Improved Learning Approach for Wine Quality Prediction Addressing Feature Dimensionality Issue

Ohjesvey Bhat, Research Scholar, Swami Vivekananda Group of Colleges, Banur, Punjab, ojesvi.bhat@gmail.com

Dr Sashi Jawla, H.O.D, ECE Dept, Swami Vivekananda Group of Colleges, Banur, Punjab, shahsijawla@gmail.com

Abstract: **Product quality certificates are being used by enterprises to market their goods. This process takes a long time and needs to be evaluated by human professionals, which makes it very costly. This paper presents an effective and efficient wine quality prediction approach that is based on Ensemble Learning (EL) Techniques. The primary objective of the proposed is to reduce error value which thereby enhances its accuracy. To accomplish this task, we have used a Wine quality dataset that is taken from UCI ML repository which comprises a total of 4898 samples and 12 features. Since the data present in the dataset is not balanced and hence data preprocessing technique is applied to it. During the preprocessing phase all the input variables and output variables are separated. After that, we employed hybrid feature selection techniques, that combine chi-Square and Principal Component Analysis (PCA) techniques, to address the dataset dimensionality concerns. Finally, ensemble learning techniques which are a combination of Random Forest (RF), XGBoost, and Gradient Boost Machine learning classifiers are used for classification. Utilizing MATLAB software, the suggested EL-based wine quality prediction model's simulations and experiments are analyzed. The simulation outcomes were obtained in terms of Mean Square Error (MSE), Mean Absolute Percentage Error (MAPE), R2 and confusion matrix.**

Keywords — Wine Quality, Feature Selection, Classification, Machine Learning, quality prediction, artificial intelligence etc.

I. INTRODUCTION

Due to research showing a positive association between wine drinking and heart rate variability, there has been a little increase in wine consumption in recent years [1]. The most popular beverage consumed worldwide is wine, and society appreciates it highly. The wine sector has an approximately $300 billion global market value. Over the past few years, wine consumption has significantly expanded due to its health advantages for human hearts as well as for pleasure ones. In order to increase output and improve the efficiency of the entire process, new techniques and methodologies are being applied across all industries today. The cost of these processes is rising over time, as are the expectations placed upon them. The chemicals used in wines, in contrast, are more or less the same, but they serve different purposes, therefore it is important to determine the type of chemicals used, which is why we utilize techniques to verify [2]. Wine, formerly seen as a luxury good, is now consistently enjoyed by a wide spectrum of customers.

Wine is a complicated beverage chemically, made up of water, ethanol, sugar, amino acids, polyphenolic compounds, anthocyanins, and other organic and inorganic components [3,4]. Portugal is the world's eleventh-largest wine producer.

A typical wine contains ethyl alcohol, sugar, acids, higher alcohols, tannins, aldehydes, esters, amino acids, minerals, vitamins, anthocyanins, minor constituents like flavoring compounds etc. [5]. The five important components of wine are Water, Alcohol, Acid, sugar and Phenolic compounds.

- **Water:** 80–90% of wine is simply plain old water. Although most of this is the water that naturally existed in the grapes, a winemaker may occasionally add water to the original grape juice if the alcohol or phenolic compounds are too strong for the grapes.
- **Alcohol:** consists of 8 to 15% of the wine's volume. The abv (alcohol by volume) is displayed on the label. 15% of the wine may be a warm climate red, and 8% could be a cool temperature white.
- **Acid:** keeps the wine from tasting floppy. It provides a tangy zing to wine Wine contains a little amount of acids and the range is 5% to.75%.

DOI : 10.35291/2454-9150.2023.0048

- **Sugar:** Brix, or sugar levels, are one measurement used by winemakers to establish whether a grape is ripe for harvest. This usually ranges from 15% to 28%.
- **Phenolic compounds:** These are some of the most important quality indications for wine since they have a huge influence on sensory characteristics like color, flavor, bitterness, and astringency. In addition to these benefits, phenolic compounds are beneficial to health because of their antioxidant, antibacterial, and vitaminic properties that help ward off cancer and cardiovascular diseases. Such particles affect a wine's flavor, fragrance, and texture even though they may be many and minute in relation to the other elements [6,7].

In order to ensure quality and prevent contamination, certification and evaluation of wine are crucial components of Portugal's wine business. In contrast to earlier periods, when resources and technology were scarce, testing and quality assertion of wines could not be accomplished. This is an important factor today because of the standards of quality and since it is difficult to remain in the market given the competition. A physicochemical test or a sensory test can be used to determine the quality of wine [8]. The first test can be established without human involvement, while the second can be accomplished with human expert guidance. The wine's quality is an important consideration for both consumers and wine-making firms. Accreditation of product quality is used by corporations to boost revenue. Wine is a commonly consumed beverage nowadays, and companies use product quality certification to increase their commercial viability. Testing for product quality used to be done at the conclusion of production. This process is quite costly since it requires a lot of time and energy, such as the use of several human specialists to assess product quality [9]. Assessing the wine's quality with human experts is challenging because each individual has a unique perspective on the test. Organizations began to count on a variety of testing tools as technology developed for the development phase. They could be able to judge the wine's quality more accurately as a result, that obviously saves a lot of money and time. Additionally, this helped to acquire a plethora of data on a variety of topics, such as the amount of various chemicals and temperature used during manufacturing, as well as the quality of the wine produced.Due to the success of ML techniques over the past 10 years, various attempts to evaluate wine quality using existing data have been made [10,11,12]. Throughout this process, the factors that directly impact the wine's quality can be changed. This gives the creator a better idea of how to modify different elements during the production process to enhance the wine quality. This may also result in wines with various tastes, and lastly, it may result in the creation of a new brand. Analysis of the key variables that affect wine quality is thus vital.

The coming up sections of the paper is categorized as: Section 2 presents the literature survey of various recently proposed wine quality prediction models along with their problem statement. Section 3 discusses the proposed work and its working methodology. Section 4 presents results obtained for the proposed model and finally Section 5 concludes the paper.

II. Literature Review

A. *Review Stage*

Please Over the past few years, a number of researchers have proposed various ML based wine quality prediction models in order to assess its quality. Some of the recently proposed wine quality prediction models are discussed here: In order to forecast the quality of wine, **K. R. Dahal, et al. [13]**, examined the effectiveness of a number of ML models, including Ridge Regression (RR), Support Vector Machine (SVM), Gradient Boosting Regressor (GBR), and multi-layer Artificial Neural Network (ANN). Results revealed that GBR outperforms all other models, with MSE, R, and MAPE values of 0.3741, 0.6057, and 0.0873, correspondingly. **Sunny Kumar, et al. [14],** proposed a wine quality prediction model in which they used methods like NB, SVM and RF. **Terry Hui-Ye Chiu, et al. [15],** a hybrid model for predicting wine quality was developed which combines at least two classifiers, such as the random forest and support vector machine. Experiments were also conducted on the wine datasets to assess the usefulness of the proposed hybrid approach and demonstrated its supremacy. **Shaw, B. et al. [16],** compared SVM, Random Forest, and Multilayer Perceptron classification methods for wine quality analysis in order to determine which classification technique produces more accurate results. **Mahima, G.U., et al. [17],** examined the wine quality by using an RF algorithm whose accuracy was further enhanced by KNN. The output of the proposed algorithm was used to categorize the wines into three categories: Fine, Medium, and Poor. **S. Aich, et al. [18],** examined a number of feature selection methods, namely simulated annealing (SA) and feature selection based on genetic algorithms (GA). The scientists used multiple performance indicators like accuracy, sensitivity, specificity, positive predictive value, and negative predictive value to compare various sets of features and supervised machine learning techniques. **Gupta, Mohit et al. [19],** utilized multiple machine learning algorithms to forecast the quality of wine, and the findings are validated using a variety of quantitative measurements. **Yogesh Gupta, et al. [20],** investigates the application of machine learning methods like linear regression, neural networks, and support vector machines for the quality of the product. Red wine and white wine databases were used for all trials. This study demonstrated that choosing a subset of traits (variables) to take into account, as opposed to all of them, can lead to improved predictions. **Xinpeng Ma, et al. [21],** predicted wine quality factors using chemometric techniques and infrared spectroscopy (IRs). Fisher Discriminant-Variable Selection (FD-VS) method, a novel variable (wavelength) selection

DOI : 10.35291/2454-9150.2023.0048

method, was developed.

From the above given literature, it is analyzed that a number of researchers have proposed various ML based wine quality prediction systems. Undoubtedly, these methods were giving good results but, we observed that there is a scope of improvement. Majority of the researchers have employed ML algorithms for predicting wine quality, however, these ML algorithms are not capable of handling large and massive datasets which causes overfitting and hence reduced accuracy of the system. Furthermore, we also observed that very few researchers have applied feature selection techniques in their work that help in overcoming dataset dimensionality issues. Additionally, there hasn't been a lot of study on cutting-edge approaches like ensemble learning that can yield meaningful results for evaluating wine quality. Given these factors, it becomes necessary to implement an effective system that can overcome the aforementioned limitations while also increasing system accuracy.

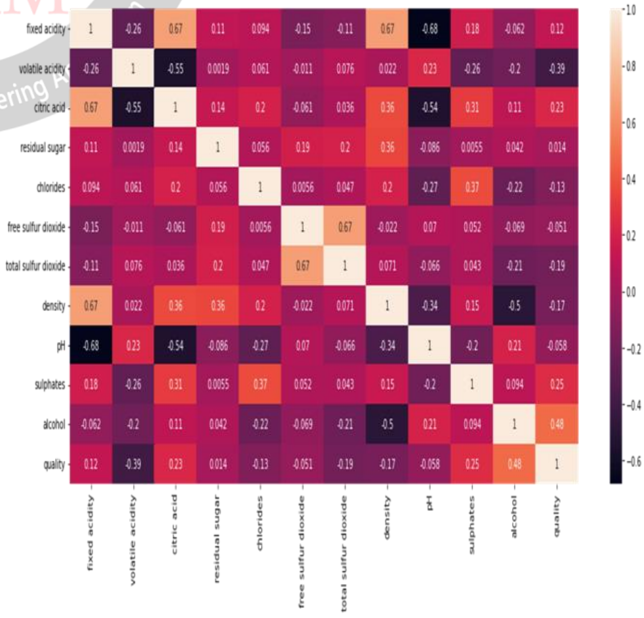
III. MATH

In order to overcome the limitations of existing wine quality prediction models, a new and effective wine quality predictive model is proposed in this paper, based on ensemble learning techniques. The key objective of the proposed EL based wine quality prediction model is to reduce error values so that overall performance of the model can be enhanced. To combat this task, we have updated mainly two phases i.e. Feature selection and Classification phase. In the proposed work, we have introduced hybrid Chi Square and Principal Component Analysis (PCA) based feature selection techniques for resolving the dataset dimensionality issues. Additionally, we used the ensemble learning method that incorporates the Random Forest (RF), XGBoost, and Gradient Boost models of machine learning classifiers. In the proposed work, the ensemble learning model is mostly used since it combines the output from three ML classifiers and produces results which are less noisy than those from individual models. The proposed model is strengthened and stabilized as a result. In the very beginning, a wine quality dataset is taken from UCI ML repository in which data is particularly related to red and white wine variants of Portuguese. This dataset is not balanced and contains a lot of unnecessary and null values that may degrade the efficacy of the system. Therefore, it is important to apply data pre-processing techniques in which input variables and target variables are separated. Following this, several wine characteristics such as density, PH, sulfates, fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, and alcohol are examined. Chi Square and PCA, a hybrid feature selection approach, are then employed to the processed dataset to address the dimensionality problems. Additionally, we have built an ensemble learning-based classification technique that groups together the three ML classifiers RF, XGBoost, and Gradient Boost. The training set is fed to the classifiers, who are subsequently fed the testing dataset to gauge how well the model performed. The wine properties are examined by the suggested ensemble learning technique, which produces a final output called Quality with a score range of 0 to 10. By producing a less noisy outcome than traditional models, ensemble learning methods guarantee great model stability and resilience. Therefore, by integrating hybrid feature selection techniques along with ensemble learning methods, the error values are reduced in the system which in turn ensures high system performance. The detailed working of proposed EL based wine quality prediction model is given in next followed up section of the paper.

A. *Methodology*

In order to achieve the desired objective of the proposed EL model, it undergoes through a number of stages like data collection, data pre-processing, feature selection, classification and performance evaluation. The brief but stepwise working of the proposed EL model is given in this section of paper.

Step 1: Data Collection: Initially, a Wine quality dataset is taken from UCI ML repository that contains two datasets of red and white wine variants taken from Portugal. The multivariate data includes 12 characteristics and a total sample size of 4898 samples. Although the dataset has sorted classes, they are not balanced. Moreover, the dataset contains a total of 11 input attributes i.e. fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, PH, sulfates and alcohol. In addition to this, there is one output variable i.e. Quality whose score range varies from 0 to 10 and is shown in figure 1.



	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
fixed acidity	1	-0.26	0.67	0.11	0.094	-0.15	-0.11	0.67	-0.68	0.18	-0.062	0.12
volatile acidity	-0.26	1	-0.55	0.0019	0.061	-0.011	0.076	0.022	0.23	-0.26	-0.2	-0.39
citric acid	0.67	-0.55	1	0.14	0.2	-0.061	0.036	0.36	-0.54	0.31	0.11	0.23
residual sugar	0.11	0.0019	0.14	1	0.056	0.19	0.2	0.36	-0.086	0.0055	0.042	0.014
chlorides	0.094	0.061	0.2	0.056	1	0.0056	0.047	0.2	-0.27	0.37	-0.22	-0.13
free sulfur dioxide	-0.15	-0.011	-0.061	0.19	0.0056	1	0.67	-0.022	0.07	0.052	-0.069	-0.051
total sulfur dioxide	-0.11	0.076	0.036	0.2	0.047	0.67	1	0.071	-0.066	0.043	-0.21	-0.19
density	0.67	0.022	0.36	0.36	0.2	-0.022	0.071	1	-0.34	0.15	-0.5	-0.17
pH	-0.68	0.23	-0.54	-0.086	-0.27	0.07	-0.066	-0.34	1	-0.2	0.21	-0.058
sulphates	0.18	-0.26	0.31	0.0055	0.37	0.052	0.043	0.15	-0.2	1	0.094	0.25
alcohol	-0.062	-0.2	0.11	0.042	-0.22	-0.069	-0.21	-0.5	0.21	0.094	1	0.48
quality	0.12	-0.39	0.23	0.014	-0.13	-0.051	-0.19	-0.17	-0.058	0.25	0.48	1

Figure 1. Output variable Quality of dataset

Step 2: Data Pre-Processing: Once the relevant information has been gathered, it is essential to pre-process

DOI : 10.35291/2454-9150.2023.0048

it because it is not balanced in nature. All input variables and output variables are divided during the pre-processing stage. Additionally, the dataset is improved and balanced by eliminating all the redundant data.

Step 3: Feature Selection: Once the data is processed, it is time to implement Feature selection technique on processed data to resolve dimensionality issues. Here, we have combined Chi Square and Principal Component Analysis (PCA) for combating this task. Only the most significant characteristics from the processed dataset that are essential to improving the system's accuracy are chosen using the hybrid feature selection method. Additionally, the hybrid feature selection method increases the speed of ML algorithms. A final feature vector is created by choosing only crucial features, and it only contains crucial information.

Step 4: Data Separation: Immediately after this, the processed data is then divided into training and testing data in the ratio of 80:20. The 80% of the given data is used for training the model and the remaining 20% of data is used for testing the efficacy of proposed EL model.

Step 5: Classification: In this phase, we have utilized ensemble learning method wherein three ML classifiers i.e. Random Forest (RF), XGBoost and gradient boost classifiers are used. The ensemble model is trained using the training data, and its performance is evaluated using the testing data. The ensemble learning model examined the various wine parameters and forecasts the wine's quality.

Step 6: Performance Evaluation: The feasibility and efficacy of the proposed EL-based wine quality estimation method are evaluated and verified in the final stage of the proposed model by contrasting it with conventional wine quality prediction models in terms of several performance metrics. The following section of this paper discusses the results that were acquired for the same.

IV. Results And Discussion

The effectiveness and performance of the suggested EL based wine quality prediction approach is examined in MATLAB Software. To prove the supremacy of the proposed approach, we compared its performance with few state of art wine quality prediction approaches in terms of MSE, MAPE, R2, and confusion matrix. This section presents a detailed discussion of various results obtained for proposed EL based wine quality prediction approach.

A. Performance Analysis

While validating the proposed model's performance, we analyzed its performance firstly in terms of confusion matrix. The graph obtained for the same is shown in figure 2. The graph for the confusion matrix clearly states that the difference between actual and predicted values in the proposed EL based wine quality prediction model is small, thereby proving the efficiency of the proposed model.

Moreover, to prove the effectiveness of suggested EL based wine quality prediction model, we also compared its performance with standard RR, SVM, GBR and ANN models in terms of their MSE values. The comparative graph obtained for MSE is shown in figure 3.

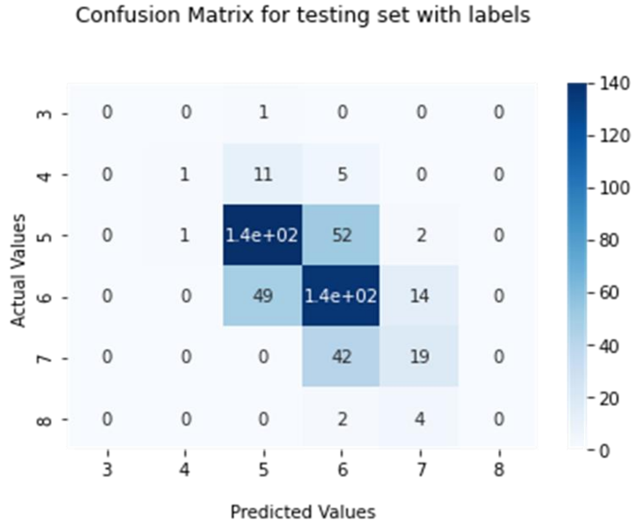


Figure 2. Confusion matrix attained in proposed model

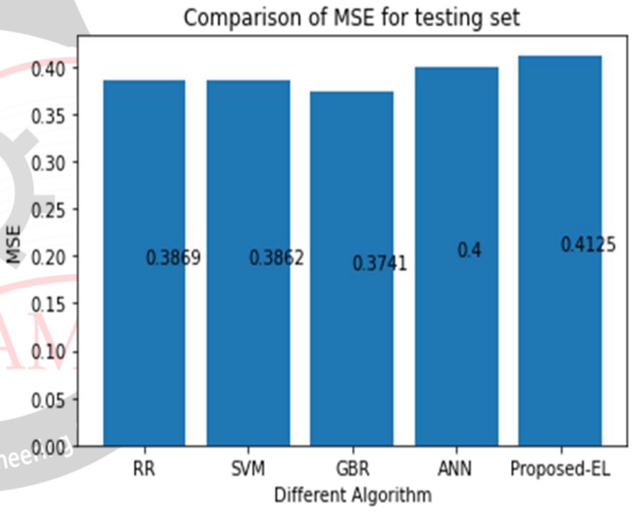


Figure 3. Comparison graph for MSE

Upon critically investigating the provided graph, we discovered that the proposed EL-based wine quality model's MSE value comes out to be 0.41. In contrast, the MSE values for the conventional RR, SVM, GBR, and ANN techniques were just 0.3869, 0.3862, 0.37, and 0.4. This demonstrates that the proposed EL model has a somewhat higher MSE value than other comparable techniques.

The reliability of the suggested EL-based wine quality prediction model was also examined, and its MAPE values were compared to those of conventional models. The resulting graph is displayed in figure 5.6. After examining the above figure, it can be seen that the classic Svm classifier had the greatest MAPE value (0.1355), trailed by the ANN (0.12), the GBR (0.0873), and the RR (0.08888), approaches. However, compared to previous similar

DOI : 10.35291/2454-9150.2023.0048

techniques, the MAPE score in the proposed EL-based wine quality prediction model was much lower at only 0.0677.

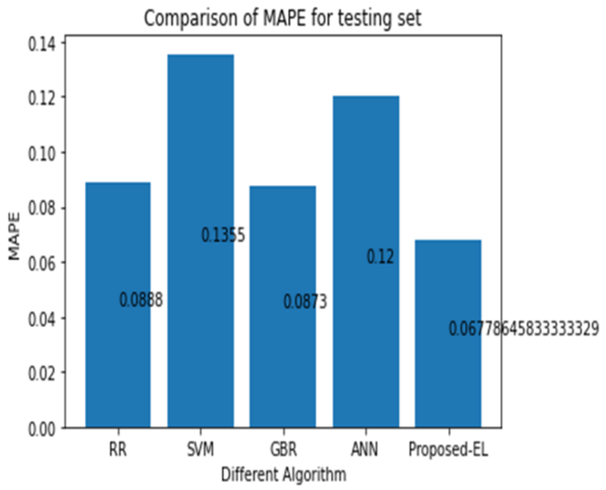


Figure 4. Comparison graph for MAPE

This decreased MAPE value demonstrates the effectiveness of the suggested EL-based wine quality prediction model because low mistakes translate into high accuracy rates.

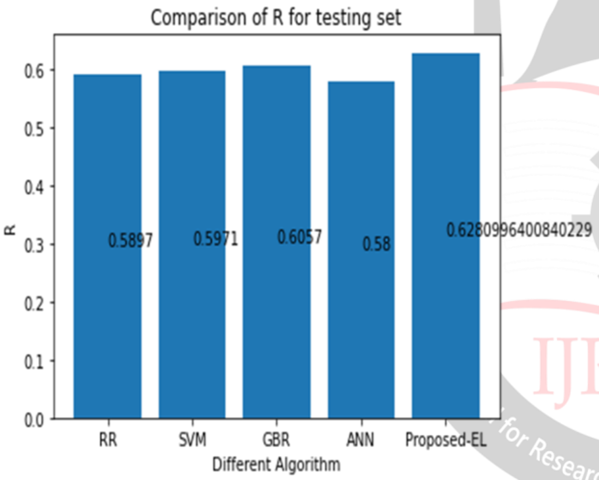


Figure 5. Comparative graph for R2

In addition to this, we have also analyzed and compared the performance of proposed EL based wine quality prediction model with conventional models in terms of their R2 values. The comparative graph obtained for the same is shown in figure 5.

Table 1: Comparison table for different parameters

Algorithms	MSE	MAPE	R^2
RR	0.3869	0.0888	0.05897
SVM	0.3862	0.1355	0.5971
GBR	0.3741	0.0873	0.6057
ANN	0.4	0.12	0.58
Proposed EL	0.4125	0.0677	0.62809

The R2 values for the RR, SVM, GBR, and ANN models were 0.58, 0.59, 0.60, and 0.58, respectively, according to the provided graph. In contrast, the proposed EL model's R2 value, that was 0.6280, was much greater than that of other models of a similar standard, including RR, SVM, GBR, and ANN. A model's goodness of fit is essentially represented by the value R2. Therefore, the value of R2 should be as highest as possible to determine the effectiveness and efficiency of wine quality prediction model. Table 1 lists the exact MSE, MAPE, and R2 values that were achieved for the suggested EL model as well as the conventional RR, SVM, GBR, and ANN models.

From the above graphs and tables, it is clear that the proposed EL based wine quality prediction model is outperforming traditional RR, SVM, GBR and ANN models in terms of MAPE, R2 values to prove its supremacy.

V. Conclusion

This paper presents a productive and efficient Wine quality prediction model that is based on Ensemble learning (EL) technique. The productivity and usefulness of the proposed EL based wine quality prediction approach is analyzed in MATLAB Software. The experimental outcomes were obtained and contrasted with traditional models in terms of performance metrics like MSE, MAPE, R2 and confusion matrix. After analyzing the results, it is observed that MSE value was 0.41 in the proposed EL model whereas, it was 0.3869, 0.3862, 0.37 and 0.4 in conventional RR, SVM, GBR and ANN models. Moreover, the value of MAPE is also analyzed whose values came out to be 0.088 in RR, 0.13 in SVM, 0.087 in GBR and 0.12 in ANN models. While as, the value of MAPE was only 0.0677 in proposed EL based wine quality prediction model. Furthermore, the value of R2 was examined whose value was 0.628 in the proposed EL based wine quality prediction model. On the other hand the value of R2 was analyzed whose values were only 0.058, 0.59, 0.60 and 0.58 in conventional RR, SVM, GBR and ANN models. These values prove the efficacy and efficiency of the proposed EL model over other similar approaches.

References

[1] Aich, Satyabrata, et al. "Prediction of quality for different type of wine based on different feature sets using supervised machine learning techniques." 2019 21st International Conference on Advanced Communication Technology (ICACT). IEEE, 2019.

[2] Uniyal, Xitiz, Prashant Barthwal, and Ashish Joshi. "Wine Quality Evaluation Using Machine Learning Algorithms." Asia-Pac. J. Converg. Res. Interchange 3.4 (2017): 1-9.

[3] Martínez-Lapuente, Leticia, Zenaida Guadalupe, and Belén Ayestarán. "Properties of wine polysaccharides." Pectins-extraction, purification, characterization and applications (2019).

[4] Jones-Moore, Hayden R., et al. "The interactions of wine polysaccharides with aroma compounds, tannins, and proteins, and their importance to winemaking." Food Hydrocolloids 123 (2022): 107150.

[5] Gutiérrez-Escobar, Rocío, María José Aliaño-González, and Emma Cantos-Villar. "Wine polyphenol content and its influence on wine quality and properties: A review." Molecules 26.3 (2021): 718.

[6] Garrido, Jorge, and Fernanda Borges. "Wine and grape polyphenols—A chemical perspective." Food research international 54.2 (2013): 1844-1858.

[7] Lorrain, Bénédicte, et al. "Evolution of analysis of polyhenols from grapes, wines, and extracts." Molecules 18.1 (2013): 1076-1100.

[8] Gupta, Ujjawal, et al. "Wine quality analysis using machine learning algorithms." Micro-Electronics and Telecommunication Engineering. Springer, Singapore, 2020. 11-18.

[9] Kothawade, Rohan Dilip. "Wine quality prediction model using machine learning techniques." (2021).

[10] Li, H., Zhang Z. and Liu, Z.J. (2017) Application of Artificial Neural Networks for Catalysis: A Review. Catalysts, 7, 306. https://doi.org/10.3390/catal7100306

[11] Shanmuganathan, S. (2016) Artificial Neural Network Modelling: An Introduction. In: Shanmuganathan, S. and Samarasinghe, S. (Eds.), Artificial Neural Network Modelling, Springer, Cham, 1-14. https://doi.org/10.1007/978-3-319-28495-8_1

[12] Jr, R.A., de Sousa, H.C., Malmegrim, R.R., dos Santos Jr., D.S., Carvalho, A.C.P.L.F., Fonseca, F.J., Oliveira Jr., O.N. and Mattoso, L.H.C. (2004) Wine Classification by Taste Sensors Made from Ultra-Thin Films and Using Neural Networks. Sensors and Actuators B: Chemical, 98, 77-82.

[13] Dahal, K. , Dahal, J. , Banjade, H. and Gaire, S. (2021) Prediction of Wine Quality Using Machine Learning Algorithms. Open Journal of Statistics, 11, 278-289. doi: 10.4236/ojs.2021.112015.

[14] S. Kumar, K. Agrawal and N. Mandan, "Red Wine Quality Prediction Using Machine Learning Techniques," 2020 International Conference on Computer Communication and Informatics (ICCCI), 2020, pp. 1-6, doi: 10.1109/ICCCI48352.2020.9104095.

[15] Terry Hui-Ye Chiu, Chien-Wen Wu, Chun-Hao Chen, "A Hybrid Wine Classification Model for Quality Prediction", Pattern Recognition. ICPR International Workshops and Challenges, vol.12664, pp.430, 2021.

[16] Shaw, B., Suman, A.K., Chakraborty, B.: Wine quality analysis using machine learning. In: Mandal, J.K., Bhattacharya, D. (eds.) Emerging Technology in Modelling and Graphics. AISC, vol. 937, pp. 239–247. Springer, Singapore (2020). https://doi.org/10.1007/978-981-13-7403-6_23

[17] Mahima, G.U., Patidar Y., Agarwal, A., Singh, K.P.: Wine quality analysis using machine learning algorithms. In: The Micro-Electronics and Telecommunication Engineering, Lecture Notes in Networks and Systems (2020). https://doi.org/10.1007/978-981-15-2329-8_2

[18] S. Aich, A. A. Al-Absi, K. Lee Hui and M. Sain, "Prediction of Quality for Different Type of Wine based on Different Feature Sets Using Supervised Machine Learning Techniques," 2019 21st International Conference on Advanced Communication Technology (ICACT), 2019, pp. 1122-1127, doi: 10.23919/ICACT.2019.8702017.

[19] Gupta, Mohit & Chandrasekaran, Vanmathi..”A Study and Analysis of Machine Learning Techniques in Predicting Wine Quality”. International Journal of Recent Technology and Engineering, May 2021.

[20] Yogesh Gupta, Selection of important features and predicting wine quality using machine learning techniques, Procedia Computer Science, Volume 125, 2018, Pages 305-312

[21] Xinpeng Ma, Jiafeng Pang, Runan Dong, Chen Tang, Yuxuan Shu, Yankun Li, Rapid prediction of multiple wine quality parameters using infrared spectroscopy coupling with chemometric methods, Journal of Food Composition and Analysis, Volume 91, 2020

DOI : 10.35291/2454-9150.2023.0048