Wine Quality Prediction using Machine Learning

*M Murugesan[1], G Prabakaran[2]
[1]Assistant Professor, Vel Tech Multi Tech Dr.RangarajanDr.Sakunthala Engineering College, TN
[2]Assistant Professor, St. Martin's Engineering College, Secunderabad, Telangana-500100
Email: murugesanmecse@gmil.com

ABSTRACT

The quality of a wine is important for the consumers as well as the wine industry. Nowadays, machine learning models are important tools to replace human tasks. There are several features to predict the wine quality but all methods are not preferable. So, our thesis work is focusing on what wine features are important to get the promising result. We will implement by using three algorithms namely Support Vector Machine (SVM), Random Forest Classifier (RFC), and Decision Tree. This project proposes solutions for a better quality of wine with accurate results. The proposed wine quality prediction was done using a Machine Learning Support Vector Machine algorithm. This improved wine quality approach was able to recognize quality of wine with high accuracy level.

Keywords: Wine Quality, Anaconda tool, Display output.

I. INTRODUCTION:

Today, a wider group of customers is enjoying wine more and more. In 2021, wine exports from all nations reached a global total of $40.7 billion. Since 2017, when wine shipments were valued at $35.4 billion, that amount in dollars represents an average 15% growth for all exporting countries1. Export sales of wines increased by 19.8% from $34.3 billion in 2020 annually. France, Italy, Spain, Chile, and Australia are the top five countries for wine exports (Fig. 1). Regarding dollar sales, that potent group of suppliers was responsible for 70.4% of the wine exported worldwide. With shipments totalling $31.1 billion, or 76.4% of the world's wine exports, Europe had the largest dollar export value of any continent in 2021. Australia and New Zealand led Oceania's sales of imported wine, which were 7.5% higher than Latin America's 7.1%, which included the Caribbean but excluded Mexico 23 . North American wine exporters provided 3.8% of the world's wine exports, while Asia delivered 3.3% ahead of 1.9% of sales of wine from African producers. In summary, the wine industry is investing in technology to improve both sales and production of wine, and quality evaluation and certification are essential factors in this process. Certification helps protect human health by preventing unlawful wine adulteration and ensuring market quality. Te certification process4 often involves evaluating wine quality using physicochemical and sensory tests, with physicochemical tests being based on laboratory measurements and sensory tests relying on human expertise. However, the relationship between physicochemical and sensory analysis is complex and not fully understood, making it challenging to classify wine accurately based on taste. Te advancement of information technology has made it possible to collect, store, and analyse large, complex datasets to improve decision-making and increase the likelihood of success. Machine learning algorithms are used to create sophisticated knowledge from unstructured data. Various machine learning algorithms5 include linear and multivariate regression, neural networks, and support vector machines. Tese algorithms have different benefits and are helpful for different types of data. However, it is essential to carefully select the appropriate

II. Related work:

Te wine business mainly uses ML techniques during the wine production process. Despite the ability of ML models to forecast wine quality based on physicochemical data, their use is quite limited and often considers small datasets. This section presents the literature review of the very popular and highly cited work. In several crucial ways, the authors' research sets itself apart from earlier investigations. First of, their use of the red wine dataset with its eleven unique physiochemical traits makes it unique and may provide insights into a dataset for wine quality forecasting that hasn't been well studied. Second, their inventive feature selection strategy may offer a special way to find the most important characteristics. It makes use of machine learning methods like Random Forest and Extreme Gradient Boosting. Te authors also want to demonstrate greater model performance, maybe by careful model calibration and hyper parameter adjustment. Tier research differs from others in terms of technique due to the use of clustering techniques for data preparation. Finally, their research goals or applications connected to wine quality prediction may provide fresh perspectives or useful uses that haven't been widely explored in earlier papers. These disparities add up to the structural and theoretical divergences between the authors' study and the body of prior work in the topic. Using objective hypothesis testing accessible at the certification stage, Cortez et al.8 aim to anticipate wine tastes. White Vinho Verde samples from northwest Portugal were included in a significant dataset. Regression analysis was used to analyse this case study. Te regression model modelled wine preference on a continuous scale from 0 to 10. An efficient and robust process that simultaneously selects variables and modelling and is directed by sensitivity analysis was used to apply three regression techniques.

Agarwal et al.9 evaluate how a deep learning algorithm forecasts for quality by employing two different convolution layers rather than focusing on various approaches. It will let winemakers use deep learning to judge how to manage their operations. Te experiment's limited data set and feature set made it impossible for a machine to choose the most helpful characteristics. By considering several feature selection techniques, such as the Recursive Feature Elimination method (RFE) and Principal Component Measurement (PCA) for feature selection, as well as non-linear decision tree-based classifiers for the analysis of performance indicators, Aich et al.10 developed a new technique. Teir investigation can aid wine specialists in understanding the crucial elements to consider when choosing high-quality wines. Gupta et al.'s11 machine learning algorithm with a user interface forecasts the wine quality by selecting the key wine factor vital for determining the wine quality. Te Random Forest method evaluates wine quality, and KNN is used to improve the model's accuracy further. Te result of the suggested model is utilized to assign the wines a Good, Average, or Bad quality rating. Te goal of Kumar et al. study12 is to determine the quality of red wine using a range of its characteristics. Methods like RF, SVM, and NB are employed, and the dataset is gathered from the sources. Te outcomes are compared between the training dataset and testing set, several performance measures are computed, and the optimum of the three techniques is therefore predicted based on the learning set outcomes. Shaw et al.13 compares the SVM, RF, and multilayer perception classification algorithms for wine quality analysis to determine which algorithm produces the most accurate results. Te multilayer perception algorithm comes in second place with an accuracy of 78.78%, followed by the SVM algorithm with an accuracy of 57.29% during our comparative analysis between those algorithms. Te RF algorithm produces the best results with an accuracy of 81.96%.

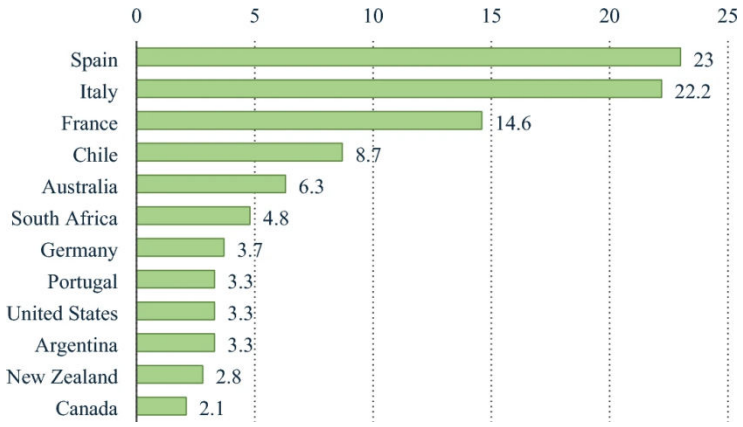

Figure1. Wine export volume in a million hectoliters in 2021.

Related work

The wine business mainly uses ML techniques during the wine production process. Despite the ability of ML models to forecast wine quality based on physicochemical data, their use is quite limited and often considers small datasets. This section presents the literature review of the very popular and highly cited work. In several crucial ways, the authors' research sets itself apart from earlier investigations. First of, their use of the red wine dataset with its eleven unique physio chemical traits makes it unique and may provide insights into a dataset for wine quality forecasting that hasn't been well studied. Second, their inventive feature selection strategy may offer a special way to find the most important characteristics. It makes use of machine learning. Methods like Random Forest and Extreme Gradient Boosting. The authors also want to demonstrate greater model performance, maybe by careful model calibration and hyper parameter adjustment. Their research differs from others in terms of technique due to the use of clustering techniques for data preparation. Finally, their research goals or applications connected to wine quality prediction may provide fresh perspectives or useful uses that haven't been widely explored in earlier papers.

These disparities add up to the structural and theoretical divergences between the authors' study and the body of Prior work in the topic. Using objective hypothesis testing accessible at the certification stage, Cortez et al. Aim to anticipate wine tastes. White Vinho Verde samples from northwest Portugal were included in a significant dataset. Regression analysis was used to analyze this case study. The regression model modeled wine preference on a continuous scale from 0 to 10. An deficient and robust process that simultaneously selects variables and modeling and is directed by sensitivity analysis was used to apply three regression techniques. Agarwal et al.9 evaluate how a deep learning algorithm forecasts for quality by employing two different convolution layers rather than focusing on various approaches. It will let winemakers use deep learning to judge how to manage their operations. he experiment's limited data set and feature set made it impossible for a machine to choose the most helpful characteristics. By considering several feature selection techniques, such as the Recursive Feature Elimination method (RFE) and Principal Component Measurement (PCA) for feature selection, as well as non-linear decision tree-based classifiers for the analysis of performance indicators, Arch et al. developed a new technique. Their investigation can aid wine specialists in understanding the crucial elements to consider when choosing high-quality wines. Gupta et al.11 machine learning algorithm with a user interface

forecasts the wine quality by selecting the key wine factor vital for determining the wine quality. he Random Forest method evaluates wine quality, and KNN is used to improve the model's accuracy further. he result of the suggested model is utilized to assign the wines a Good, Average, or Bad quality rating. he goal of Kumar et al. Study is to determine the quality of red wine using a range of its characteristics. Methods like RF, SVM, and NB are employed, and the dataset is gathered from the sources. The outcomes are compared between the training dataset and testing set, several performance measures are computed, and the optimum of the three techniques is therefore predicted based on the learning set outcomes. Shaw et al. compares the SVM, RF, and multilayer perception classification algorithms for wine quality analysis to determine which algorithm produces the most accurate results. The multilayer perception algorithm comes in second place with an accuracy of 78.78%, followed by the SVM algorithm with an accuracy of 57.29% during our comparative analysis between those algorithms. he RF algorithm produces the best results with an accuracy of 81.96%. Bhardwaj et al. examined the chemical and physicochemical data from New Zeala.

Material and methods

This study's analysis was completed using the Google Colab notebook, Python version 3.8.16. The operating system that was installed on the system was Windows 10 64-bit. An NVIDIA Ge Force 1 GB graphics card and an Intel i5-Core 2.5 GHz processor with 8 GB RAM round out the hardware specifications. his work is implemented by programming with Python, proposing the framework, and using the classifier provided by the Scikit-Learn package.

Datasets

There are two datasets regarding the red and white varieties of Portuguese "Vinho Verde" wine. This project uses only the red wine dataset (RWD) Paulo Cortez's website and the UCI Machine Learning Repository have available datasets. his study examined the physicochemical variables as input and sensory variables as output from available due to logistical and privacy concerns (for instance, there is no data about grape types, wine brands, wine selling price, etc.). There are far more average wines than good or bad ones, and the classes are organized but unbalanced.

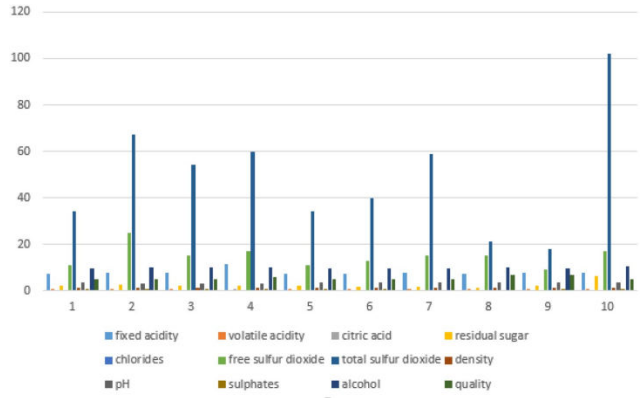

Figure 2. Snapshot of the RWD.

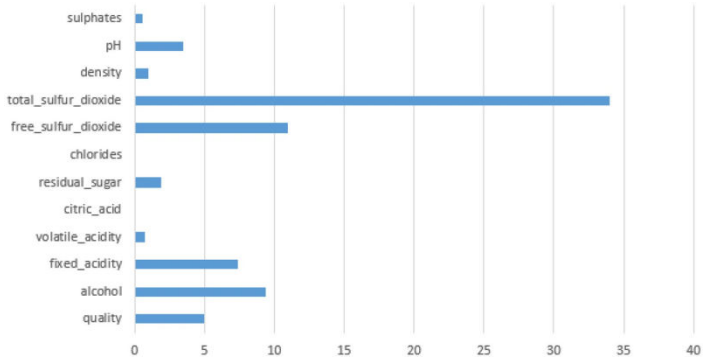

Figure 3. Plot of physicochemical attributes of RWD.

Machine Learning analysis

To establish the correlation between the various features to understand their relationships better, we plot a correlation plot between all the features of RWD, as represented in Fig. 6. As seen from this figure that some features that are strongly correlated to quality. So, these variables are also the most important features in the ML Analysis and models. To compare the performance of various classifiers, we have already transformed the output variable to a binary as "good quality" If quality = 7 and 'bad quality' if if quality < 7. The feature variables (X) will also be separated from the target variable (Y) into different data frames. To cross-validate the ML models and assess their delicacy, we have divided the data into training and test sets at 80% and 20%. We have compared vie different machine learning models, DT, RF, Adam Boost, Gradient Boost, and XG Boost, for their accuracy. An ensemble machine learning approach called gradient boosting is well known for its exceptional predictive power in regression and classification applications. This technique works by sequentially assembling a group of weak learners, frequently decision trees. It begins with an initial forecasting, usually a straightforward one, and then moves on to find and it additional weak learners who are precisely targeted at the residuals or mistakes generated by the current ensemble. With each iteration, these fresh recruits are carefully chosen to cut down on mistakes and eventually increase the model's accuracy. By merging the predictions from each weak learner, the final prediction is made, creating a powerful and incredibly accurate predictive model. In operations research, strategic planning, and ML, decision trees are a common model. Although they are simple to construct and intuitive, decision trees are inaccurate. An ensemble learning method based on decision trees is called random forests. Several decision trees are constructed using random forests utilizing bootstrapped datasets of the original information, which then randomly select a subset of the variables for each step of the decision tree. The model then selects the mode of each decision tree's predictions. By relying on a "majority wins" approach, the likelihood of a single tree making a mistake is reduced.

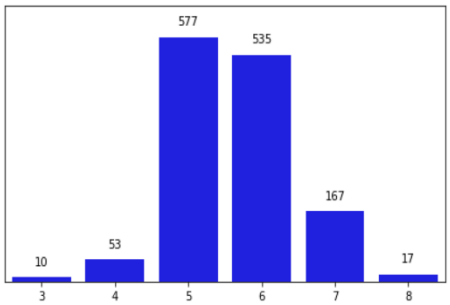

Figure 4. Plot for wine quality.

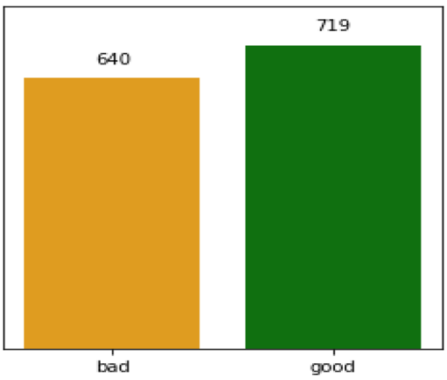

Figure 5. Comparison plot for two classes of RWD.

Feature selection

In this section, we used these two ML methods to extract the best three features from eleven features, and we then executed ML analysis on the returned features. The quality of the wine was estimated using ML techniques. As already stated, the RWD is categorized as a binary classification problem. The default parameters for each ML classier were utilized. We graphed the feature importance based on the RF model in Fig. 7 and the XG Boost model in Fig. 8. While they vary slightly, the top 3 features are the same: alcohol, sulfates, and volatile acidity.

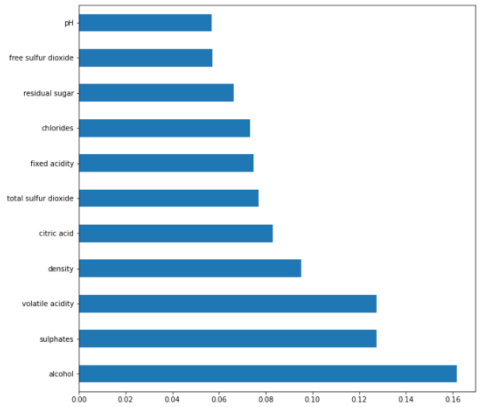

Figure 6. Feature importance based on the RF model. he top four features are alcohol, sulphates, volatile acidity and density.

By examining these descriptive statistics from Tables 4 and 5, we can infer that superior-quality wines typically include higher amounts of alcohol, lower volatile acidity, higher levels of sulfates, and higher levels of residual sugar.

Conclusion

Interest in the wine industry has grown recently, which begs for industrial expansion. To increase wine production and sales, corporations are investing in cutting-edge technologies. For each of these procedures, wine quality certification is essential and necessitates expert human wine testing. We utilized samples from the red wine dataset (RWD) with eleven distinct physiochemical properties. With the initial sample of RWD, vie ML models were trained and evaluated. We evaluated the effectiveness of the RF and XG Boost classifiers based on accuracy, recall, F1 scores, and support before introducing them as ML models to predict wine quality. Using these two ML methodologies, the top three features are chosen

from a total of eleven features, and ML analysis is performed on the other features. Various plots are used to represent the feature importance based on the XG Boost model and RF. Wine quality was predicted using significant characteristics (also known as essential

References

1. Wang, L., Cheng, Y. & Wang, Z. Risk management in sustainable supply chain: A knowledge map towards intellectual structure, logic diagram, and conceptual model. Environ. Sci. Pollut. Res. 29(44), 66041–66067. https:// doi. org/ 10. 1007/ S11356- 022- 22255-X (2022).
2. Loose, S. M. and Pabst, E. Current state of the German and international wine markets. he German and International Wine Markets. https:// www. resea rchga te. net/ publi cation/ 32340 2029_ Curre nt_ state_ of_ the_ German_ and_ inter natio nal_wine_marke ts (Accessed 29 December 2022) (2018).
3. Bansla, N., Kunwar, S. & Gupta, K. Social engineering: A technique for managing human behavior. J. Inf. Technol. Sci. https:// doi. org/ 10. 5281/ ZENODO. 25808 22 (2019).
4. Ingrassia, M. et al. Visitor's motivational framework and wine routes' contribution to sustainable agriculture and tourism. Sustain-ability 14(19), 12082. https:// doi. org/ 10. 3390/ SU141 912082 (2022).
5. Jain, K., Singh, A., Singh, P. & Yadav, S. An improved supervised classification algorithm in healthcare diagnostics for predicting upload habit disorder. Int. J. Reliab. Qual. E-Healthc. 11(1), 116. https:// doi. org/ 10. 4018/ IJRQEH. 297088 (2022).
6. Dev, V. A. & Eden, M. R. Gradient boosted decision trees for lithology classiication. Comput. Aided Chem. Eng. 47, 113–118. https:// doi. org/ 10. 1016/ B978-0- 12- 818597- 1. 50019-9 (2019).
7. Qian, H., Wang, B., Yuan, M., Gao, S. & Song, Y. Financial distress prediction using a corrected feature selection measure and gradient boosted decision tree. Expert Syst. Appl. 190, 116202. https:// doi. org/ 10. 1016/J. ESWA. 2021. 116202 (2022).
8. Cortez, P. et al. Using data mining for wine quality assessment. Lect. Notes Comput. Sci. (Incl. subser. Lect. Notes Artif. Intell. Lect. Notes Bioinform.) 5808, 66–79. https:// doi. org/ 10. 1007/ 978-3- 642- 04747-3_ 8/ COVER (2009).
9. Agrawal, G. & Kang, D.-K. Wine quality classiication with multilayer perceptron. Int. J. Internet Broadcast. Commun. 10(2), 25–30. https:// doi. org/ 10. 7236/ IJIBC. 2018. 10.2.5 (2018).
10. Aich, S., Al-Absi, A. A., Hui, K. L., Lee, J. T. and Sain, M. A classiication approach with diferent feature sets to predict the quality of diferent types of wine using machine learning techniques. In International Conference on Advanced Communication Technology, ICACT , vol. 2018, 139–143. https:// doi. org/ 10. 23919/ ICACT. 2018. 83236 74 (2018).
11. Gupta, Y. Selection of important features and predicting wine quality using machine learning techniques. Procedia Comput. Sci.125, 305–312. https:// doi. org/ 10. 1016/J. PROCS. 2017. 12. 041 (2018).
12. Kumar, S., Agrawal, K. and Mandan, N. Red wine quality prediction using machine learning techniques. In 2020 International Conference on Computer Communication and Informatics, ICCCI 2020. https:// doi. org/ 10. 1109/ ICCCI 48352. 2020. 91040 95 (2020)
13. Shaw, B., Suman, A. K. & Chakraborty, B. Wine quality analysis using machine learning. Adv. Intell. Syst. Comput. 937, 239–247. https:// doi. org/ 10. 1007/ 978- 981- 13- 7403-6_ 23/ COVER (2020).
14. Bhardwaj, P., Tiwari, P., Olejar, K., Parr, W. & Kulasiri, D. A machine learning application in wine quality prediction. Mach. Learn. Appl. 8, 100261. https:// doi. org/ 10. 1016/J. MLWA. 2022. 100261 (2022).
15. Tiwari, P. et al. Understanding quality of Pinot Noir wine: Can modelling and machine learning pave the way?. Foods 11(19), 3072.https:// doi. org/ 10. 3390/ FOODS 11193 072/ S1 (2022).
16. Mahima, U. G., Patidar, Y., Agarwal, A. & Singh, K. P. Wine quality analysis using machine learning algorithms. Lect. Notes Netw. Syst. 106, 11–18. https:// doi. org/ 10. 1007/ 978- 981- 15-2329-8_ 2/ COVER (2020).
17. Ma, X. et al. Rapid prediction of multiple wine quality parameters using infrared spectroscopy coupling with chemometric methods.J. Food Compos. Anal. 91, 103509. https:// doi. org/ 10. 1016/J. JFCA. 2020. 103509 (2020).

18. Prez-Magario, S. & Gonzalez-SanJose, M. L. Prediction of red and rosé wine CIELab parameters from simple absorbance measurements. J. Sci. Food Agric. 82(11), 1319–1324. https://doi.org/10.1002/JSFA.1191 (2002).
19. Corsi, A. & Ashenfelter, O. Predicting Italian wine quality from weather data and expert ratings. J. Wine Econ. 14(3), 234–251. https:// doi. org/ 10. 1017/ JWE. 2019. 41 (2019).
20. Croce, R. et al. Prediction of quality parameters in straw wine by means of FT-IR spectroscopy combined with multivariate data processing. Food Chem. 305, 125512. https:// doi. org/ 10. 1016/J. FOODC HEM. 2019. 125512 (2020).
21. Astray, G. et al. Prediction models to control aging time in red wine. Molecules 24(5), 826. https:// doi. org/ 10. 3390/ MOLEC ULES240508 26 (2019).
22. Mienye, I. D. & Sun, Y. A survey of ensemble learning: Concepts, algorithms, applications, and prospects. IEEE Access 10, 99129–99149. https:// doi. org/ 10. 1109/ ACCESS. 2022. 32072 87 (2022).