.LinearRegression

```
Attributes:   3
              Arm_Length
              res_height
              body_weight
Test mode:    10-fold cross-validation

=== Classifier model (full training set) ===


Linear Regression Model

body_weight =

     24.9365 * Arm_Length +
      3.7678 * res_height +
   -122.6128

Time taken to build model: 0.15 seconds

=== Cross-validation ===
=== Summary ===

Correlation coefficient                  0.9656
Mean absolute error                      6.6927
Root mean squared error                  8.3469
Relative absolute error                 24.5497 %
Root relative squared error             25.9964 %
Total Number of Instances            10000
```

Status: OK

Fig. 2: Dataset in execution

3. Results and discussions

There is a total of 10 variables of the quality of slipper and the collection is patterned in Cortez et al. (2009) with the aid of Asuncion and Newman (2007) as stated in the methodology section. The variable quality rating is the dependent or predicted variable, whereas the other 9 variables are independent or predictor variables. Data was converted from a comma-separated values (CSV) file type to an attribute-relation file format in this experiment (ARFF). The notepad-converted csv file to ARFF is shown in the image below. The descriptive statistics of the data set are shown in Table 2. Table 2 provides a comprehensive overview of the descriptive statistics pertaining to the attributes under consideration. Notably, the "Arch Support" attribute exhibits the lowest standard deviation (s=0.01), signifying minimal variability, whereas the attribute "Total Water Resistance" presents the highest standard deviation (s=42.50), indicating substantial variability within the dataset.

Model's performance hinges on the correlation coefficient. The experimental outcomes unveil that the Random Forest algorithm attains the highest correlation coefficient (r=0.76), followed by the k-nearest neighbor (r=0.65), Decision Tree (r=0.53), Linear Regression (r=0.51), Multi-layer Perceptron (r=0.51), and Support Vector Machine (r=51).

Table 2: Descriptive statistics of the dataset

Attributes description	Range	Mean	Standard deviation
Comfort	1-10	6.82	0.81
Warmth	1-8	0.22	0.11
Temperature-regulation	1-10	0.31	0.11
Appropriate sole	3-10	6.32	5.02
Arch support	4-8	0.03	0.01
Easy to wash	2-9	35.21	17.00
Water resistance	3-10	138.11	42.23
Antibacterial fabric	3-8	0.93	0.01
Moisture-wicking	2-8	3.12	0.102
Quality	1-10	0.45	0.24

In Table 3, an extensive summary of the performance output related to slipper quality assessment is presented. The WEKA application utilizes equations (1), (2), (3), (4), and (5) to compute performance metrics associated with the provided datasets. The holistic evaluation of the

The outcomes of the correlation study between input and output variables are encapsulated in Table 4. It is noteworthy that the "Easy to Wash" input exhibits a positively moderate association (r=0.013), while the "Antibacterial Fabric" input demonstrates an inversely moderate correlation (r=-0.313). Conversely, the remaining input variables display negligible or no discernible relationships with the output variable.

Table 3: Performance measures

Algorithm	Correlation coefficient	MAE	RMSE	RAE	RRSE	Rank
Linear regression	0.51	0.58	0.75	87.23%	85.04%	4th
k-nearest neighbors	0.65	0.41	0.75	61.73%	85.60%	2nd
Decision tree	0.53	0.57	0.74	85.19%	84.29%	3rd
Support vector machine	0.51	0.58	0.75	87.03%	85.39%	6th
Multi-layer perceptron	0.51	0.60	0.78	90.57%	88.41%	5th
Random forest	0.76	0.41	0.59	62.22%	66.92%	1st

Table 4: Correlation analysis between the inputs and the output

Independent variable(input)	Dependent variable/output	Pearson R
Comfort	quality	-0.110
Warmth	quality	-0.192
Temperature-regulation	quality	-0.013
Appropriate sole	quality	-0.102
Arch support	quality	-0.213
Easy to wash	quality	0.013
Water resistance	quality	-0.172
Antibacterial fabric	quality	-0.313
Moisture-wicking	quality	0.102

4. Conclusions

Data mining is one of the approaches that may be used to investigate data. The WEKA applications were used to process data from the UCI Machine Learning Repository (Asuncion and Newman, 2007) in order to forecast the performance of a model. Furthermore, data mining and related tools, such as WEKA, aid in the production of sound judgments based on the results, which are utilized to benefit both consumers and product producers (Gupta, 2018).

The study has conducted a comprehensive examination of machine learning methodologies for predictive modeling, with a primary focus on assessing their accuracy and determining the most proficient supervised learning tools currently available for model development. The experiments were conducted employing the WEKA application, utilizing a dataset comprising slipper samples.

This dataset encompasses a diverse range of physical attributes, consisting of one dependent variable and nine independent variables. These independent variables are hypothesized to exert a substantial influence on the dependent variable, which pertains to slipper quality.

The study's outcomes reveal that the predictive modeling of slipper quality exhibits a moderate positive association with the inherent qualities of the slippers. Specifically, the Random Forest algorithm achieved the highest correlation coefficient (r=0.76), securing the top position, followed by the k-nearest neighbor (r=0.65), Decision Tree (r=0.53), Linear Regression (r=0.51), Multi-layer Perceptron (r=0.51), and Support Vector Machine (r=51).

As part of prospective research directions, further exploration of the factors influencing slipper quality is warranted. Future investigations may encompass the exploration of alternative machine learning techniques for slipper quality prediction, alongside the incorporation of fresh datasets, with the overarching objective of enhancing the precision of prediction models.

Compliance with ethical standards

Ethical consideration

All processes used to analyze data sets from a specific source complied with ethical guidelines. The data sources are properly credited and listed in the reference section.

Conflict of interest

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

References

Asuncion A and Newman D (2007). UCI machine learning repository. Available online at: https://archive.ics.uci.edu/

Beltrán NH, Duarte-Mermoud MA, Vicencio VAS, Salah SA, and Bustos MA (2008). Chilean wine classification using volatile organic compounds data obtained with a fast GC analyzer. IEEE Transactions on Instrumentation and Measurement, 57(11): 2421-2436. https://doi.org/10.1109/TIM.2008.925015

Buccola ST and VanderZanden L (1997). Wine demand, price strategy, and tax policy. Applied Economic Perspectives and Policy, 19(2): 428-440. https://doi.org/10.2307/1349750

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. https://doi.org/10.1016/j.dss.2009.05.016

Er Y and Atasoy A (2016). The classification of white wine and red wine according to their physicochemical qualities. International Journal of Intelligent Systems and Applications in Engineering, 4(Special Issue-1): 23-26. https://doi.org/10.18201/ijisae.265954

Frank E, Hall M, Holmes G, Kirkby R, Pfahringer B, Witten IH, and Trigg L (2010). WEKA-A machine learning workbench for data mining. In: Maimon O and Rokach L (Eds.), Data Mining and Knowledge Discovery Handbook. Springer, Boston, USA. https://doi.org/10.1007/978-0-387-09823-4_66

Gupta Y (2018). Selection of important features and predicting wine quality using machine learning techniques. Procedia Computer Science, 125: 305-312. https://doi.org/10.1016/j.procs.2017.12.041

Gutiérrez AJ, Rubio C, Moreno IM, González AG, Gonzalez-Weller D, Bencharki N, and Revert C (2017). Estimation of dietary intake and target hazard quotients for metals by consumption of wines from the Canary Islands. Food and Chemical Toxicology, 108: 10-18. https://doi.org/10.1016/j.fct.2017.07.033 **PMid:28733233**

Holmes G, Donkin A, and Witten IH (1994). WEKA: A machine learning workbench. In Proceedings of ANZIIS'94-Australian New Zealand Intelligent Information Systems Conference, IEEE, Brisbane, Australia: 357-361. https://doi.org/10.1109/ANZIIS.1994.396988

Khalafyan AA, Temerdashev ZA, Akin'shina VA, and Yakuba YF (2021). Data on the sensory evaluation of the dry red and white wines quality obtained by traditional technologies from European and hybrid grape varieties in the Krasnodar Territory, Russia. Data in Brief, 36: 106992. https://doi.org/10.1016/j.dib.2021.106992 **PMid:33889695 PMCid:PMC8050733**

Kumar S, Agrawal K, and Mandan N (2020). Red wine quality prediction using machine learning techniques. In the International Conference on Computer Communication and Informatics, IEEE, Coimbatore, India: 1-6. https://doi.org/10.1109/ICCCI48352.2020.9104095

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. https://doi.org/10.1016/S0003-2670(03)00301-5

McClendon L and Meghanathan N (2015). Using machine learning algorithms to analyze crime data. Machine Learning and Applications: An International Journal, 2(1): 1-12. https://doi.org/10.5121/mlaij.2015.2101

Smith DV and Margolskee RF (2006). Making sense of taste. Scientific American, 16(3): 84-92. https://doi.org/10.1038/scientificamerican0906-84sp

Sun LX, Danzer K, and Thiel G (1997). Classification of wine samples by means of artificial neural networks and discrimination analytical methods. Fresenius' Journal of Analytical Chemistry, 359: 143-149. https://doi.org/10.1007/s002160050551

Vlassides S, Ferrier JG, and Block DE (2001). Using historical data for bioprocess optimization: Modeling wine characteristics using artificial neural networks and archived process information. Biotechnology and Bioengineering, 73(1): 55-68. https://doi.org/10.1002/1097-0290(20010405)73:1<55::AID-BIT1036>3.0.CO;2-5 **PMid:11255152**

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. https://doi.org/10.1021/jf0725575 **PMid:18167072**