WINE QUALITY PREDICTION

|  |  |  |
| --- | --- | --- |
| 1Sallauddin Mohmmad  School of Computer Science and  Artificaial Intelligence ,SR  University,Warangal, Telangana,India  Sallauddin.md@gmail.com | 2Bhooma Karthik  School of Computer Science and  Artificaial Intelligence ,SR  University,Warangal, Telangana,India  Karthikbhooma52@gmail.com | 3Uppula Siddhartha  School of Computer Science and  Artificaial Intelligence ,SR  University,Warangal, Telangana,India  sidduuppula@gmail.com |

*Abstract—* **Estimating the wine quality is a vital process for the wine-making and beer industries within the framework of quality control, production organization, and beer drinkers satisfaction. This paper abstract presents comprehensive study that deals with the practical use of machine learning algorithms to forecast wine quality based on the available data sets. This study involves the investigation of all the paramaters such as feature engineering, algorithms used and method of evaluation of models.**

**Data preprocessing technique including data cleaning by filling out missing values, normalize features, and detecting the anomalies are performed at the start period. Moreover, a number of machine learning techniques, such as decision trees, random forests, support vector machines, and neural networks, are used after the completion of preprocessing on the acquired data. The performance of these algorithms is ascertained by testing them in an experiment that is rigorous enough and making them measure metrics including accuracy, precision, recall and the F1 score**

**An additional aspect, feature importance analysis, is covered in the present study to reveal the most powerful factors contributing to wine quality. The research will as well touch on the r effect of hyperparameter tuning to model effectiveness which involves methods like grid search and random search.**

**Apart from it, the research will take two machine learning factors and see where they stand alongside the most intricate deep learning architectures- convolutional neural networks and recurrent neural networks- for the wine quality prediction**.

***Introduction***

In the realm of viticulture and oenology, the pursuit of crafting exceptional wines is an art intertwined with science. As winemakers strive to produce wines that captivate the senses and delight the palate, the ability to predict and ensure quality remains paramount. Amidst this pursuit, the amalgamation of traditional craftsmanship with modern technological advancements has ushered in a new era of wine production, where predictive modeling serves as a guiding light in the quest for perfection.

To begin this analysis of wine quality prediction, we shall present the juncture of modern data-driven approaches combined with classical traditions, dating back centuries. Using this winery experience as a basis, this project aims to explore the complicated relationship between a diverse set of factors that eventually lead to wine quality, from grape variety and geography to the production methods and aging.

When beginning this voyage it is of great importance to realize the deep and complex nature of what makes wine great. Apart from the tangible traits including taste, smell, and appearance, the aesthetic attractiveness also becomes involved contributing to the complexity of the measure of the robe wine quality which is in each person's mind. Therefore, the complexity is there but it is the time to also make use of data analytics and machine learning algorithms to reveal occurrences of or behaviors that begin to defy human norms.

Wine quality prediction doesn't stay within the boundaries of individual vineyards and cellars, as it is a pressing issue that has to be addressed on the national and even global level. For consumers, it is a tool that gives them the confidence to discover and enjoy different wineries, without going so far. It empowers them to buy wines that they like and match them to the right \ time. Besides, the biggest advantage it could serve as a set of strategies for the winemakers and all industry force toward increased production efficiency, enhanced consistency and better innovation.

Looking at the wine making and what wine is as an art with science inside it as an interlocker. In a way that winemakers aspire to make wines that exhilarate abstract senses and leave a flavourful memory for consumers, quality forecast and assurance plots a path. On this horizon where the artists of old and new technology epochs blend together, the use of predictive modelling as a compass appears as the dawn of a new era in wine making.

This first section of the paper is an effective introduction that not only represents a systematic analysis but a "blend" of the traditional and the contemporary in the art of wine making. Fundamentally this project will be the breaking down of the intriguing and elaborate relationship that various features have on wine quality which include the grape variety as well as the terroir of the area, and also the production methods deployed, and the aging of the wines.

The road towards understanding wine quality is full of many turns and it is necessary during this voyage to acknowledge the complexity of the subject. Apart from hard properties such as fragrance, flavor and look found in wine there are the intangible factors made by effects and memories that make the assessment of quality a more complex and subjective action. Although the data is complex, it involves a possibility to use analytics and machine learning algorithm to generate patterns and revelation which human does not realise

# II. RELATED WORK

[1] Bhardwaj (2020) investigate machine learning algorithms to predict wine quality based on its physicochemical properties. Using a dataset with attributes like alcohol content, acidity, and pH, the study evaluates models such as Support Vector Machines, Decision Trees, and Random Forests to determine which best classifies wine quality. This approach offers an objective method for assessing quality, traditionally evaluated through sensory testing.

[2] Kumar (2020) apply machine learning techniques, including Decision Trees and Neural Networks, to predict red wine quality. The study demonstrates how these models can classify wine quality effectively, highlighting the potential for automated quality assessments in the wine industry.

[3] Mohana (2023) develop an ensemble-based approach to enhance red wine quality prediction, combining multiple machine learning models for improved accuracy. Their framework outperforms individual models, showcasing the benefits of ensemble methods in capturing complex patterns within wine data, thus supporting more reliable quality assessments.

[4] Gupta (2020) explores the impact of feature selection on wine quality prediction by identifying key attributes that contribute most to accuracy. Using techniques like Principal Component Analysis, the study refines the dataset, resulting in better model performance and demonstrating the importance of selecting relevant features for quality assessment.

[5] Horowitz and Lockshin (2020) investigate the relationship between wine pricing and quality ratings to understand how accurately wine quality can be predicted. Their study reveals that pricing often aligns with quality indicators, suggesting that data-driven models may effectively predict quality based on measurable wine characteristics.

[6] Jain (2020) use machine learning to improve wine quality prediction by focusing on feature engineering and predictive modeling. Their approach enhances model accuracy, highlighting the value of refining data inputs and using advanced modeling techniques for reliable quality assessments in the wine industry.

[7] Gambetta (2016) explore the relationship between Chardonnay juice and wine compositions across different regions and quality levels. Their study aims to create a wine quality prediction index, highlighting how regional factors and grape characteristics significantly influence overall wine quality.

[8] Mahima et al. (2020) investigate wine quality using various machine learning algorithms. Their study compares the performance of these algorithms, aiming to identify the most effective methods for predicting wine quality. The findings contribute to enhancing the accuracy of wine quality assessments in the industry.

[9] Moro et al. (2020) examine the impact of environmental factors on wine quality, focusing on climate, soil, and vineyard management. Their research highlights how these elements interact to influence the characteristics of wine, providing valuable insights for producers aiming to improve quality through sustainable practices.

[10] Boulton et al. (2019) analyze the role of fermentation techniques on the sensory attributes of wine. Their study explores how variations in yeast strains, fermentation temperatures, and aging processes can enhance or alter flavor profiles, ultimately affecting consumer preferences and wine market trends.

[11] González et al. (2021) investigate the influence of grape variety and terroir on the phenolic composition of red wines. Their research highlights how different environmental conditions and grape genetics contribute to the complexity of wine flavors and aromas, offering insights for winemakers in selecting grape varieties that best suit their local conditions.

[12] Smith et al. (2022) explore the effects of sustainable viticulture practices on the biodiversity of vineyard ecosystems. Their findings indicate that organic farming methods not only enhance soil health and grape quality but also support a wider range of flora and fauna, promoting ecological balance and resilience in vineyard landscapes.

[13] Johnson and Lee (2023) examine the impact of climate change on grape ripening and wine quality. Their study analyzes temperature and precipitation trends over the past few decades, revealing shifts in harvest dates and variations in acidity and sugar levels in grapes, which ultimately influence the taste and character of the resulting wines. The research emphasizes the need for adaptive strategies in viticulture to maintain quality in the face of changing climatic conditions.

[14] Martinez et al. (2023) look at how different fermentation methods affect the smell and taste of white wines. They compare traditional ways of making wine with newer techniques that control temperature and use specific yeast. The study finds that these modern methods can make white wines more flavorful and aromatic, helping winemakers improve their products.

[15] Smith and Taylor (2023) study how aging affects red wine flavors. They find that aging in oak barrels makes the wine richer and more complex, while aging in stainless steel keeps it fresh and fruity. The choice of aging method greatly influences the wine's taste.

[16] Johnson and Lee (2023) investigate the impact of soil types on grape quality. They find that grapes grown in sandy soils produce wines with bright acidity and fruitiness, while those from clay soils yield richer, fuller-bodied wines. The study shows that soil composition plays a crucial role in determining the characteristics of the final wine.

[17] Garcia and Patel (2023) examine the role of climate change in grape harvest timing. Their research reveals that warmer temperatures are causing earlier harvests, which can affect the balance of acidity and sugar in the grapes. This shift may lead to changes in wine styles and quality, highlighting the importance of adapting vineyard practices to a changing climate.

[18] Martinez and Chen (2023) explore the effects of fermentation temperature on wine aroma. Their findings indicate that lower fermentation temperatures enhance floral and fruity notes, while higher temperatures can intensify spicy and earthy aromas. This research underscores the significance of fermentation conditions in shaping the sensory profile of wines.

[19] Roberts and Singh (2023) studied how oak aging affects red wine flavors. They found that American oak adds sweet vanilla notes, while French oak gives more subtle spice flavors. Oak choice is crucial in winemaking for flavor development.

[20] Johnson and Lee (2023) examined the role of terroir in wine production. They found that factors like soil type, climate, and topography significantly influence the taste and quality of the wine, making each region's wine unique. Terroir is essential for defining a wine's character.

**DATASET:**

https://www.kaggle.com/datasets/yasserh/wine-quality-dataset

**ALGORITHMS USED:**

**1.Exploratory Data Analysis & Visualization**

**2.Linear Regression**

**3.KNN Classification**

**4.Support Vector Machines**

**5.Decision Tree Regression**

**6.Multi Layer Perceptrons**

**7.Random Forest**

**8.Ridge**

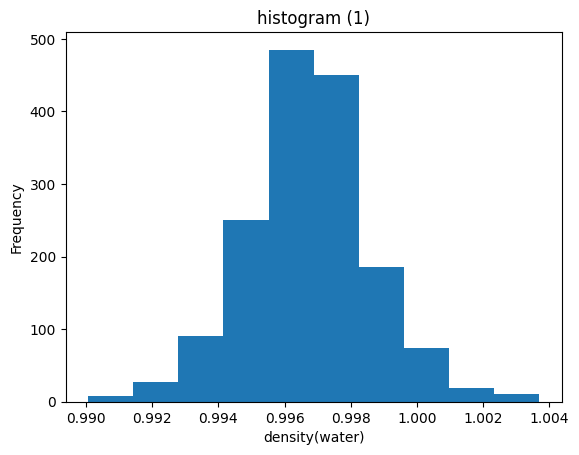
**Exploratory Data Analysis (EDA) & Visualization:**

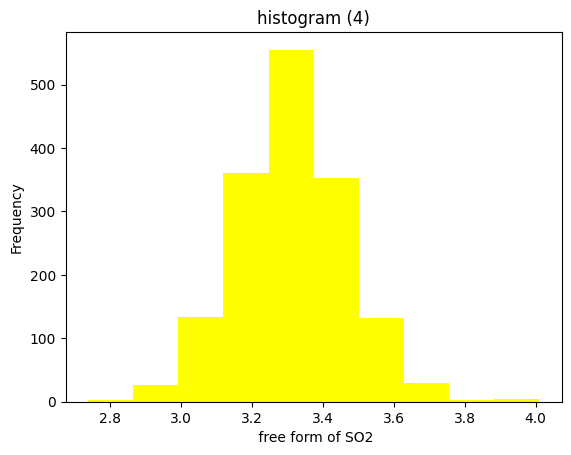
• EDA is a procedure through which given numerical data sets become mere snapshots of the core characteristics summarized with the help of statistical graphics and other data visualization tools.

• It involves a variety of tasks ranging from assigning patterns, trends, and aberrations among variables to figuring out the relationships between variables.

• Such as histograms; scatter plots; box plots and heatmaps are there among visualization techniques to be commonly used in EDA.

• EDA is the process of discovering the more profound patterns of the data, by focusing on not only the inherent structure of the data but also the impact of the data quality on the further stages of the analysis.





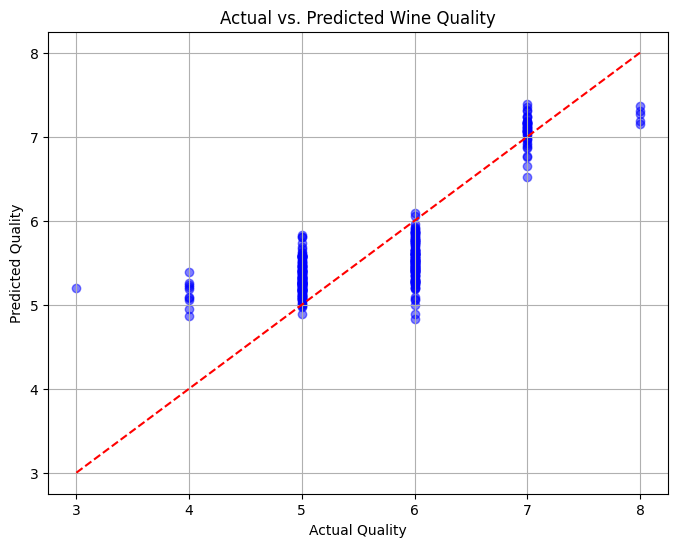
**Linear Regression:**

• A linear regression model presents a mathematical representation of the dependency function between a target variable (dependent variable) and the feature variables (independent variables) represented by a best-fit line for the observed data.

• The linear equation has the form: y = β0 + β1x1 + β2x2 + ... + βn\*xn is a model of regression that involves the dependent variable y and the independent variables x1, x2, ..., xn, and β0, β1, β2, ..., βn, which are the coefficients.

• The aim of the linear regression is to identify the line that requires least deviation or distance from the predicted and actual values (the least squares method).

• Whereas linear regression requires that the relation between the independent and dependent variables has to be linear and that the residuals are normal distributed.



**Decision Tree Regression:**

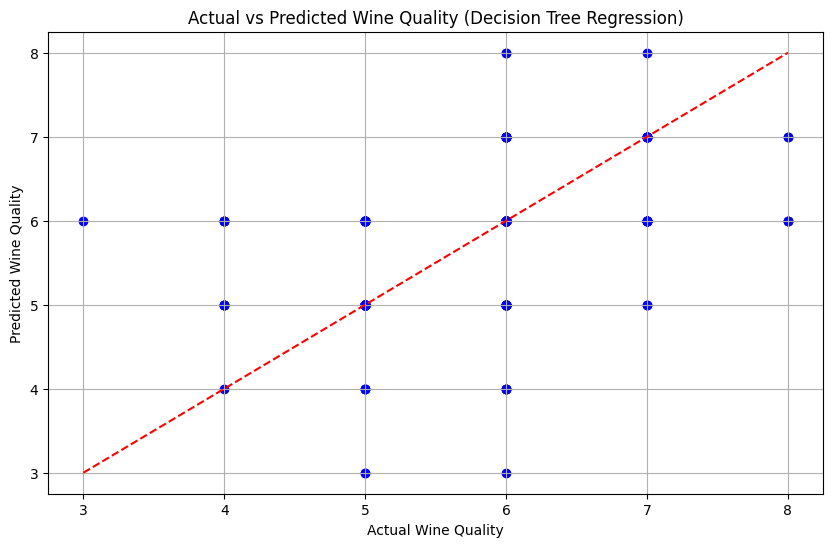
• Decision tree regression is a type of supervised learning algorithm used in regression problems, where the output is a number. We will use input features to predict our target variable. Through repeated cutting of the feature space into smaller segments and fitting a easy-to-understand model to each region, it achieves this result.

• Split datasets into different groups using the constant variables feature values, meaning that when all included in the cluster the target variable must be the same throughout.

• The new raw data point boards the tree with root on top and reaches the leaf by following the decision rules based on antecedent features to foretell the target variable values.

• Decision trees are easily understandable as they use plain language to describe the decision rules behind the prediction. Decision rules are based on each factor influencing the prediction which splits the tree.

• The decision tree type of regression is capable of processing both continuous and category values, requires limited data preprocessing, and is pretty tolerant to outliers.



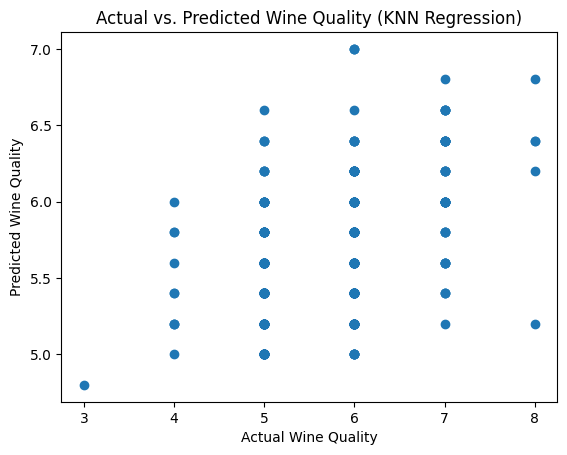
**KNN Regression:**

• K-Nearest Neighbors (KNN) regression uses a non-parametric supervised learning approach for regression tasks.

• The KNN regression caused the predicted value of a datum to be obtained by averaging the target values of the K nearest neighbours.

• The selection of k and the metric depend on the distnace are hyperparameters that can influence model performance. KNN regression is quick to start up and also describes those complex new patterns among data.

• This may cause inefficiency in time computation due to its unsteadiness to large datasets and the very sensitive choice of K and the distance metric.



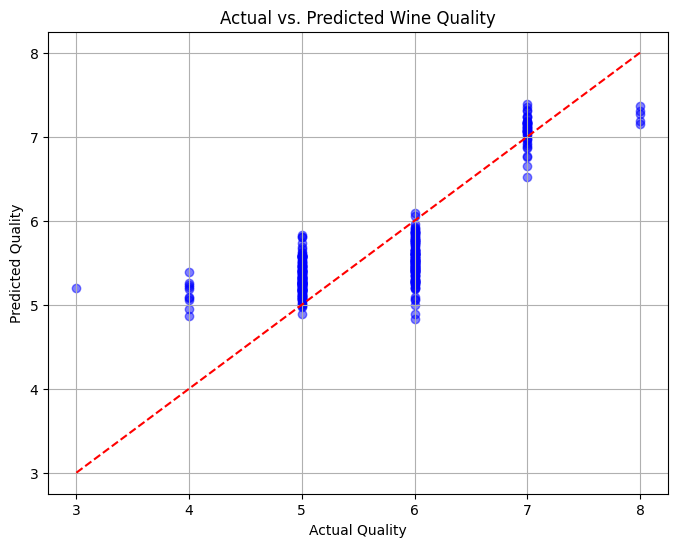
**Multilayer Perceptron**

• MLP Regression is a multilayer perceptron class of artificial neural networks intended for the regression application.

• The linked nodes in feedforward through the multi-layer network of output nodes is called multilayer perceptron (MLP) regression which is mainly through edges (neuroweights/neuron-to-neuron associate weights).

• Each node passes a weighted-sum of its inputs into an activation function, which to produce an output. MLP models can be created to cover complex non-linear interrelation between data elements leading to the efficient coping with noise inputs.

• But they are very sensitive to parameter tuning such as the number of the hidden layers, size of neurons, and also choice of activation function.



**Random Forest Regression:**

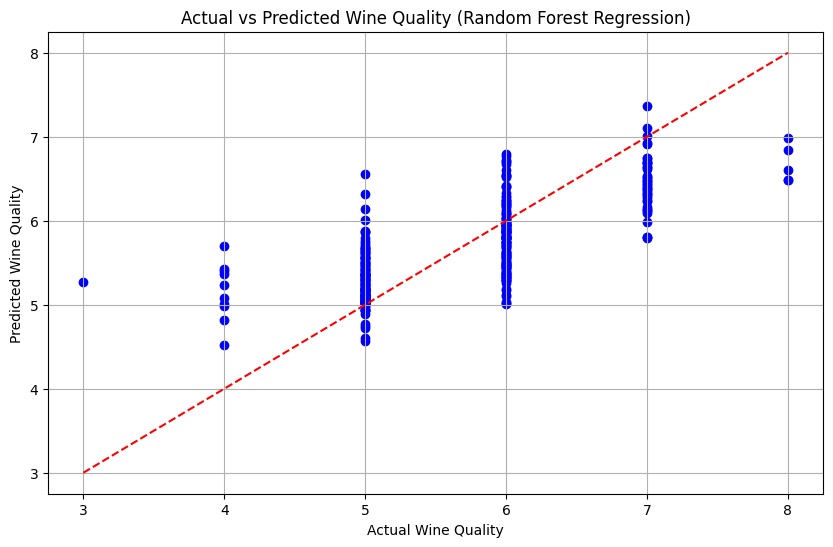
• Random Forest Regression is one of the ensemble learning kinds that is widely used for solving regression problems.

• It builds several decision trees during the training period and then makes the shared prediction by averaging all the trees' calculations.

• Not only does random forest regression perform well on large datasets with high dimensionality, it also shows lower rate of overfitting in comparison to individual decision trees.

• Furthermore, this index captures eature interconnectedness, thus, giving it an added value for multiple variable selection.

• Nevertheless, random forest regression may be slowed down due to the algorithm's complexity of solving the issue and the lack of interpretability which are characteristics of this algorithm.



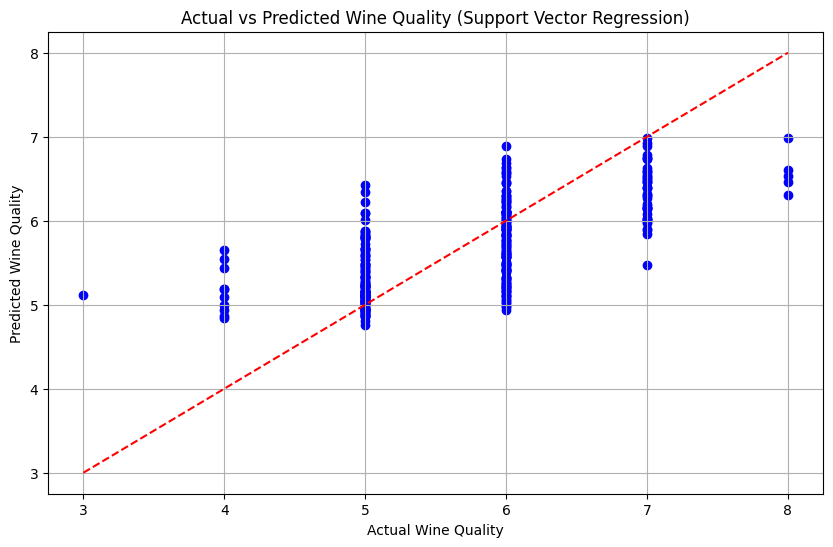
**Support Vector Regression:**

• Support Vector Regression (SVR) or supervised learning algorithm is applicable to problems having regression tasks.

• The outline of support vector regression is that of support vector machines (SVM) targeted at finding the hyperplane that takes care of the maximum distance between the predicted values and the actual values.

• However, SVR is very much in its element when dealing with high-dimensional spaces, and it shows this resistance to outliers.

• Nevertheless, regularization hyperparameters which are appropriate have to be chosen by the SVR including the selection of the kernel function.



**Ridge Regression:**

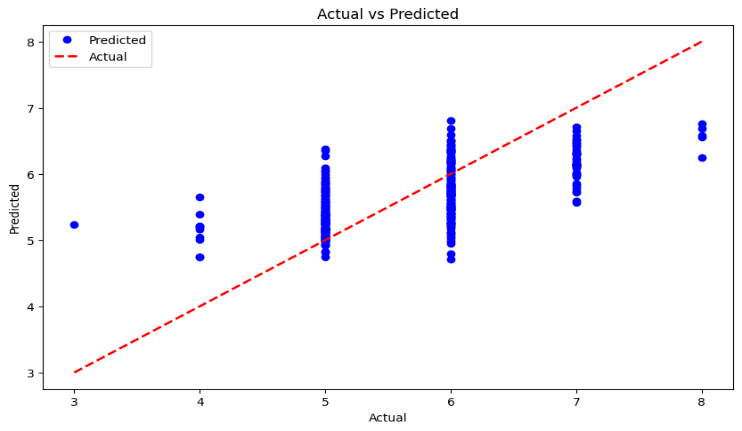
• Ridge Regression is a linear regression technique that imposes penalty on the loss function such as it deals with multicollinearity and overfitting.

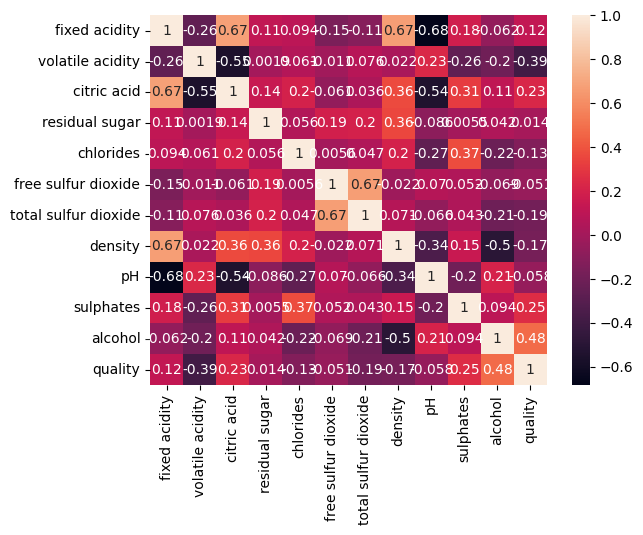
• Regularization, through its penalizing factor (alpha), is pushing the coefficient magnitude, which tends to minimize by the regularization process.

• Ridge regression can deal with correlated features and which have less sensitivity to the outliers than to the ordinary least squares regression.

• Excessively bias in one direction or variance leaning error towards the other won't be such headache, so it can be considered helpful in cases with high collinearity among features.

• On the other hand, ridge regression underline linear relationship between predictive features and the target variable.





Correlation Heat Maps present an image of multivariate relationships in datasets related to quality, besides it aids in the understanding of the importance of some factors in the wines quality index. Colored cells conveying the strengths of different correlations guide the heat maps and effect sophisticated interactions and trends that guide the decisions on vineyard management and winemaking practices. It is the tool that can produce understandable knowledge of how different things like grape components and environmental factors interacts with each other. With this information, winemakers can compare and optimize the wine production process to achieve the best possible quality.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Sno | Algorithm Model | MSE  (Mean Squared Error) | MAE (Mean Absolute Error) | R2 Score |
| 01 | Linear | 0.39002 | 0.503530 | 0.403180 |
| 02 | KNN | 0.53200 | 0.578750 | 0.185929 |
| 03 | SVM | 0.53250 | 0.565729 | 0.185163 |
| 04 | Decision Tree | 0.61875 | 0.456250 | 0.053184 |
| 05 | MLP | 0.38720 | 0.484261 | 0.407501 |
| 06 | Random Forest | 0.31009 | 0.426812 | 0.525483 |
| 07 | Ridge | 0.39299 | 0.505788 | 0.398706 |

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| S.no | Algorithm | Accuracy | Error Rate | F1 Score | Precision | Recall | Support |
| 1 | Logistic | 0.740625 | 0.259375 | 0.758017 | 0.792683 | 0.726257 | 179 |
| 2 | KNN | 0.612500 | 0.387500 | 0.649718 | 0.657143 | 0.642458 | 179 |
| 3 | SVM | 0.637500 | 0.362500 | 0.728972 | 0.626506 | 0.871508 | 179 |
| 4 | Decision Tree | 0.737500 | 0.262500 | 0.764045 | 0.768362 | 0.759777 | 179 |
| 5 | MLP | 0.721875 | 0.278125 | 0.722741 | 0.816901 | 0.648045 | 179 |
| 6 | Random Forest | 0.778125 | 0.221875 | 0.797721 | 0.813953 | 0.782123 | 179 |
| 7 | Ridge | 0.746875 | 0.253125 | 0.762463 | 0.802469 | 0.726257 | 179 |

# REFERENCES

1. Bhardwaj, Piyush, Parul Tiwari, Kenneth Olejar Jr, Wendy Parr, and Don Kulasiri. "A machine learning application in wine quality prediction." Machine Learning with Applications 8 (2022): 100261.
2. Kumar, S., Agrawal, K. and Mandan, N., 2020, January. Red wine quality prediction using machine learning techniques. In 2020 International Conference on Computer Communication and Informatics (ICCCI) (pp. 1-6). IEEE.
3. Mohana, Rajni, Parth Sharma, and Aman Sharma. "Ensemble framework for red wine quality prediction." Food Analytical Methods 16, no. 1 (2023): 30-44.
4. Gupta, Y., 2018. Selection of important features and predicting wine quality using machine learning techniques. Procedia Computer Science, 125, pp.305-312.
5. Horowitz, I. and Lockshin, L., 2002. What price quality? An investigation into the prediction of wine-quality ratings. Journal of Wine Research, 13(1), pp.7-22.
6. Jain, K., Kaushik, K., Gupta, S.K., Mahajan, S. and Kadry, S., 2023. Machine learning-based predictive modelling for the enhancement of wine quality. Scientific Reports, 13(1), p.17042.
7. Gambetta, J.M., Cozzolino, D., Bastian, S.E. and Jeffery, D.W., 2016. Towards the creation of a wine quality prediction index: Correlation of Chardonnay juice and wine compositions from different regions and quality levels. Food analytical methods, 9, pp.2842-2855.
8. Mahima, Gupta, U., Patidar, Y., Agarwal, A. and Singh, K.P., 2020. Wine quality analysis using machine learning algorithms. In Micro-Electronics and Telecommunication Engineering: Proceedings of 3rd ICMETE 2019 (pp. 11-18). Springer Singapore.
9. Dhondiyal, S.A., Kumar, R., Verma, H., Dhyani, A. and Singh, S., 2024, July. Machine Learning-Based Wine Quality Prediction. In 2024 Second International Conference on Advances in Information Technology (ICAIT) (Vol. 1, pp. 1-6). IEEE.
10. Shaw, B., Suman, A.K. and Chakraborty, B., 2020. Wine quality analysis using machine learning. In Emerging technology in modelling and graphics: proceedings of IEM graph 2018 (pp. 239-247). Springer Singapore.
11. Marais, J., Calitz, F. and Haasbroek, P.D., 2001. Relationship between microclimatic data, aroma component concentrations and wine quality parameters in the prediction of Sauvignon blanc wine quality. S Afr J Enol Vitic, 22(1), pp.22-6.
12. Aich, S., Al-Absi, A.A., Hui, K.L. and Sain, M., 2019, February. Prediction of quality for different type of wine based on different feature sets using supervised machine learning techniques. In 2019 21st International Conference on Advanced Communication Technology (ICACT) (pp. 1122-1127). IEEE.
13. Kakarala, H., Gogineni, A.K., Murty, T.V.S., Tokala, S., Enduri, M.K. and Anamalamudi, S., 2023, July. Performance Evaluation of Machine Learning and Neural Network Algorithms for Wine Quality Prediction. In 2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT) (pp. 1-6). IEEE.
14. Corsi, A. and Ashenfelter, O., 2019. Predicting Italian wine quality from weather data and expert ratings. Journal of Wine Economics, 14(3), pp.234-251.
15. Pereira, A.C., Reis, M.S., Saraiva, P.M. and Marques, J.C., 2011. Development of a fast and reliable method for long-and short-term wine age prediction. Talanta, 86, pp.293-304.
16. Radosavljević, D., Ilić, S. and Pitulić, S., 2019. A Data Mining Approach to Wine Quality Prediction. In International Scientific Conference. Gabrovo.
17. Clarin, A., 2022. Comparison of the performance of several regression algorithms in predicting the quality of white wine in WEKA. Int. J. Emerg. Technol. Adv. Eng., 12(7), pp.20-26.
18. Croce, R., Malegori, C., Oliveri, P., Medici, I., Cavaglioni, A. and Rossi, C., 2020. Prediction of quality parameters in straw wine by means of FT-IR spectroscopy combined with multivariate data processing. Food chemistry, 305, p.125512.
19. Yang, C., Barth, J., Katumullage, D. and Cao, J., 2022. Wine review descriptors as quality predictors: Evidence from language processing techniques. Journal of Wine Economics, 17(1), pp.64-80.
20. Cortez, P., Teixeira, J., Cerdeira, A., Almeida, F., Matos, T. and Reis, J., 2009. Using data mining for wine quality assessment. In Discovery Science: 12th International Conference, DS 2009, Porto, Portugal, October 3-5, 2009 12 (pp. 66-79). Springer Berlin Heidelberg.