Application of Linear Discriminant Analysis and k-Nearest Neighbors Techniques to Recommendation Systems

JAVIER BILBAO, IMANOL BILBAO
Applied Mathematics Department,
University of the Basque Country (UPV/EHU),
Bilbao School of Engineering, Pl. Ing. Torres Quevedo, 1, 48013, Bilbao,
SPAIN

Abstract: - Among the different techniques of Machine Learning, we have selected various of them, such as SVM, CART, MLP, kNN, etc. to predict the score of a particular wine and give a recommendation to a user. In this paper, we present the results from the LDA and kNN techniques, applied to data of Rioja red wines, specifically with Rioja Qualified Denomination of Origin. Principal Component Analysis has been used previously to create a new and smaller set of data, with a smaller number of characteristics to manage, contrast, and interpret these data more easily. From the results of both classifiers, LDA and kNN, we can conclude that they can be useful in the recommendation system.

Key-Words: - Machine Learning, recommendation systems, LDA, kNN, Principal Components Analysis, classification regions.

Received: May 29, 2023. Revised: January 2, 2024. Accepted: January 24, 2023. Published: March 4, 2024.

1 Introduction

Currently, Machine Learning techniques are varied and are applied to different fields of science. In addition, they can also be applied to industry. One of those possible applications is the wine industry and the field of enology. The production and consumption of wine in the world is currently of great importance in certain countries, such as Spain, Italy, Greece, Chile, France, etc., [1], [2], [3], [4].

Wine has a tradition in society that goes back thousands of years, integrating itself into the culture of different societies and forming part of everyday life in different classes at different levels, [5], [6]. However, it is still the wine experts who generally mark the quality of wines, also based on laboratory analysis assessments by authorities and wine producers, [7], [8].

This personal experience when tasting a wine depends on each person. Even if it is the same wine, vintage, and production, even the same bottle, each person can feel different nuances and classify the same wine in different ways. Therefore, it is interesting to be able to predict the rating of a wine based on each person.

The mathematical models that can be used to predict each person are very diverse. But if we also want it to be done automatically, statistical techniques based on Machine Learning (ML) can be very valid and are postulated as suitable tools for, among other things, generating personalized predictive models automatically, [9], [10]. These techniques allow making a prediction based on a database of previously collected data and, in our case, focus on the evaluation of wines. In this article, we focus on the prediction of personal evaluation.

By applying ML techniques, it is possible to predict what would be the rating of a wine given by a certain person, as long as the data related to that specific wine is provided (wine characteristics) and there is also a background of the person's tastes.

Machine Learning techniques are usually divided taking into account the type of learning into two main groups, which are unsupervised and supervised, [11].

For unsupervised learning, only the characteristics that identify the product to be compared, which in our case is wine, would be necessary. The score given to the wine in previous tests is not necessary. Although the effectiveness of these unsupervised techniques may normally be lower than supervised learning techniques, the contribution they make to the study can be very useful in simplifying the number of functionalities that are used. Some of these techniques are Linear Discriminant Analysis (LDA) and Principal Component Analysis (PCA), [12].

In contrast, in supervised learning, it is necessary to include the results of previous evaluations, already known, so that the methods can be them and

train with them. Some of the techniques of this type of Machine Learning are support vector machines (SVM) [13], [14], classification and regression trees (CART) [15], [16], k-nearest neighbor (kNN), [17] multilayer perceptron (MLP) [18], [19], Naïve Bayes classifiers (NBC) [20], [21], linear regression (LR) [22] and logistic regression [23].

They are different techniques that can be more or less interesting depending on the problem to be analyzed, its characteristics, the available data, etc. However, sometimes the analysis can obtain better results if several of these techniques are combined, [24], [25].

For the study of wines, one of the characteristics that is usually chosen to try to classify this product is the concentration of anthocyanins, [26], [27].

This article aims to explain the applicability of different Machine Learning techniques to make meaningful recommendations, individually for each person, referring to Rioja red wines, specifically with Rioja Qualified Denomination of Origin (DOC).

2 Data Sets

As a prior step to the study of Machine Learning techniques, different sets of wine characteristics were obtained. These characteristics were the following:

- the characteristics obtained in the analysis of the wines, which, originally, were a total of 62;
- 21 components derived from anthocyanins;
- the PCA components that express 99%, 95%, and 90% of the variability in the data (16 components in the PCA90 set, 23 components in the PCA95 set, and 37 in PCA99);
- the class-independent Fisher discriminant of 3 components for each taster;
- the class-dependent Fisher discriminant with 12 components per taster; and finally,
- three sets of selected characteristics: on the one hand, from the first taster, 19 characteristics from his data; from the second taster, 36 characteristics of the data of that second taster, these two subsets forming two LDA selections; and the third set of selected features was QDA selection, with a total of 21 features.

3 Principal Component Analysis

Principal Components Analysis (PCA) is a Machine Learning technique that fits within the category of unsupervised learning. Based on the characteristics of a data set, this technique allows creation a new set of characteristics, smaller in number and, therefore, easier to manage, contrast, and interpret. To achieve this objective, a linear transformation is used first, to then select a smaller number of those characteristics but without losing important information for the study, [28], [29].

Focusing briefly on the mathematics behind this method, the final objective of PCA is to find an orthogonal matrix that allows a change of characteristics, from the original ones to a new set, in such a way that the characteristics of the new set are not correlated with each other and all of this in order of decreasing variance. This means that this set of new characteristics will have a diagonal covariance matrix and the elements of its main diagonal will be ordered from largest to smallest. The variance captured by each component is represented by the eigenvalue associated with each eigenvector *i*. In this way, the average error committed when approximating the original data with the new set will coincide with the sum of the eigenvalues of the components not selected in the study.

Let X be a random vector of *r* variables (*r*-dimensional), each with *n* observations, which can be expressed as deviations from the mean or standardized:

$$X = \left(X_1, X_2, X_3, ..., X_r \right)' \tag{1}$$

The steps to apply the algorithm are as follows:

- In the data, taking the characteristics *i* one by one, its arithmetic mean is calculated separately, and also a measure of its variance, such as the standard deviation. Subsequently, each characteristic is normalized.
- Then, using the normalized data, the covariance matrix is calculated.

$$S = \frac{1}{r} X_{norm}^T \cdot X_{norm} \tag{2}$$

- Then, the eigenvalues of the matrix *S* and the associated eigenvectors are obtained. The most widely used algorithm for this step is the singular value decomposition and singular vector decomposition.
- Next, the *n* eigenvectors of *S* associated with the largest eigenvalues are taken. Thus, the matrix $U_{reduced}$ is generated.
- Finally, the data for the new vector space is obtained. These will be the reduced (transformed) ones through this technique. For

the training data set, we will therefore have $X_{reduced} = X \cdot U_{reduced}$.

In this way, the original data are compressed and, in addition, its representation is usually possible due to the reduced number of dimensions.

4 The Classifiers

The Machine Learning techniques used were the following: QDA, LDA, NBC (Naïve Bayes), CART, kNN, Multi-Layer Perceptron (MLP) and Probabilistic Neural Networks (PNN). Using these techniques, different families of classifiers were generated. Matlab R2019b was used to carry out the calculations and the study.

The mission of a classifier is to correctly assign data represented by a vector of *d* characteristics, to one of *c* different categories, which have been previously defined. We will use $x = (x_1, x_2, ..., x_d)$ to designate the data and $C_1, C_2, ..., C_c$ to designate the categories or classes. Most Machine Learning techniques search for and assign the category to which the lowest risk is associated if an error occurs. That is, for a certain data *x*, the techniques, in general, first, calculate the risk or consequences of making a wrong decision by incorrectly selecting the category C_i using the expression:

$$R(C_i|x) = \sum_{j=1}^{c} L(C_i, C_j) P(C_j|x) = \\ = \sum_{j=1}^{c} L(C_i, C_j) P(C_j) P(x|C_j) \quad (3)$$

where $L(C_i, C_j)$ represents the losses if it is decided to classify an element as C_i, when in reality it belongs to class C_j. After that, the techniques choose the C_i that minimizes $R(C_i|x)$.

Different expressions can be used for the loss function, depending on how you want to show the result of a bad recommendation.

Normally, the standard expression is used, but two cases tend to get more attention: the first one is when a sample that has the worst possible classification is classified as a positive sample; and, the second case, is when you have an ordinary (or even negative) sample and it is misclassified as positive.

If what happens is that a sample is classified as the best of all the samples, when the truth is that it does not belong to that optimal class, the result would be recommendations that would be misleading and that would cause the users to withdraw their trust in the recommendation system. This may lead to discontinuation of the recommender system or even to penalties.

Generally, some degree of distrust is generated, which is directly proportional to the distance between the categories, but that degree of distrust is not always the same. Fundamentally, the user usually shows more interest in positive recommendations than in negative recommendations. We propose for this case the loss matrix shown in Table 1, where the classification mistakes are taken into account.

In our case, and because the number of samples was not large, the classifiers were validated using Leave One Out (LOO).

Table 1. Loss matrix in which the consequences of the classification mistakes are taken into account.

Real / Classified	**Bad**	**Medium**	**Good**	**Excellent**
Bad	0	2	3	4
Medium	1	0	1	3
Good	1,5	1	0	2
Excellent	3,5	2,5	1,5	0

5 Linear Discriminant Analysis

The Linear Discriminant Analysis (LDA) technique is usually used to classify each of the samples with the assumption that the probability distribution $P(x|C_j)$ is a multivariate Gaussian with mean m_j, and that the covariance matrix is the same for all distributions of the different classes. This covariance matrix is represented as:

$$S'_{\text{intra}} = U^T \cdot S_{\text{intra}} \cdot U \quad (4)$$

with *U* taken from the class-independent Fisher discriminant algorithm.

The decision is made according to the so-called Fisher discriminant functions:

$$\log P(C_j|x) = -\frac{1}{2}(x - m_j) \cdot \\ \cdot S'^{-1}_{\text{int}ra} \cdot (x - m_j)^T + \\ + \log P(C_j) - \frac{1}{2}\log|S'_{\text{int}ra}| - \\ - \frac{n}{2}\log 2\pi \quad (5)$$

If we had standard losses, the decision limits obtained using this technique are the hyperplanes

equidistant from the centroids of the different classes (Figure 1).

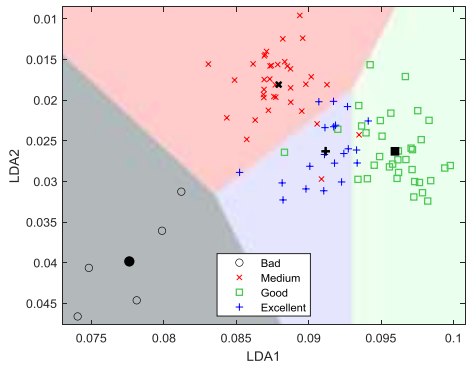

Fig. 1: Decision regions of the not-validated LDA classifier in the first two independent Fisher principal components for the data of the first taster

In our case, when we use the proposed loss matrix, we try to ensure the predictions are classified as positive and negative by decreasing the regions assigned to these classes (Figure 2).

This way of operating can be interpreted in a way that, for the boundaries of those particular regions, we are increasing their safety margin.

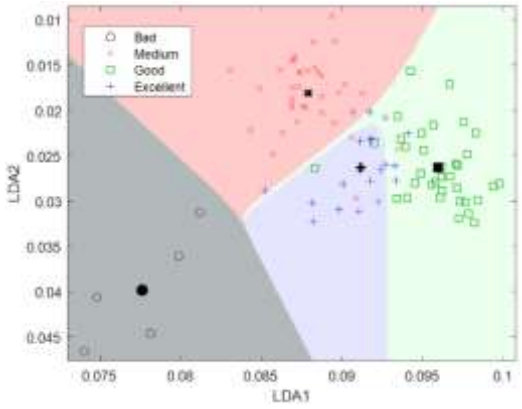

Fig. 2: Decision regions of the not-validated LDA classifier in the first two independent Fisher principal components for the data of the first taster with a non-standard loss matrix

The classifiers validated through LOO that use the first three principal components of the data of the first taster obtain an overall accuracy of 90.62%.

All samples in the Truly Bad category have been correctly classified. However, this technique has had some problems with the samples rated with the highest score by the taster. If we take the 20 samples that obtained the highest score, 1 of them was classified as Good, and 2 as Medium; This implies that 15% of the samples that should be recommended as a priority to a potential user would be lost and would be classified with a lower score. However, the biggest problem that has been found is that this technique recommends three samples as having a maximum score when their true classification is simply Medium. Something similar happens with one of the samples in the Good category, which is also wrongly classified among the best. This can cause great disappointment to the user of the system, since 19% of the samples that the classifier issues as positive recommendations would not be positive (Figure 3).

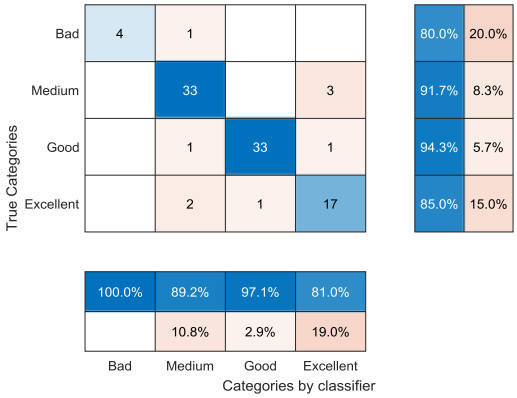

Fig. 3: Confusion matrix of the 3 independent Fisher components of the main class LDA classifier, validated by LOO, for the first taster data with the standard loss matrix

With the data from the second taster, the accuracy of the validated classifier is even better, at 95.31%. These good results with these classifiers may be due to the high separability of the classes.

On the other hand, we have verified that when using PCA to eliminate noise in the data before the Fisher method, no advantage is obtained.

6 k-Nearest Neighbors

The k-Nearest Neighbors (kNN) algorithm is a simple algorithm that falls under the category of supervised learning. It is often used for classification, [30], [31]. Using this algorithm, the decision regions that separate each class can be constructed, without first needing to estimate the density function.

To build these decision or classification regions, all the training cases provided are saved by the algorithm to later compare the distances between the new sample that we want to classify and every one of the training cases. Thus, the algorithm obtains an ordering with the k nearest neighbors. The category assigned to the new input sample is that which is in the majority among those k nearest neighbors.

In the reference [32], several ideas to build a kNN classifier that supports a non-standard loss

matrix are proposed. If we estimate the probability $P(C_j|x)$ using the number of nearest neighbors in each category, we can then calculate the risk to finally make our choice. In this case, the relative frequency of occurrence of each category in the subset of the selected *k* nearest neighbors is taken as an estimate of $P(C_j|x)$.

The number of neighbors to choose and the metric used to measure distances are the parameters in our design. We will previously select the distance metric, which can be Euclidean or another, such as the Manhattan distance, and we will build two different sets of classifiers. Once we preselect the metric, we reserve a sample to test the classifiers using LOO. To choose the most favorable option, of the samples that remain unused, we reserve one to validate the classifier and select the number of neighbors. The rest of the samples will be used as training samples.

Following this procedure, in the case of the first data set, we obtain 96 different classifiers with 95 validated variations for each one according to the number of neighbors. Using each of these variations, we can construct the confusion matrix and calculate the total losses by applying the loss matrix over the confusion matrix. Finally, we proceed to calculate the average of the costs among the 95 experiments carried out, choosing the optimal number of neighbors (Figure 4).

We want to highlight here that the samples that had been reserved to test the final classifier have not been used to obtain this value.

The decision regions (or, in general, classification regions) are different depending on whether Euclidean distance classifiers or Manhattan distance classifiers are used.

Figure 5 shows the different classification regions in the case of the Manhattan distances for a classifier with the loss matrix of Table 1 on the first data set built using only the first two Fisher principal components. As it can be seen, such a metric distance tends to form separation regions parallel to the axes along which distances are measured.

This vision of the problem shows us that, in this case, the classifiers with the proposed losses also tend to increase the regions of the intermediate categories, such as Medium and Good, reducing at the extremes, that is, in the classifications called Bad and Excellent.

This acts again as an increase in the guard zone when making significant predictions, trying to discard doubtful cases from the Excellent and Bad categories and ensuring in some way that the resulting predictions of those categories are more reliable. This increase translates into a greater extension of the regions associated with the intermediate categories.

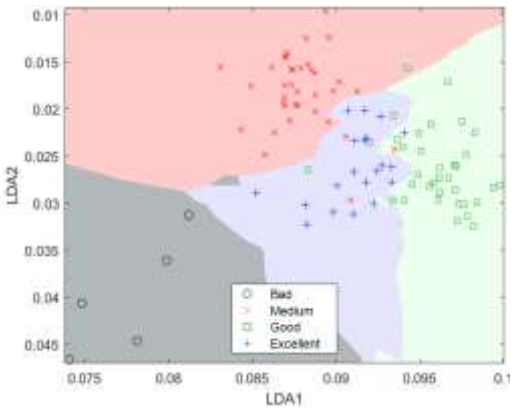

Fig. 4: Classification regions using the first two class-independent Fisher principal components for *k*=4 along with standard lossy Euclidean distance

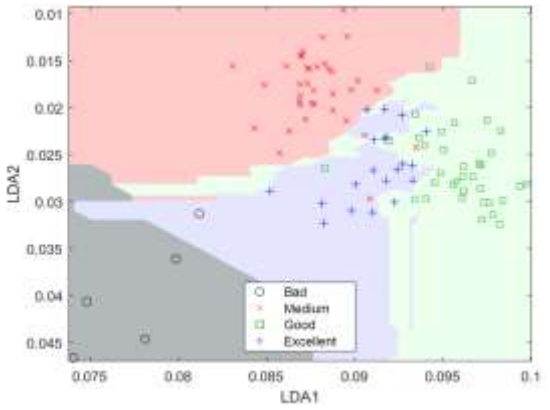

Fig. 5: Classification regions using the first two class-independent Fisher principal components for *k*=4 along with the Manhattan distance with proposed losses

Table 2 shows the results obtained after applying the kNN technique to the second data set. These results can be considered excellent since they represent the characteristics in the independent components of each Fisher class with more than 95% accuracy. The use of different distances or metrics does not seem too relevant since the results are very similar, although it is true that it provides slight improvements in the classifiers.

The differences between using one metric or another are smaller than the margin of error that arises when validating the data.

Figure 6 and Figure 7 show the confusion matrices and it can be seen that they coincide for both types of metrics when the number of neighbors is optimal.

We also want to highlight that the proposed loss classifier provides correct recommendations in 100% of the cases, losing only one case that was originally Excellent in the process.

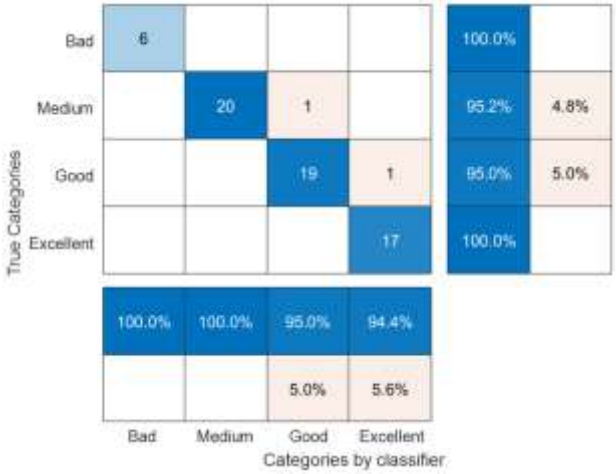

Fig. 6: Confusion matrix of the kNN classifier validated using LOO on the second data set, without losses

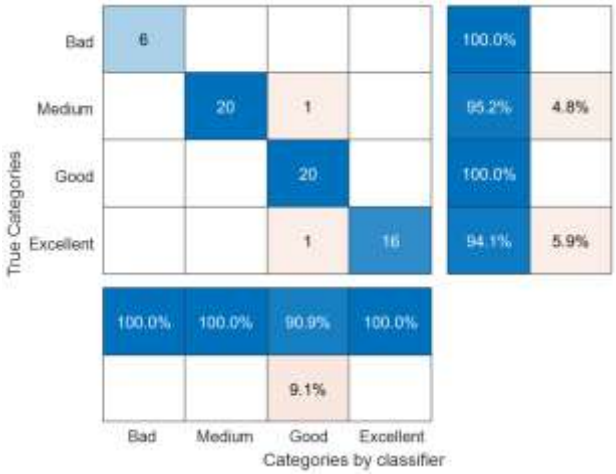

Fig. 7: Confusion matrix of the kNN classifier validated using LOO on the second data set, with losses

If we consider the first data set, the application of the kNN technique obtains lower precision results, slightly lower than 89% of global accuracy.

7 Conclusion

This article presents a comparison of various Machine Learning techniques applied to the classification of red wines from Rioja.

The novelty focuses on the applicability and also on the results of the PCA, LDA, and kNN techniques, comparing the results obtained with each of these techniques on the same data.

The scores of the wines have been grouped into four different categories. This has made it easier for the samples to be classified according to the opinions of the different tasters.

Furthermore, it has been demonstrated that four factors are sufficient to characterize the wines, in this case, red wines from Rioja, to create a recommendation system. These factors have been: anthocyanin derivatives, alcoholic content, tannins, and anthocyanins.

If we take into account the representation spaces of the samples, there is no significant advantage in the classifiers if we apply the PCA technique to the original data. Furthermore, it does not improve the classification results if the PCA technique is applied before the Fisher decomposition.

More reliable meaningful predictions can be made if we use the proposed loss matrix for classifier generation. In this way, greater accuracy is obtained per significant category. Unfortunately, it is necessary to reduce the total number of recommended wines to achieve this greater precision, that is, some possible recommendations must be lost.

When we apply the proposed loss matrix, the positive and negative predictions are ensured, and at the same time the regions assigned to these classes are decreased.

Both classifiers, LDA and kNN, can be useful in the recommendation system. On the one hand, the kNN classifier without losses in the LDA components with 4 neighbors and Euclidean distance offers the second best classification rate for wines categorized as Excellent (88.2%). In this process, only 20% and 25% of the original samples are lost. On the other hand, the LDA classifier with standard losses offers an intermediate level: its success rates are 81% in the Excellent category and losses of 20% and 15% of the original samples.

Table 2. Design parameters and general results of the kNN classifiers taking into account the second data set

		Standard Losses		Proposed Losses	
	Distance	Optimum number of neighbors	Accuracy	Optimum number of neighbors	Accuracy
Originals	Euclidean	18	35.40	13	40.63
	Manhattan	13	37.50	13	37.50
Anthocyanin derivatives	Euclidean	4	43.75	15	39.06
	Manhattan	15	45.31	15	40.63
LDA_2 Selection	Euclidean	3	43.75	6	37.50
	Manhattan	2	48.44	2	48.44
QDA Selection	Euclidean	11	43.75	18	37.50
	Manhattan	11	42.19	11	40.63
99%	Euclidean	2	43.75	16	37,50
	Manhattan	1	45.31	1	45,31
95%	Euclidean	2	43.75	2	43,75
	Manhattan	2	45.31	2	45,31
90%	Euclidean	2	45.31	2	45,31
	Manhattan	2	46.88	2	46,88
	Euclidean	5	96.88	9	96.88
	Manhattan	9	96.88	11	96.88
Fischer's dependent	Euclidean	2	46.88	9	43.75
	Manhattan	9	60.94	8	59.38

Acknowledgement:
We would like to thank the work and generosity of the members of the QAProdNat research group, especially Dr. Noelia Prieto Perea and Dr. Luis Ángel Berrueta Simal, whose previous work served as an invaluable basis for this research, and Dr. Oihane Elena Albóniga Díez, the catalyst of this relationship.

References:

[1] G. Vazquez Vicente, V. Martin Barroso, F. J. Blanco Jimenez, Sustainable tourism, economic growth and employment—The case of the wine routes of Spain, *Sustainability*, vol. 13, no 13, 2021, pp. 7164.

[2] B. Marco-Lajara, P. Seva-Larrosa, J. Martínez-Falcó, F. García-Lillo, Wine clusters and Protected Designations of Origin (PDOs) in Spain: an exploratory analysis, *Journal of Wine Research*, vol. 33, no 3, 2022, pp. 146-167.

[3] J. P. Torres, J. I. Barrera, M. Kunc, S. Charters, The dynamics of wine tourism adoption in Chile, *Journal of Business Research*, vol. 127, 2021, pp. 474-485.

[4] C. Yang, C. Menz, H. Fraga, S. Costafreda-Aumedes, L. Leolini, M. C. Ramos, M. C., D. Molitor, C. van Leeuwen, J. A. Santos, Assessing the grapevine crop water stress indicator over the flowering-veraison phase and the potential yield lose rate in important European wine regions, *Agricultural Water Management*, vol. 261, 2022, pp. 107349. https://doi.org/10.1016/j.agwat.2021.107349.

[5] J. A. Santos, H. Fraga, H., A. C. Malheiro, J. Moutinho-Pereira, L. T. Dinis, C. Correia, M. Moriondo, L. Leolini, C. Dibari, S. Costafreda-Aumedes, T. Kartschall, C. Menz, D. Molitor, J. Junk, M. Beyer, H. R. Schultz, A review of the potential climate change impacts and adaptation options for European viticulture, *Applied Sciences*, vol. 10, no 9, 2020, pp. 3092. https://doi.org/10.3390/app10093092.

[6] E. Pijet-Migoń, P. Migoń, Linking wine culture and geoheritage—Missing opportunities at European UNESCO World Heritage sites and in UNESCO Global Geoparks? A survey of web-based resources, *Geoheritage*, vol. 13, no 3, 2021, pp. 71.

[7] V. Santos, P. Ramos, N. Almeida, E. Santos-Pavón, Developing a wine experience scale: a new strategy to measure holistic behaviour of wine tourists, *Sustainability*, vol. 12, no 19, 2020, pp. 8055.

[8] I. Dos Santos, G. Bosman, J. L. Aleixandre-Tudo, W. du Toit, Direct quantification of red wine phenolics using fluorescence spectroscopy with chemometrics, *Talanta*, vol. 236, 2022, pp. 122857.

[9] M. Torrisi, G. Pollastri, Q. Le, Deep learning methods in protein structure prediction, *Computational and Structural Biotechnology Journal*, vol. 18, 2020, pp. 1301-1310.

[10] F. Huang, Z. Cao, J. Guo, S. H. Jiang, S. Li, Z. Guo, Comparisons of heuristic, general statistical and machine learning models for landslide susceptibility prediction and

mapping, *Catena,* vol. 191, 2020, pp. 104580. https://doi.org/10.1016/j.catena.2020.104580.

[11] M. Alloghani, D. Al-Jumeily, J. Mustafina, A. Hussain, A. J. Aljaaf, A systematic review on supervised and unsupervised machine learning algorithms for data science, *Supervised and unsupervised learning for data science*, 2020, pp. 3-21. https://doi.org/10.1007/978-3-030-22475-2_1.

[12] D. K. Choubey, M. Kumar, V. Shukla, S. Tripathi, V. K. Dhandhania, Comparative analysis of classification methods with PCA and LDA for diabetes, *Current diabetes reviews*, vol. 16, no 8, 2020, pp. 833-850. https://doi.org/10.2174/1573399816666200123124008.

[13] B. E. Boser, I. M. Guyon and V. N. Vapnik, A training algorithm for optimal margin classifiers, *Proceedings of the fifth annual workshop on Computational learning theory - COLT 92*, 1992.

[14] C. Cortes and V. Vapnik, Support-vector networks, *Machine Learning*, vol. 20, 1995, pp. 273-297.

[15] L. Breiman, J. Friedman, C. J. Stone and R. A. Olshen, *Classification and Regression Trees*, Taylor & Francis, 1984.

[16] L. Breiman, Random forests, *Machine learning*, vol. 45, 2001, pp. 5-32.

[17] T. Cover and P. Hart, Nearest neighbor pattern classification, *IEEE Transactions on Information Theory*, vol. 13, 1967, pp. 21-27.

[18] D. E. Rumelhart, G. E. Hinton and R. J. Williams, *Learning internal representations by error propagation*, 1985.

[19] B. Widrow and M. A. Lehr, 30 years of adaptive neural networks: perceptron, madaline, and backpropagation, *Proceedings of the IEEE*, vol. 78, 1990, pp. 1415-1442.

[20] R. O. Duda, P. E. Hart and D. G. Stork, *Pattern Classification*, Wiley John & Sons, 2000.

[21] P. Langley, W. Iba, and K. Thompson, An analysis of Bayesian classifiers, *Proceedings of the Tenth National Conference on Artificial Intelligence*, 1992, pp. 223–228.

[22] D. Maulud, A. M. Abdulazeez, A review on linear regression comprehensive in machine learning, *Journal of Applied Science and Technology Trends*, vol. 1, no 4, 2020, pp. 140-147. https://doi.org/10.38094/jastt1457.

[23] D. W. Hosmer Jr, S. Lemeshow and R. X. Sturdivant, *Applied logistic regression*, John Wiley & Sons, 2013.

[24] C. El-Hajj, P. A. Kyriacou, A review of machine learning techniques in photoplethysmography for the non-invasive cuff-less measurement of blood pressure, *Biomedical Signal Processing and Control*, vol. 58, 2020, pp. 101870.

[25] M. Elbadawi, S. Gaisford, A. W. Basit, Advanced machine-learning techniques in drug discovery, *Drug Discovery Today*, vol. 26, no 3, 2021, pp. 769-777. https://doi.org/10.1016/j.drudis.2020.12.003

[26] Y. Ju, L. Yang, X. Yue, Y. Li, R. He, S. Deng, X. Yang, Y. Fang, Anthocyanin profiles and color properties of red wines made from Vitis davidii and Vitis vinifera grapes, *Food Science and Human Wellness*, vol. 10, no 3, 2021, pp. 335-344. https://doi.org/10.1016/j.fshw.2021.02.025.

[27] A. B. Bautista-Ortín, J. I. Fernández-Fernández, J. M. López-Roca, E. Gómez-Plaza, The effects of enological practices in anthocyanins, phenolic compounds and wine colour and their dependence on grape characteristics, *Journal of Food Composition and Analysis*, vol. 20, no 7, 2007, pp. 546-552.

[28] I. Bilbao, J. Bilbao, C. Feniser, A. Borsa, Practical data mining applied in steel coils manufacturing, *Acta Technica Napocensis-Series: Applied Mathematics, Mechanics, and Engineering*, vol. 63, no 3, 2020.

[29] Y. M. Sebzalli, X. Z. Wang, Knowledge discovery from process operational data using PCA and fuzzy clustering, *Engineering Applications of Artificial Intelligence*, 14, 2001. https://doi.org/10.1016/S0952-1976(01)00032-X.

[30] I. Revilla, S. Pérez-Magariño, M. L. González-SanJosé and S. Beltrán, Identification of anthocyanin derivatives in grape skin extracts and red wines by liquid chromatography with diode array and mass spectrometric detection, *Journal of Chromatography A*, vol. 847, 1999, pp. 83-90. https://doi.org/10.1016/S0021-9673(99)00256-3.

[31] N. Katanić, K. Fertalj, Improving Physical Security with Machine Learning and Sensor-Based Human Activity Recognition, *WSEAS Transactions on Information Science and Applications*, vol. 14, pp. 1-9, 2017.

[32] Z. Qin, A. T. Wang, C. Zhang, S. Zhang, S., Cost-Sensitive Classification with k-Nearest Neighbors, *Knowledge Science, Engineering and Management*, Springer, Berlin,

Heidelberg, 2013. https://doi.org/10.1007/978-3-642-39787-5_10.

Contribution of Individual Authors to the Creation of a Scientific Article (Ghostwriting Policy)

Conceptualization, J.B. and I.B.; methodology, J.B. and I.B.; software, I.B.; validation, J.B. and I.B.; formal analysis, J.B. and I.B.; investigation, J.B. and I.B.; resources, J.B. and I.B.; data curation, I.B.; writing—original draft preparation, J.B. and I.B.; writing—review and editing, J.B. and I.B.; visualization, J.B.; supervision, J.B.; project administration, J.B. All authors have read and agreed to the published version of the manuscript.

Sources of Funding for Research Presented in a Scientific Article or Scientific Article Itself

No funding was received for conducting this study.

Conflict of Interest

The authors have no conflicts of interest to declare.