Cureus Journal of Computer Science

Open Access Original Article

Alcohol Quality Analysis Using Machine Learning Regression Technique

Aditya S. Baheti [1] , Ankush D. Sawarkar [1] , Ubaid A. Shaikh [1] , Deepti D. Shrimankar [2]

1. Information Technology, Shri Guru Gobind Singhji Institute of Engineering and Technology, Nanded, IND 2. Computer Science and Engineering, Visvesvaraya National Institute of Technology, Nagpur, IND

Corresponding authors: Aditya S. Baheti, 2021bit002@sggs.ac.in, Ankush D. Sawarkar, adsawarkar@sggs.ac.in

Review began 08/18/2024
Review ended 11/06/2024
Published 11/12/2024

DOI: https://doi.org/10.7759/s44389-024-00227-1

Abstract

This research paper investigates the predictive modeling of alcohol quality machine learning regression analysis. The study employs a comprehensive approach to preprocessing, modeling, and evaluation to achieve accurate predictions. Initial data exploration reveals the dataset's structure and identifies potential issues such as missing values and outliers. Preprocessing steps include outlier removal using the interquartile range method and feature scaling to ensure uniformity. The dataset is split into two parts, i.e., training and testing, for development and evaluation. Regression models are built using the ordinary least squares method, incorporating features transformed to address skewness and scaled to improve model performance. Model diagnostics are conducted to assess assumptions including multicollinearity and normality of residuals. The final model exhibits strong predictive capability, as evidenced by high R-squared values and low mean squared error. Visualization techniques illustrate the relationship between actual and predicted values, providing insight into model accuracy. Overall, this paper demonstrates the effectiveness of linear regression in predicting alcohol content with an accuracy of 92%, offering valuable implications for quality control and production optimization in the alcohol industry along with other regression models such as gradient boosting with an accuracy of 90% and decision tree regressor with an accuracy of 83%.

Categories: AI applications, Data Engineering, Machine Learning (ML)
Keywords: wine quality, machine learning, linear regression, decision tree regressor, gradient boosting

Introduction

Machine learning regression is a statistical technique used for predicting continuous numerical outcomes based on input variables. In predictive modeling, regression analysis plays an important role in understanding and forecasting various phenomena across different domains. Its significance lies in uncovering relationships between variables and making valid predictions, thereby aiding decision-making processes. Machine learning regression models, including linear, ridge, and lasso regression, are widely employed in diverse fields such as finance, healthcare, marketing, and engineering to forecast outcomes like stock prices, patient health outcomes, consumer behavior, and product performance [1]. The emergence of machine learning has revolutionized regression analysis by enabling the progress of more complex and flexible models capable of handling large datasets and capturing nonlinear relationships. In today's data-driven world, where organizations are provided with huge amounts of data, leveraging regression analysis for predictive modeling is invaluable for gaining deeper insights, making informed decisions, and driving innovation and growth.

Wine quality is of paramount importance in the beverage industry due to its widespread demand and competitive market landscape. Ensuring wine quality is crucial for both consumers and producers, prompting the adoption of proactive quality control measures throughout the production cycle. Technological advancements have facilitated various methodologies for quality testing, resulting in significant time and budget efficiency while enhancing product quality [2]. This research aims to develop and evaluate regression models for forecasting alcohol content in wine datasets, assess preprocessing techniques' effectiveness, investigate predictive capabilities, explore regression analysis feasibility for quality prediction in the alcohol industry, and offer practical recommendations for producers. The study begins with a comprehensive literature review, followed by methodology, experimental setup, results and discussions, and conclusion sections, providing insights into the practical implications for wine producers.

Reviewing the work done on alcohol quality prediction using machine learning regression reveals a growing body of research exploring the predictive capabilities of regression models in assessing alcohol quality based on physiochemical properties [1]. Studies have employed various algorithms, which includes linear regression, support vector, and random forest regression, to forecast wine attributes such as alcohol content, acidity levels, and sensory scores [2]. These algorithms have been used on datasets containing diverse wine samples from various regions and varieties, highlighting the importance of considering factors such as grape variety, climate, and production methods in wine quality assessment. However, gaps in existing literature include limited exploration of advanced regression techniques tailored specifically for wine quality prediction and a lack of emphasis on the interpretability and robustness of regression models in the context of wine production (Table *1*). The current study focuses on addressing these gaps by leveraging

advanced regression methodologies, rigorous data preprocessing techniques, and comprehensive model evaluation strategies to improve the accuracy and reliability of wine quality prediction models, thereby contributing to the advancement of wine production practices and quality assurance measures.

Materials And Methods

Authors	Methodology	Key Findings	Ref.
Dahal et al. (2021)	This work illustrates how statistical investigation can be utilized to distinguish the components that mainly control the wine quality prior to the generation. This will aid wine producing company to control the quality of the wine before production.	This work shows an elective approach that could be utilized to actuate the wine quality, and thus it can be an incredible beginning point to screen the variables on which the quality of alcohol is dependent.	[2]
Cortez et al. (2009)	This study proposes a information mining approach to foresee human wine taste preferences that is based on effectively accessible, analytical tests at the final step. A huge dataset is considered, with white and red Vinho Verde tests.	The proposed data-driven approach is on the objective tests, and therefore it can be coordinated into a choice back framework, helping the speed and quality of the oenologist execution. The show seems to be utilized to move forward the preparing of oenology students.	[3]
Satyabrata, et al. (2018)	In this paper a new approach has been proposed by considering diverse feature selection calculation such as principal component analysis as well as recursive feature elimination approach (RFE) approach for feature choice and nonlinear choice tree-based classifiers for analyzing the execution measurements.	In this study, we have utilized nonlinear classifiers to foresee the quality of two types of wines by achieving great classification accuracies extending from 94.51% to 97.79%. Irregular woodland classifier shows the highest precision of 94.51% when anticipating the quality of red wine with feature based on RFE sets; simultaneously, the classifier shows highest precision of 97.79% when foreseeing the quality of wine with RFE-based highlight sets.	[4]
Bhardwaj et al. (2022)	In this research, the major objective is to foresee wine quality by producing engineered information and develop a machine learning show based on this engineered information and accessible experimental information collected from diverse and differing locales over New Zealand.	This paper actualized a Smote algorithm utilizing 12 tests, and remaining tests were utilized for testing. We presented several feature related scenarios in this research to improve models' execution. We tried machine learning model without feature choice, feature chosen utilizing XGBoost, random forest, gradient boosting, extra trees classifier, and essential variables.	[5]

TABLE 1: Summary of work done for alcohol quality analysis using regression technique.

In this research, the dataset used is the openly available wine quality dataset that is sourced from the GitHub repository. This repository hosts a vast collection of datasets widely employed by the machine learning community. We opted for the white wine data over red wine, given its prevalent usage. The white wine data comprise many physiochemical properties: fixed acidity, volatile acidity, sulfur dioxide, chlorides, pH level, density, etc. In addition to these properties, sensory scores were collected from multiple taste testers who rated each wine sample on a scale of 0 to 10. The median score was recorded and serves as the response variable. Various statistical analyses were conducted to comprehend the dataset's characteristics, as outlined in Figure *1*.

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
count	4898.000000	4898.000000	4898.000000	4898.000000	4898.000000	4898.000000	4898.000000	4898.000000	4898.000000	4898.000000	4898.000000	4898.000000
mean	6.854788	0.278241	0.334192	6.391415	0.045772	35.308085	138.360657	0.994027	3.188267	0.489847	10.514267	5.877909
std	0.843868	0.100795	0.121020	5.072058	0.021848	17.007137	42.498065	0.002991	0.151001	0.114126	1.230621	0.885639
min	3.800000	0.080000	0.000000	0.600000	0.009000	2.000000	9.000000	0.987110	2.720000	0.220000	8.000000	3.000000
25%	6.300000	0.210000	0.270000	1.700000	0.036000	23.000000	108.000000	0.991723	3.090000	0.410000	9.500000	5.000000
50%	6.800000	0.260000	0.320000	5.200000	0.043000	34.000000	134.000000	0.993740	3.180000	0.470000	10.400000	6.000000
75%	7.300000	0.320000	0.390000	9.900000	0.050000	46.000000	167.000000	0.996100	3.280000	0.550000	11.400000	6.000000
max	14.200000	1.100000	1.660000	65.800000	0.346000	289.000000	440.000000	1.038980	3.820000	1.080000	14.200000	9.000000

FIGURE 1: Descriptive statistics of the white wine data.

System architecture and workflow

During the data preprocessing stage, we first loaded the dataset from the 'alcoholcontentquality.csv' file using Pandas. The flowchart of system architecture is provided in Figure *2* for visual understanding of the process. To identify missing values, we visualized the dataset using the missingno.matrix() function. Outliers were then detected using the interquartile range (IQR) method, with those lying beyond 1.5 times the IQR removed (Figure *3*). Next, we applied min-max scaling to selected features such as 'citric acid', 'density', 'residual sugar', and 'sulphates' using the sklearn.preprocessing.minmax_scale function. Skewed features, specifically 'residual sugar' and 'sulphates', underwent log transformation to improve their distribution, with new columns created to store the transformed features. To identify multicollinearity, we calculated the variance inflation factor (VIF) for the features (Figure *4*). Additionally, we checked for statistical assumptions such as linearity, homoscedasticity, and normality of residuals using ordinary least squares regression. Skewness and kurtosis were also calculated for selected features using the skew() function.

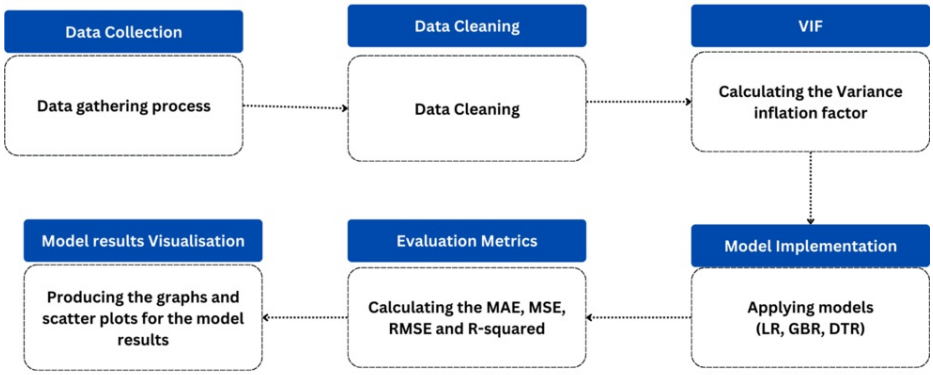

FIGURE 2: Flowchart of model implementation and result analysis.

VIF: variance inflation factor; MAE: mean absolute error; MSE: mean squared error; RMSE: root mean square error; LR: linear regression; GBR: gradient boosting regression; DTR: decision tree regression

The dataset comprises a continuous target variable (alcohol quality) and multiple physiochemical properties of wine. Regression models are applicable for datasets where the relationship between the target variable and features can be approximated by a linear function [1].

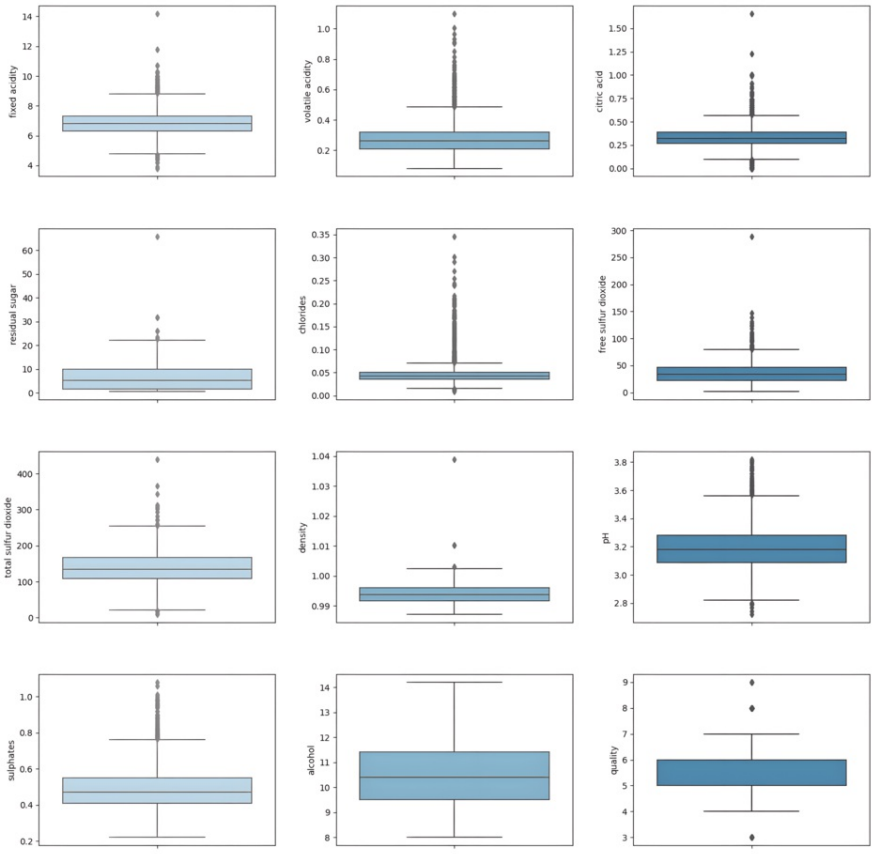

FIGURE 3: Box plot of the variables of the white wine data.

Linear regression models allow transparent interpretations of the link between independent variables and the target variable, crucial for understanding the impact of physiochemical properties on alcohol content [1]. VIF identifies multicollinearity among features, with steps taken to address it, ensuring the reliability of regression coefficients.

Linear regression coefficients signify the importance of features in predicting the target variable, aiding in identifying significant physiochemical properties affecting alcohol content in wine. Performance metrics such as R-squared (R2), mean squared error (MSE), and mean absolute error (MAE) assess model performance, providing baseline metrics for differentiation with more advanced models.

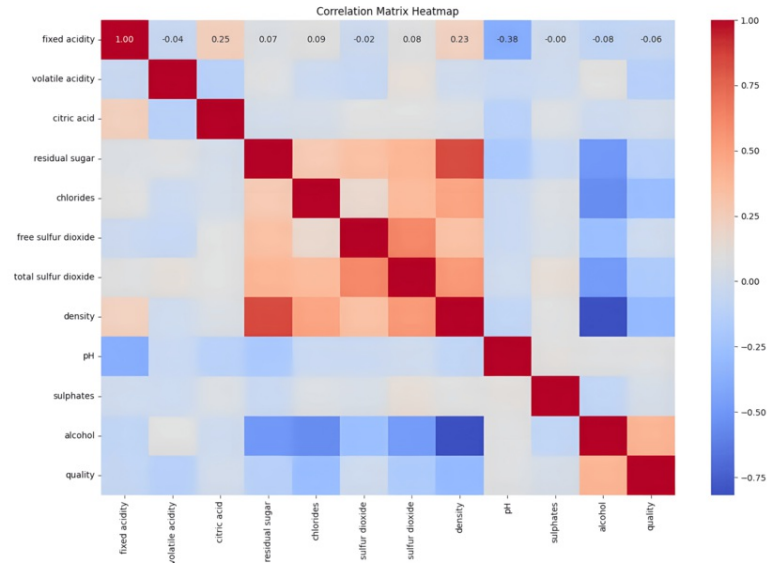

FIGURE 4: Heatmap of collinearity between the features of dataset.

R2: R2 demonstrates the quantity of variance in the response variable (alcohol quality), which is shown by the predictor variable [6].

R2 values lie in the range of 0 to 1. R2 value of 1 indicates that the model correctly predicts the target variable based on the predictor variables [6]. Equation 1 shows the formula for the calculation of R2, where SSR is the sum of squared residuals, and SST is the total sum.

$$R2 = \sum \frac{(1 - SSR)}{(1 - SST)} \qquad (1)$$

MSE: It evaluates the average squared difference between the actual values and the predicted values of the target variable [6].

Smaller values of MSE show better model performance, with 0 being the value representing perfect predictions [6]. Equation 2 shows the mathematical formula for the calculation of MSE.

$$MSE = \sum \frac{(x - \hat{x})^2}{n} \qquad (2)$$

MAE: It evaluates the average absolute difference between the actual and predicted values of the target variable [6].

Smaller values of MAE show better model performance, with 0 being the ideal value representing perfect predictions [6]. Equation 3 shows the mathematical formula for the calculation of MAE.

$$MAE = \sum \frac{x - \hat{x}}{n} \qquad (3)$$

The model is implemented using the Python programming language, using library scikitlearn for implementing decision tree regressor, linear regression, and gradient boost regressor models. The implementation does not require significant computational resources and can run on standard personal computers or laptops. With a moderate dataset size of 4,898 samples, computational requirements are reasonable for the tasks performed. Any Python-compatible integrated development environment or text editor, such as Jupyter Notebook, PyCharm, and VSCode, is suitable for coding. The code can be executed locally or in a cloud-based environment with Python support. Ensure that all necessary Python packages and dependencies are installed using pip or conda for successful code execution. K-fold cross-validation is employed, dividing the data into k equal-sized folds for training and validation. The model gets trained for k times, each time using k-1 folds as data for training and the remaining fold as data for validating the prediction. Evaluation metrics are averaged from all results to obtain the final performance metrics [7].

Results

The experiments conducted in the study yielded significant results, showcasing the predictive performance of various regression models such as linear regression, gradient boosting, and decision tree regression in estimating alcohol quality based on physiochemical properties of white wine samples (Figure *5*). Performance metrics such as R2, MAE, and MSE were utilized for evaluating model performance. Comparative analysis for various models such as linear regression, decision tree regressor, and gradient boost regressor provided insights into their effectiveness in capturing the relationship between independent variables and alcohol content. The results demonstrated that linear regression models exhibited satisfactory performance, with acceptable R2 value of 0.92114 and low error metrics of value: 0.2623 MAE and 0.112255 MSE (Table *2*). Moreover, the comparative analysis highlighted the differences between model complexity and predictive accuracy, with more simple linear regression performing competitively against more complex alternatives. These findings underscored the suitability of linear regression for predicting alcohol quality in white wine samples based on physiochemical properties, offering valuable insights for wine quality assessment and production optimization.

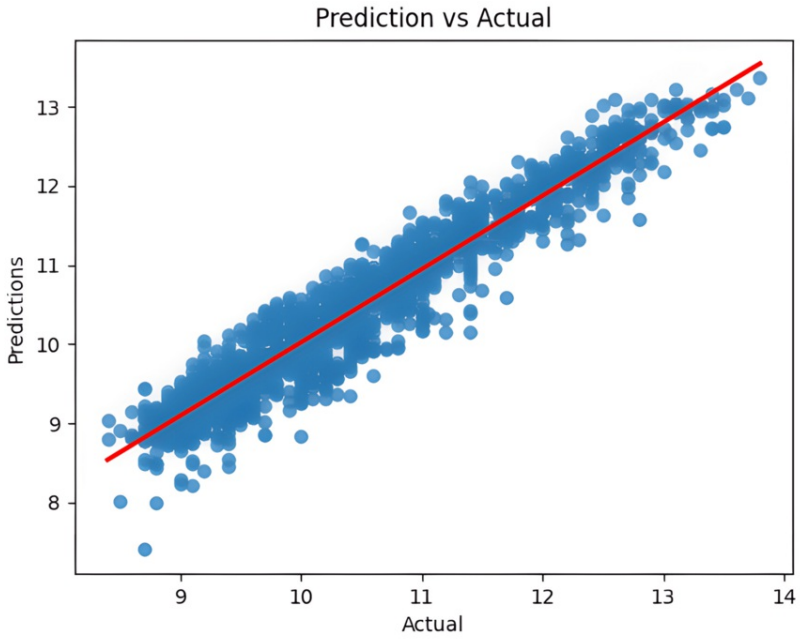

FIGURE 5: Actual vs. prediction plot for linear regression.

Regression Model	Mean Squared Error	R-squared	Mean Absolute Error
Linear Regression	0.1122	0.9211	0.2623
Decision Tree Regressor	0.8324	0.8324	0.3054
Gradient Boosting Regressor	0.1308	0.9080	0.2803

TABLE 2: Evaluation results of models implemented.

Discussion

The results presented in the table illustrate the performance of different regression models in predicting alcohol quality in white wine samples. The linear regression model achieved the lowest MSE (0.1122) and the highest R2 value (0.9211), indicating that it explains approximately 92% of the variance in alcohol content. This model also recorded the lowest MAE (0.2623), suggesting its predictions are closest to the actual values compared to the other models. The superior performance of linear regression underscores its effectiveness in capturing the relationship between physiochemical properties and alcohol quality, aligning with the high accuracy reported in the study. In contrast, the decision tree regressor, with an MSE of 0.8324 and an R2 value of 0.8324, demonstrates lower predictive accuracy and explanatory power. The gradient

boosting regressor, while offering an R2 value of 0.9080 and an MSE of 0.1308, does not surpass the linear regression model in overall performance.

These results underscore the robustness of linear regression models for predicting alcohol quality in wine and validate the choice of this technique in the study. The comparison with decision tree and gradient boosting algorithms highlights the linear regression model's advantages in terms of interpretability and predictive accuracy, reinforcing its utility in wine quality assessment. The findings suggest that while more complex models like gradient boosting can offer competitive performance, the simplicity and clarity of linear regression provide substantial benefits for practical applications and decision-making in the wine industry. Overall, the results validate the study's hypothesis and demonstrate the effectiveness of linear regression in predicting alcohol quality based on physiochemical properties.

Comparing our findings with previous studies reveals both similarities and differences in approaches and outcomes. Like prior research, our study emphasizes the significance of physiochemical properties in predicting alcohol quality in wine, particularly highlighting the influence of acidity, sulfur dioxide levels, and alcohol percentage [8]. These outcomes align with previous literature, which underscores the importance of these factors in wine quality assessment [6]. However, our research contributes insights by employing rigorous data preprocessing techniques and implementing linear regression models, demonstrating satisfactory predictive performance and enhancing interpretability. Additionally, while previous research has explored similar relationships between physiochemical properties and alcohol content, our research extends the analysis by leveraging cross-validation techniques to ensure model robustness and validity [5]. Overall, while our findings corroborate existing knowledge, our methodology and approach offer important contributions to the field, enhancing the understanding and application of regression analysis in wine quality assessment.

The study makes notable contributions to the field of machine learning regression by showcasing the application of regression analysis in predicting alcohol quality in white wine samples based on physiochemical properties using three regression models. Employing rigorous data preprocessing techniques underscores the importance of data quality and preparation in regression modeling, offering insights into best practices for handling datasets with multiple features. Furthermore, the study highlights the interpretability of linear regression models, providing transparent interpretations of relationships between independent variables and the target variable [1]. The utilization of cross-validation techniques ensures the robustness of the regression models, contributing to advancements in methodologies for evaluating model performance and generalization to unseen data [9]. These contributions underscore the significance of regression analysis in addressing real-world problems and advancing the recognition and use of machine learning techniques across various domains.

Building upon these insights, future studies could explore broader wine types and regions to enhance generalizability, investigate advanced regression techniques for capturing complex relationships, integrate additional data sources for deeper insights into wine quality factors, and conduct longitudinal studies to assess the impact of evolving production methods [7,10]. These avenues hold promise for advancing regression analysis in wine quality assessment and production optimization, contributing to industry practices and consumer satisfaction.

Conclusions

The research outcomes offer significant insights into the relationship between physiochemical properties and alcohol content in white wine samples. Through meticulous data preprocessing, including outlier removal, feature engineering, and normalization, coupled with the implementation of linear regression models, the study uncovered influential factors affecting alcohol quality. Notably, properties such as acidity, sulfur dioxide levels, and alcohol percentage significantly influenced alcohol quality in wine. Supported by robust cross-validation techniques, the regression models demonstrated satisfactory performance with accuracies of 92% using linear regression algorithm, 90% using gradient boosting algorithm, and 83% using decision tree algorithm and an MAE value of 0.3 in predicting alcohol quality based on these properties. These findings provide valuable insights for wine quality assessment and production processes, enabling producers to optimize methods and enhance product quality by leveraging an understanding of the impact of physiochemical properties on alcohol quality.

Appendices

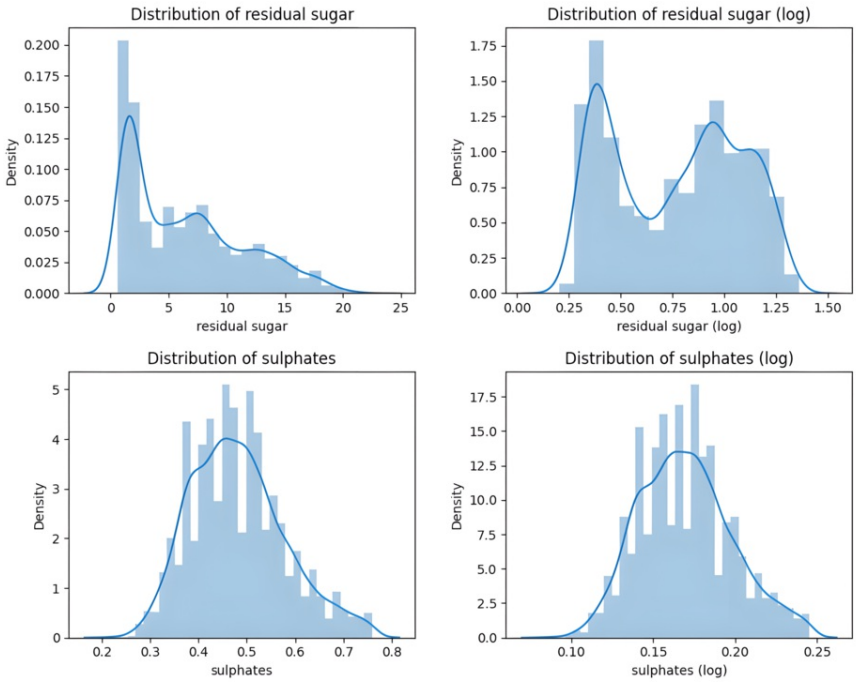

FIGURE 6: Distribution plots.

The distribution plots (Figure *6*) indicate the comparison of original and transformed distributions of specific variables. These also asses the impact of transformations on the data distribution, potentially making it more suitable for modeling.

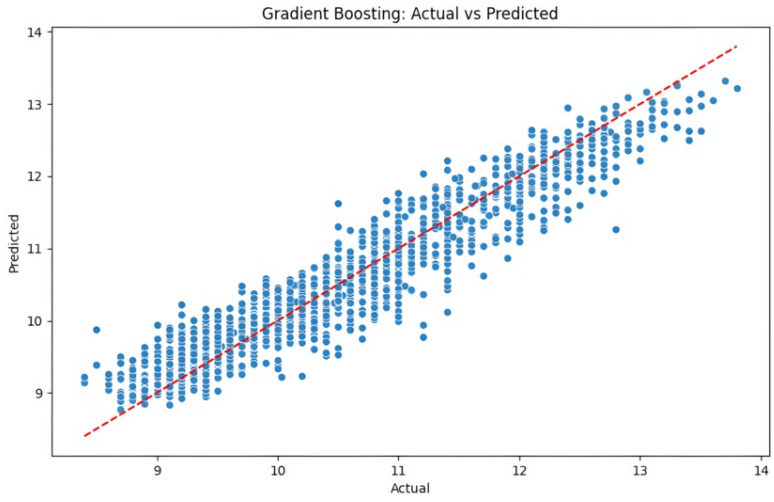

FIGURE 7: Actual vs. prediction - gradient boosting.

Results of the gradient boosting regression to predict target values, evaluating its performance using the visual method, are provided in Figure *7*, which shows the plot of actual vs. predicted outcomes of the model after training.

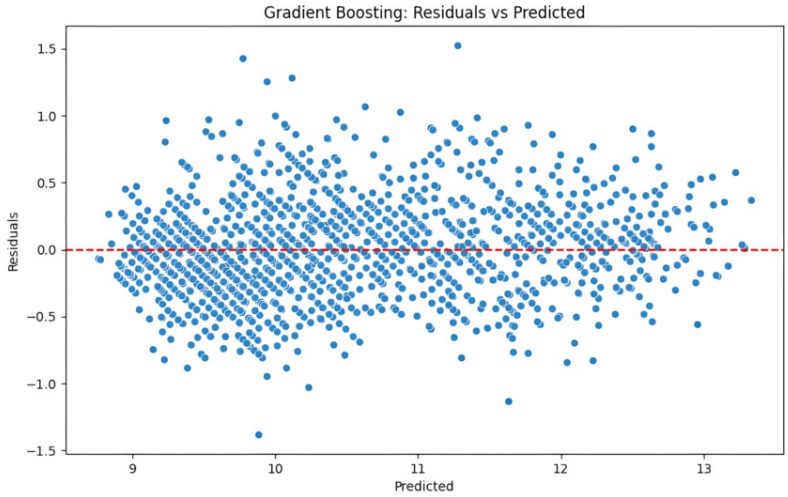

FIGURE 8: Residuals vs. prediction - gradient boosting.

A residuals plot helps evaluate the model's accuracy by visualizing the differences between predicted and actual values (Figure *8*). Ideally, the residuals (prediction errors) should be randomly scattered around zero, indicating that the model captures the data pattern well without systematic bias. If patterns or trends appear in the residuals, it might suggest that the model has not fully captured certain aspects of the data, indicating potential areas for improvement or the need for additional features or different modeling techniques.

Additional Information

Author Contributions

All authors have reviewed the final version to be published and agreed to be accountable for all aspects of the work.

Concept and design: Ankush D. Sawarkar, Aditya S. Baheti, Ubaid A. Shaikh

Acquisition, analysis, or interpretation of data: Ankush D. Sawarkar, Ubaid A. Shaikh, Deepti D. Shrimankar

Drafting of the manuscript: Ankush D. Sawarkar, Aditya S. Baheti, Ubaid A. Shaikh

Critical review of the manuscript for important intellectual content: Ankush D. Sawarkar, Aditya S. Baheti, Ubaid A. Shaikh, Deepti D. Shrimankar

Supervision: Ankush D. Sawarkar, Ubaid A. Shaikh

Disclosures

Human subjects: All authors have confirmed that this study did not involve human participants or tissue. **Animal subjects:** All authors have confirmed that this study did not involve animal subjects or tissue. **Conflicts of interest:** In compliance with the ICMJE uniform disclosure form, all authors declare the following: **Payment/services info:** All authors have declared that no financial support was received from any organization for the submitted work. **Financial relationships:** All authors have declared that they have no financial relationships at present or within the previous three years with any organizations that might have an interest in the submitted work. **Other relationships:** All authors have declared that there are no other relationships or activities that could appear to have influenced the submitted work.

Acknowledgements

The authors are thankful to the Director, Shri Guru Gobind Singhji Institute of Engineering and Technology (SGGSIE&T), Nanded, India, for providing the necessary facilities for this work.

References

1. Schneider A, Hommel G, Blettner M: Linear regression analysis: part 14 of a series on evaluation of scientific

publications.. Deutsches Arzteblatt International. 2010, 107:776-782. 10.3238/arztebl.2010.0776
2. Dahal K, Dahal J, Banjade H, Gaire S: Prediction of wine quality using machine learning algorithms. Open Journal of Statistics. 2021, 11:278-289. 10.4236/ojs.2021.112015
3. Cortez P, Cerdeira A, Almeida F, Matos T, Reis J: Modeling wine preferences by data mining from physicochemical properties. Decision Support Systems. 2009, 47:547-553. 10.1016/j.dss.2009.05.016
4. Aich S, Al-Absi AA, Hui KL, Lee JT, Sain M:A classification approach with different feature sets to predict the quality of different types of wine using machine learning techniques. International Conference on Advanced Communication Technology (ICACT). 2018, 1:1-2. 10.23919/ICACT.2018.8323673
5. Bhardwaj P, Tiwari P, Olejar Jr K, Parr W, Kulasiri D:A machine learning application in wine quality prediction. Machine Learning with Applications. 2022, 8:100261. 10.1016/j.mlwa.2022.100261
6. Monro TM, Moore RL, Nguyen M-C, Ebendorff-Heidepriem H, Skouroumounis GK, Elsey GM, Taylor DK: Sensing free sulfur dioxide in wine. Sensors. 2012, 12:10759-10773. 10.3390/s120810759
7. Gupta M, Vanmathi C: A study and analysis of machine learning techniques in predicting wine quality. International Journal of Recent Technology and Engineering (IJRTE). 2021, 10:314-319. 10.35940/ijrte.a5854.0510121
8. Gruenewald PJ, Ponicki WR, Holder HD, Romelsjö A: Alcohol prices, beverage quality, and the demand for alcohol: quality substitutions and price elasticities. Alcoholism: Clinical and Experimental Research. 2006, 30:96-105. 10.1111/j.1530-0277.2006.00011.x
9. Liang W, Chih H, Chikritzhs T: Predicting alcohol consumption patterns for individuals with a user-friendly parsimonious statistical model. International Journal of Environmental Research and Public Health. 2023, 20:2581. 10.3390/ijerph20032581
10. Bowler A, Escrig J, Pound M, Watson N: Predicting alcohol concentration during beer fermentation using ultrasonic measurements and machine learning. Fermentation. 2021, 7:34. 10.3390/fermentation7010034