big data and cognitive computing

MDPI

Article

Exploring Predictive Modeling for Food Quality Enhancement: A Case Study on Wine

Cemil Emre Yavas [1,*], Jongyeop Kim [1], Lei Chen [1], Christopher Kadlec [1] and Yiming Ji [2]

[1] Department of Information Technology, Georgia Southern University, Statesboro, GA 30460, USA; jongyeopkim@georgiasouthern.edu (J.K.); lchen@georgiasouthern.edu (L.C.); ckadlec@georgiasouthern.edu (C.K.)
[2] College of Computing and Software Engineering, Kennesaw State University, Kennesaw, GA 30060, USA; yji9@kennesaw.edu
* Correspondence: cy02470@georgiasouthern.edu

Abstract: What makes a wine exceptional enough to score a perfect 10 from experts? This study explores a data-driven approach to identify the ideal physicochemical composition for wines that could achieve this highest possible rating. Using a dataset of 11 measurable attributes, including alcohol, sulfates, residual sugar, density, and citric acid, for wines rated up to a maximum quality score of 8 by expert tasters, we sought to predict compositions that might enhance wine quality beyond current observations. Our methodology applies a second-degree polynomial ridge regression model, optimized through an exhaustive evaluation of feature combinations. Furthermore, we propose a specific chemical and physical composition of wine that our model predicts could achieve a quality score of 10 from experts. While further validation with winemakers and industry experts is necessary, this study aims to contribute a practical tool for guiding quality exploration and advancing predictive modeling applications in food and beverage sciences.

Keywords: wine quality prediction; machine learning; polynomial ridge regression; linear regression; RMSE; accuracy; human perception

check for updates

Academic Editor: Domenico Ursino

Received: 6 December 2024
Revised: 21 February 2025
Accepted: 24 February 2025
Published: 26 February 2025

Citation: Yavas, C.E.; Kim, J.; Chen, L.; Kadlec, C.; Ji, Y. Exploring Predictive Modeling for Food Quality Enhancement: A Case Study on Wine. *Big Data Cogn. Comput.* **2025**, *9*, 55. https://doi.org/10.3390/bdcc9030055

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

The quality assessment of wine is traditionally determined by human experts, who rate it on a scale from 0 to 10. While this subjective measure is widely accepted, it is inherently limited by the constraints of human perception. The primary aim of this study is to explore the potential for predictive modeling to quantify wine quality based on measurable physicochemical properties. This approach provides a reproducible framework that is not only useful for understanding quality determinants but also enables the prediction of hypothetical wine compositions that might achieve the highest possible quality scores. These predictions aim to guide future experimentation and quality enhancement efforts in the wine industry.

The predictive modeling of wine quality aligns with broader trends in food and beverage research, where machine learning techniques have been increasingly employed to understand consumer preferences [1]. For instance, Yu et al. [2] utilized a hybrid partial least squares–artificial neural network (PLS-ANN) model to predict consumer liking scores for green tea beverages. Similarly, Sudha et al. [3] developed a fermented millet sprout milk beverage by combining physicochemical studies with consumer acceptability data.

We began by analyzing the 11 variables in the dataset, representing various physicochemical properties of the wines. To address skewness in the data, which can adversely

affect the performance of predictive models, we applied log transformations to the skewed variables. This preprocessing step stabilized variance and facilitated more linear relationships among the variables.

To investigate the relationships between the physicochemical properties and wine quality, we initially employed a linear regression model. However, recognizing that wine quality is unlikely to have a purely linear relationship with its underlying properties, we expanded our analysis to explore quadratic and cubic relationships. Specifically, we fit second-degree and third-degree polynomial models to capture potential nonlinearities and interactions among the variables.

Through this comprehensive exploration, we found that the relationship between the variables and wine quality was best captured by a combination of linear, quadratic, and logarithmic terms. Building on this insight, we optimized the model by evaluating all possible subsets of the 11 variables—2047 combinations in total. For each subset, we calculated the performance of the second-degree polynomial model, splitting the dataset into a training sample (80%) and a test sample (20%) to ensure robustness and reproducibility. Using a random state of 42, we ensured consistent results across experiments.

Our analysis revealed that comparable accuracy could be achieved with a reduced set of seven variables, leading to the selection of an optimized second-degree polynomial ridge regression model. This approach balanced predictive accuracy with simplicity, reducing the risk of overfitting by minimizing the number of input variables.

The predictive modeling of wine quality also extends beyond traditional applications by exploring hypothetical compositions. While the dataset includes wines rated up to a maximum score of 8, we extended our analysis to predict quality scores approaching 10. These exploratory predictions aim to identify theoretical compositions that could guide future experimentation, offering new perspectives on the complex interplay of variables that contribute to wine quality. We explicitly acknowledge the limitations of extrapolating beyond the observed data range and emphasize that these predictions are intended as a foundation for further validation rather than definitive outcomes.

This study focuses on developing a robust and interpretable predictive model for wine quality based on physicochemical properties. While comparisons with other machine learning algorithms are included, they are presented to validate the effectiveness of the proposed model and highlight its unique contributions. The primary goal is not to provide a comprehensive tutorial on algorithmic capabilities, but to showcase how mathematical modeling can offer actionable insights into wine quality and inspire further research in this domain.

In summary, this study demonstrates how polynomial ridge regression can serve as an effective tool for wine quality prediction, providing actionable insights and a framework for exploring quality enhancements. By extending these methods to hypothetical compositions, we hope to inspire future applications in other domains, such as food recipe optimization, and contribute to advancements in the understanding of product quality. Future work should include validation of the proposed methodology with winemakers and experts to ensure practical applicability and alignment with industry standards. Such validation could not only refine the model but also provide additional insights into the sensory and contextual factors that influence wine quality, ultimately strengthening the study's contributions to both research and industry.

2. Background

A paper by K. R. Dahal et al. [4], titled "Prediction of Wine Quality Using Machine Learning Algorithms", was published in the open journal of *Statistics* on 18 March 2021. The paper explores the use of machine learning algorithms to predict the quality of wine

based on various parameters. The authors compare the performance of four different ML models: ridge regression (RR), support vector machine (SVM), gradient boosting regressor (GBR), and a multi-layer artificial neural network (ANN). They find that the GBR model performs best, with metrics such as mean squared error (MSE), correlation coefficient (R), and Mean Absolute Percentage Error (MAPE) of 0.3741, 0.6057, and 0.0873 respectively. The paper demonstrates how statistical analysis can help identify key components influencing wine quality before production, aiding manufacturers in improving quality [5].

In the introduction, the authors highlight the importance of wine quality for both consumers and producers. Historically, wine quality was tested after production, which was costly if the quality was poor. With advancements in technology, manufacturers started testing during production, saving time and money. Machine learning has been used to determine wine quality using available data.

The data description and preprocessing section explains that the study uses the red wine dataset from the UCL Machine Learning Repository, containing 11 physicochemical properties and sensory scores from blind taste testers. The authors analyze the Pearson correlation coefficient to identify significant variables affecting quality, noting that alcohol has the highest correlation (0.435). They also address outliers and feature scaling before training the models.

In the methodology section, the paper describes the four algorithms used: ridge regression (RR), which is similar to linear regression but with a shrinkage penalty; support vector machine (SVM), which is kernel-based regression using the radial basis kernel (RBF); gradient boosting regressor (GBR), which is an ensemble algorithm building sequential weak learners; and artificial neural network (ANN), which is composed of layers of neurons using nonlinear transformations.

The results and discussion section reports that the authors use metrics such as MSE, MAPE, and R to evaluate the models. GBR performs best on the test dataset, with the highest R and lowest MSE and MAPE. ANN underperforms due to the small, skewed dataset. The authors identify alcohol and sulfates as the most important features controlling wine quality.

In conclusion, the paper demonstrates the effectiveness of machine learning in predicting wine quality, with GBR being the best-performing model. The authors conclude that machine learning provides an alternative approach to determining wine quality and screening key variables before production.

A paper by Terry Hui-Ye Chiu et al. [6], titled "A Generalized Wine Quality Prediction Framework by Evolutionary Algorithms", was published in *The International Journal of Interactive Multimedia and Artificial Intelligence* on 21 April 2021. The paper focuses on developing a framework that combines different classifiers and their hyperparameters using genetic algorithms to predict wine quality. This approach addresses the variability in wine datasets and offers a robust method for wine quality prediction. The authors propose a hybrid model that evolves through genetic operations to optimize prediction performance.

To evaluate their approach, the authors conducted experiments on wine datasets and demonstrated the effectiveness of the proposed method. The framework allows for automatic discovery of suitable classifiers and hyperparameters, which is crucial for optimizing the prediction results. The results showed that the proposed approach performed better than several other models, highlighting its utility in predicting wine quality effectively.

A paper by Yogesh Gupta [7], titled "Selection of Important Features and Predicting Wine Quality Using Machine Learning Techniques", was published in *Procedia Computer Science* on 31 December 2017. The paper explores the use of machine learning algorithms to predict wine quality using various features. The author examines the dependency of wine quality on different physicochemical characteristics and employs linear regression, neural

networks, and support vector machines for the analysis. The study shows that by selecting important features, better prediction results can be achieved.

A paper by Piyush Bhardwaj et al. [8], titled "Machine learning application in wine quality prediction", was published in the journal *Machine Learning with Applications* on 28 January 2022. The paper focuses on predicting wine quality using machine learning techniques with both synthetic and experimental data. The authors collected 18 Pinot noir wine samples with 54 different characteristics from diverse regions in New Zealand. They utilized synthetic data and various machine learning models, including Adaptive Boosting (AdaBoost), Random Forest (RF), and gradient boosting (GBOOST), among others. The AdaBoost classifier showed 100% accuracy in predicting wine quality. The study demonstrates that machine learning can effectively predict wine quality, particularly for New Zealand Pinot noir wines, by focusing on essential variables.

A paper by Amalia Luque et al. [9], titled "Determining the Importance of Physicochemical Properties in the Perceived Quality of Wines", was published in *IEEE Access* on 18 October 2023. The paper explores how the quality of wine, which holds significant economic, nutritional, and cultural value, can be improved by understanding the impact of physicochemical properties on perceived quality.

The authors used several metrics to analyze the importance of different wine attributes, including a novel metric based on the Jensen–Shannon Divergence (JSD). They found that JSD performed better than previous metrics and demonstrated that the main physicochemical attributes influencing red wine quality were citric acidity, alcohol, sulfates, and fixed acidity, while for white wine, the key attributes were alcohol, free sulfur dioxide, and pH.

In addition, other studies have explored consumer preferences in food and beverage contexts, such as Yu et al. [2], who developed a partial least squares–artificial neural network (PLS-ANN) model to predict consumer liking scores for green tea beverages, and Sudha et al. [3], who integrated physicochemical properties with acceptability data for fermented millet sprout milk beverages. These studies highlight the broad applicability of machine learning models in predicting consumer-relevant metrics.

Knowledge Gaps and Study Motivation

While previous studies have demonstrated the utility of machine learning models in predicting wine quality, several gaps remain unaddressed. For example, Dahal et al. [4] and Bhardwaj et al. [8] achieved high predictive accuracies but did not explore the potential of these models to predict wine qualities beyond the observed range of scores. Similarly, Chiu et al. [6] focused on optimizing classifier performance but did not examine the interpretability of their models in relation to the physicochemical attributes that influence quality. Furthermore, while Gupta [7] and Luque et al. [9] emphasized feature importance, they did not explore systematic methods to optimize the feature selection process for simpler, more interpretable models.

These limitations highlight the need for a methodological framework that not only predicts wine quality but also extends its applicability to hypothetical scenarios and provides interpretable insights into the contributions of physicochemical properties. This study aims to address these gaps by introducing a polynomial ridge regression model optimized through exhaustive feature combination analysis. Additionally, the study explores the possibility of predicting hypothetical quality scores, thereby expanding the scope of existing research.

3. Methodology

3.1. Wine Dataset

This paper will utilize the renowned datasets provided by the University of California Irvine (UCI) Machine Learning Repository [10], with a particular focus on the Wine Quality Dataset [11]. This dataset includes entries from two subsets of Vinho Verde wines originating from the north of Portugal, categorized into red and white varieties. Our analysis will primarily focus on the subset consisting of 1599 red wine samples, the physicochemical features of which are presented in Table 1. Each of these samples is analyzed based on 11 distinct physicochemical traits and assigned a quality rating that ranges from 0 (indicating very poor quality) to 10 (indicating excellent quality), with each rating represented as an integer.

Table 1. Units Units of physicochemical features in the Wine Quality Dataset.

Feature	Unit
Fixed Acidity	g (tartaric acid)/dm^3
Volatile Acidity	g (acetic acid)/dm^3
Citric Acid	g/dm^3
Residual Sugar	g/dm^3
Chlorides	g (sodium chloride)/dm^3
Free Sulfur Dioxide	mg/dm^3
Total Sulfur Dioxide	mg/dm^3
Density	g/cm^3
pH	unitless
Sulfates	g (potassium sulfate)/dm^3
Alcohol	% volume
Quality	score (0–10)

3.2. References to Data Processing Methods

The methodological framework of this study involves the application of several data processing and modeling techniques, including ridge regression, polynomial regression, neural networks, Random Forest, and Principal Component Analysis (PCA). To provide readers with resources for the further exploration of these techniques, we have included references to foundational studies:

- Ridge Regression and Polynomial Regression: Hoerl, A. E., & Kennard, R. W. *Ridge Regression: Biased Estimation for Nonorthogonal Problems. Technometrics*, 12(1), 1970, pp. 55–67. [12]
- Neural Networks: Rumelhart, D. E., Hinton, G. E., & Williams, R. J. *Learning Representations by Back-Propagating Errors. Nature*, 323(6088), 1986, pp. 533–536. [13]
- Random Forests: Breiman, L. *Random Forests. Machine Learning*, 45(1), 2001, pp. 5–32. [14]
- Principal Component Analysis (PCA): Pearson, K. *On Lines and Planes of Closest Fit to Systems of Points in Space. Philosophical Magazine*, 2(11), 1901, pp. 559–572. [15]

These foundational works provide a detailed theoretical basis for the methods employed and serve as valuable resources for readers interested in the underlying principles of data processing and modeling techniques.

3.3. Methodology Overview

This study follows a systematic methodology to process, analyze, and model the wine quality dataset. The complete workflow is visually represented in Figure 1, which illustrates the preprocessing steps, feature engineering, model training, and final selection.

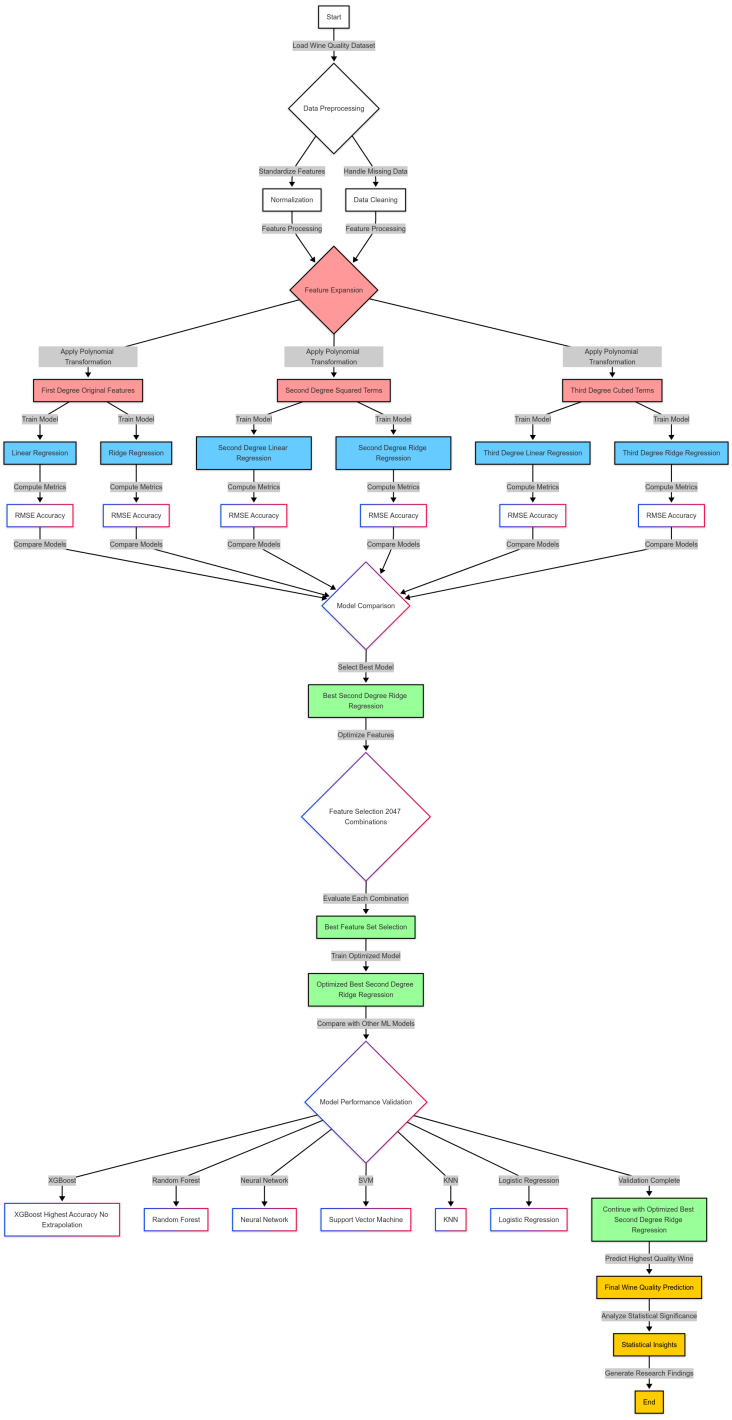


Figure 1. Flowchart of the methodology applied in this study. The arrows indicate the sequential steps in the workflow, from data preprocessing to model selection and evaluation. The boxes are color-coded for clarity: red represents feature processing and expansion, blue denotes machine learning models, green highlights the final selected model, and yellow marks critical decision points such as model validation and statistical analysis.

The methodology consists of the following key stages:

3.3.1. Data Preprocessing and Feature Engineering

The initial phase involves loading the wine quality dataset, handling missing data, and standardizing features through normalization. Following this, feature expansion is performed via polynomial transformation, generating first-degree (original), second-degree (squared terms), and third-degree (cubed terms) features.

3.3.2. Model Training and Evaluation

After feature expansion, multiple models are trained, including linear regression, ridge regression, and polynomial variants of these models. Each model is evaluated using Root Mean Square Error (RMSE) and accuracy metrics to assess predictive performance.

3.3.3. Feature Selection and Optimized Model

To refine the model, a feature selection process evaluates all 2047 possible combinations of the 11 wine characteristics. The optimal subset of features is determined using second-degree ridge regression, balancing accuracy and model simplicity. The best-performing equation is identified and validated against various machine learning models such as XGBoost, Random Forest, neural networks, and support vector machines.

3.3.4. Final Model and Prediction

Some models lack extrapolation capability, making them unsuitable for predicting wine quality outside the training range. Therefore, the best second-degree ridge regression model is selected as the final predictive model. This model is then used to predict the highest-quality wine and analyze the statistical significance of the selected features.

3.4. Exploratory Data Analysis

We conducted exploratory data analysis (EDA) to investigate the relationships between the different features in the Portuguese red wine dataset. This subsection provides detailed explanations of the statistical tests applied, as well as graphical representations summarizing the data.

3.4.1. Descriptive Statistics

Descriptive statistics provide a summary of the dataset. As shown in Table 2, the dataset contains a total of 1599 samples with a variety of chemical properties (e.g., fixed acidity, volatile acidity, alcohol, etc.). The mean alcohol content is 10.42%, and the average wine quality is 5.64. Each feature varies significantly in range, as evident from the minimum, maximum, and standard deviation values.

Table 2. Descriptive Descriptive statistics of the wine dataset.

Feature	Count	Mean	Std	Min	25%	50%	75%	Max
Fixed Acidity	1599	8.32	1.74	4.60	7.10	7.90	9.20	15.90
Volatile Acidity	1599	0.53	0.18	0.12	0.39	0.52	0.64	1.58
Citric Acid	1599	0.27	0.19	0.00	0.09	0.26	0.42	1.00
Residual Sugar	1599	2.54	1.41	0.90	1.90	2.20	2.60	15.50
Chlorides	1599	0.09	0.05	0.01	0.07	0.08	0.09	0.61
Free Sulfur Dioxide	1599	15.87	10.46	1.00	7.00	14.00	21.00	72.00
Total Sulfur Dioxide	1599	46.47	32.90	6.00	22.00	38.00	62.00	289.00
Density	1599	0.997	0.002	0.990	0.9956	0.9968	0.9978	1.0037
pH	1599	3.31	0.15	2.74	3.21	3.31	3.40	4.01
Sulfates	1599	0.66	0.17	0.33	0.55	0.62	0.73	2.00
Alcohol	1599	10.42	1.07	8.40	9.50	10.20	11.10	14.90

3.4.2. Correlation Analysis

The correlation matrix Figure 2 shows the pairwise Pearson correlation coefficients between the features. Pearson's correlation coefficient r is calculated as

$$r = \frac{\sum (X - \overline{X})(Y - \overline{Y})}{\sqrt{\sum (X - \overline{X})^2 \sum (Y - \overline{Y})^2}} \quad (1)$$

where X and Y are the variables and $\overline{X}$ and $\overline{Y}$ are their means. It can be observed that alcohol has a moderately positive correlation with wine quality (0.48), whereas volatile acidity shows a negative correlation with wine quality (−0.39). The strong positive correlation between fixed acidity and citric acid (0.67) suggests that these two features are closely related.

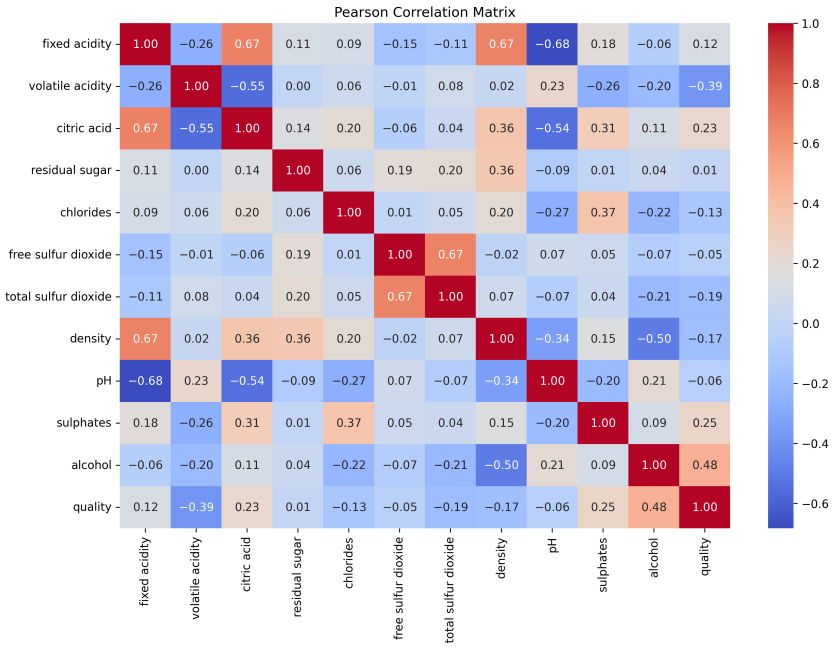


Figure 2. Pearson correlation matrix of wine features.

3.4.3. Variance Inflation Factor (VIF)

The variance inflation factor (VIF) is a statistical measure used to detect multicollinearity in regression analysis. It quantifies the extent to which the variance of an estimated regression coefficient increases due to correlations among predictors. A VIF value greater than 10 is often considered indicative of severe multicollinearity, while values exceeding 5 may indicate potential concerns [16–19]. The VIF is computed as the ratio of the variance of the estimated coefficients in a model to the variance of the coefficients in a hypothetical model devoid of multicollinearity, effectively measuring the inflation in variance caused by multicollinearity among predictors [20,21]. Researchers typically evaluate VIF values in conjunction with tolerance statistics, where a tolerance below 0.1 signifies multicollinearity issues [17,19,22]. Consequently, VIF serves as a vital diagnostic tool in regression modeling, enhancing the reliability of statistical inferences drawn from the data [23,24]. To check for multicollinearity, we calculated the variance inflation factor (VIF) for each feature. VIF is given by

$$\mathrm{VIF} = \frac{1}{1 - R^2} \quad (2)$$

where R^2 is the coefficient of determination of a regression of one predictor on all other predictors. A VIF value above 10 typically indicates multicollinearity. The results, displayed

in Table 3, show that fixed acidity and density have relatively high VIF values of 7.77 and 6.34, respectively, suggesting that these variables may be correlated with others in the dataset.

Table 3. Variance inflation factor (VIF) for all wine features.

Feature	VIF
Fixed Acidity	7.77
Density	6.34
Citric Acid	3.13
Alcohol	3.03
pH	3.33
Total Sulfur Dioxide	2.19
Volatile Acidity	1.95
Residual Sugar	1.81
Chlorides	1.67
Free Sulfur Dioxide	1.52
Sulfates	1.48

3.4.4. Principal Component Analysis (PCA)

Principal Component Analysis (PCA) is a widely utilized statistical technique for dimensionality reduction, particularly in the fields of data analysis and machine learning. It transforms high-dimensional data into a lower dimensional form while preserving as much variance as possible. This is achieved by identifying the principal components, which represent the directions of maximum variance within the dataset. PCA is especially effective in simplifying complex datasets, facilitating the visualization and interpretation of the underlying structure of the data [25,26].

In practical applications, PCA is employed across various domains, including image processing, where it assists in feature extraction and face recognition by reducing the dimensionality of image data while retaining essential features [27–29]. Furthermore, PCA can enhance the performance of machine learning algorithms by alleviating the curse of dimensionality, thereby improving computational efficiency and model accuracy [30,31]. Its versatility and effectiveness make PCA a fundamental tool in data science and analytics. We calculate the principal components by solving the eigenvalue decomposition problem of the covariance matrix Σ, such that

$$\Sigma \mathrm{v} = \lambda \mathrm{v} \tag{3}$$

where v are the eigenvectors (principal components) and λ are the eigenvalues (variance captured by each component). The PCA identifies the directions (components) in which the data varies the most, with the first few components typically capturing most of the variability in the dataset.

In our analysis, the first two principal components (PC1 and PC2) were extracted, which are linear combinations of the original features in the dataset. These components can be described as follows:

- First Principal Component (PC1): This is the linear combination of the original features that captures the maximum variance in the dataset. Essentially, PC1 represents the direction in the multi-dimensional feature space along which the data points exhibit the greatest spread or variation. Mathematically, it can be expressed as

$$\mathrm{PC1} = w_{11} \cdot \mathrm{Feature}_1 + w_{12} \cdot \mathrm{Feature}_2 + w_{13} \cdot \mathrm{Feature}_3 + \cdots + w_{1k} \cdot \mathrm{Feature}_k \tag{4}$$

where w_{1i} are the weights (or loadings) assigned to each original feature in forming the first principal component.

- Second Principal Component (PC2): This is the linear combination of the original features that captures the next highest amount of variance, subject to the constraint that it is orthogonal (uncorrelated) to PC1. PC2 represents the direction of the second greatest spread or variation in the data. It is calculated similarly to PC1 but with a different set of weights:

$$\text{PC2} = w_{21} \cdot \text{Feature}_1 + w_{22} \cdot \text{Feature}_2 + w_{23} \cdot \text{Feature}_3 + \cdots + w_{2k} \cdot \text{Feature}_k \quad (5)$$

The loadings (w_{ij}) in the equations above indicate the contribution of each feature to the corresponding principal component. For example, a high absolute value of w_{ij} signifies that the feature has a significant influence on the corresponding component. These loadings are distinct from the scores, which represent the projection of the original data points onto the principal components.

The scatter plot in Figure 3 illustrates the PCA scores for the first two principal components, PC1 and PC2, for the wine dataset. Each data point represents a sample wine, and the coordinates of these points correspond to their scores on PC1 and PC2. The scores are calculated by multiplying the original feature values with the loadings. The points are colored according to their quality scores, ranging from 3 (lowest quality) to 8 (highest quality), as evaluated by expert tasters based on 11 physicochemical parameters. This visualization highlights how the samples are distributed in the reduced two-dimensional space.

The overlapping regions in the plot suggest that there is no clear separation between high- and low-quality wines based solely on these two components, indicating that wine quality may depend on more complex interactions between features. The visual patterns in this plot provide insights into the relationships among the features and their impact on wine quality.

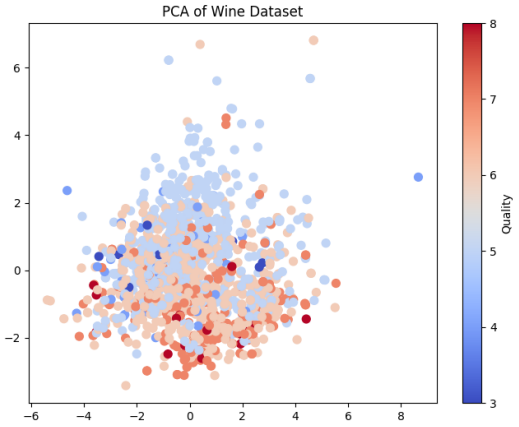


Figure 3. Principal Component Analysis (PCA) of the wine dataset. Data points are colored according to wine quality scores, ranging from 3 (lowest) to 8 (highest), as defined by expert evaluations based on physicochemical properties.

3.4.5. Distribution and Normality Tests

Figure 4 illustrates the distribution of all features in the dataset through histograms combined with kernel density estimation (KDE). KDE estimates the probability density function of a random variable. We performed normality tests using the Shapiro–Wilk test, where the test statistic W is defined as

$$W = \frac{\left(\sum_{i=1}^{n} a_i x_{(i)}\right)^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \quad (6)$$

where $x_{(i)}$ are the ordered data points and a_i are constants derived from the covariance matrix. The Shapiro–Wilk test evaluates whether a sample comes from a normally distributed population, with higher values of W indicating closer conformity to a normal distribution.

The Shapiro–Wilk test is a widely utilized statistical method for assessing the normality of data distributions. Developed by Samuel Shapiro and Martin Wilk in 1965, it is particularly effective for small sample sizes, typically recommended for samples of fewer than 50 observations [32,33]. The test operates by comparing the observed distribution of data to a normal distribution, calculating a W statistic that reflects how well the data conforms to normality [34].

Research indicates that the Shapiro–Wilk test has superior power compared to other normality tests, such as the Kolmogorov–Smirnov and Anderson–Darling tests, particularly in detecting deviations from normality due to skewness or kurtosis [35,36]. Its application is crucial in various fields, including psychology and medicine, where normality assumptions underpin many statistical analyses [33,37]. However, it is essential to interpret the results cautiously, as the test can be sensitive to sample size and outliers [38,39].

The results of the Shapiro–Wilk test indicate that almost all features in the dataset deviate significantly from normality, as evidenced by very low *p*-values. A selection of key results includes

- Fixed Acidity: $W = 0.942$, p-value $= 1.52 \times 10^{-24}$
- Volatile Acidity: $W = 0.974$, p-value $= 2.69 \times 10^{-16}$
- Citric Acid: $W = 0.955$, p-value $= 1.02 \times 10^{-21}$
- Residual Sugar: $W = 0.566$, p-value $= 0.0$
- Chlorides: $W = 0.484$, p-value $= 0.0$
- Free Sulfur Dioxide: $W = 0.902$, p-value $= 7.70 \times 10^{-31}$
- Total Sulfur Dioxide: $W = 0.873$, p-value $= 3.57 \times 10^{-34}$

These results show that the distributions of features such as residual sugar, chlorides, and total sulfur dioxide are particularly non-normal. The combination of histograms and KDE plots in Figure 4 visually confirms these deviations, as the distributions appear skewed and non-symmetrical. This deviation from normality has implications for the choice of statistical methods, as many parametric tests and models assume normally distributed data.

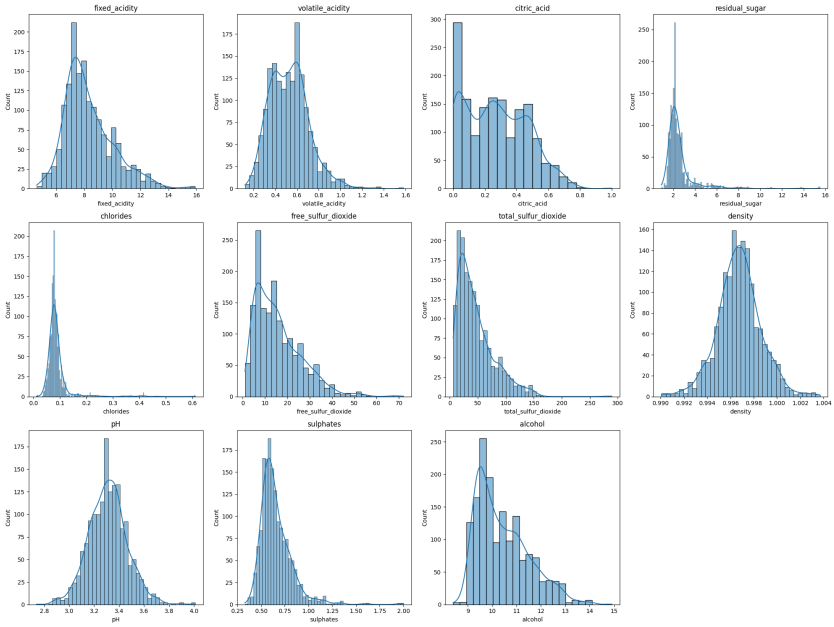


Figure 4. Histograms and KDE of wine features.

3.4.6. Boxplots

To explore potential outliers and the spread of each feature, we generated boxplots. Boxplots visualize the distribution of the data through quartiles. The interquartile range (IQR) is calculated as

$$\mathrm{IQR} = \mathrm{Q}_3 - \mathrm{Q}_1 \tag{7}$$

where Q_3 and Q_1 are the 75th and 25th percentiles, respectively. Figure 5 displays standardized boxplots of all features, allowing for direct comparison despite the differences in scale across the features. Additionally, a raw boxplot of the unstandardized features is provided in Figure 6. Both figures highlight significant outliers, especially for total sulfur dioxide and chlorides.

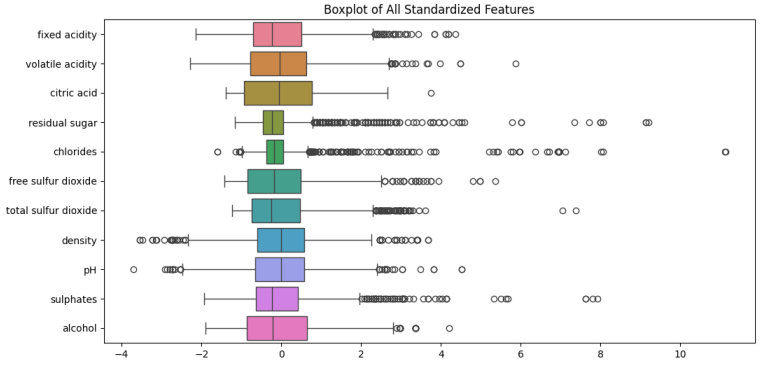


Figure 5. Standardized boxplot of all features.

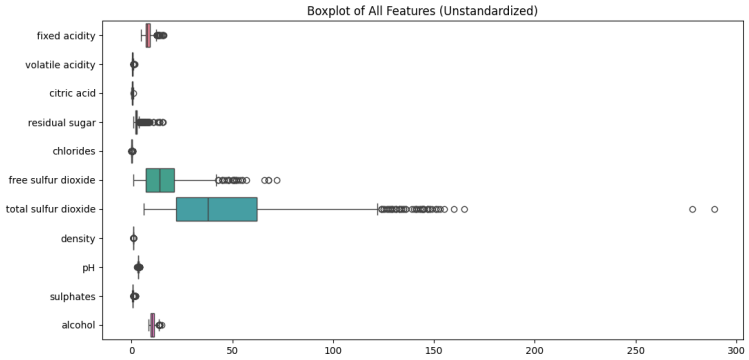


Figure 6. Boxplot of all features (unstandardized).

3.4.7. ANOVA Results

We conducted one-way ANOVA tests to determine the effect of each feature on wine quality. One-way ANOVA (Analysis of Variance) is a statistical method used to determine if there are significant differences between the means of three or more independent groups. This technique is particularly useful when comparing multiple groups to assess whether at least one group mean is different from the others, based on a single independent variable. The fundamental principle of one-way ANOVA is to analyze the variance within each group and between groups, allowing researchers to infer population mean differences from sample data.

The one-way ANOVA test operates under certain assumptions, including the normality of data distribution and homogeneity of variances across groups. When these assumptions are met, the test can effectively identify significant differences among group means. In practical applications, one-way ANOVA is often followed by post hoc tests, such as Tukey's HSD, to determine which specific groups differ from each other [40,41]. This method is

widely utilized across various fields, including biology, psychology, and social sciences, demonstrating its versatility and importance in statistical analysis [42–44]. The test statistic (F) for ANOVA is calculated as

$$F = \frac{MS_{between}}{MS_{within}} \tag{8}$$

where $MS_{between}$ is the mean sum of squares between groups and MS_{within} is the mean sum of squares within groups. The results, shown below, indicate that all features significantly impact wine quality, as evidenced by their p-values being less than 0.05. This suggests that variations in these features are associated with differences in wine quality.

- Fixed Acidity: $F = 3126.21$, p-value $= 0.0$
- Volatile Acidity: $F = 60,979.24$, p-value $= 0.0$
- Citric Acid: $F = 66,691.96$, p-value $= 0.0$
- Residual Sugar: $F = 5810.01$, p-value $= 0.0$
- Chlorides: $F = 75,227.40$, p-value $= 0.0$
- Free Sulfur Dioxide: $F = 1522.99$, p-value $= 8.82 \times 10^{-273}$
- Total Sulfur Dioxide: $F = 2462.15$, p-value $= 0.0$
- Density: $F = 527,69.96$, p-value $= 0.0$
- pH: $F = 12,785.32$, p-value $= 0.0$
- Sulfates: $F = 58,190.52$, p-value $= 0.0$
- Alcohol: $F = 20,494.88$, p-value $= 0.0$

The high F-statistics for features such as chlorides, volatile acidity, and citric acid indicate that these features have particularly strong effects on the wine quality. These results suggest that variations in these chemical properties contribute significantly to the differences in wine quality.

3.5. Distribution Analysis of Wine Quality

We conducted an exploratory analysis to determine the distribution of the wine quality variable. This analysis involved fitting several distributions, including normal, log-normal, Weibull, gamma, and exponential, to the empirical data. The fitted distributions were then compared using goodness-of-fit tests and visualized alongside the empirical data for a comprehensive understanding of the underlying distribution. The results of the statistical tests and the fitted distribution plots are presented in this subsection.

3.5.1. Descriptive Statistics

The descriptive statistics of the wine quality variable are shown in Table 4. The dataset contains 1599 samples, with a mean quality score of 5.64 and a standard deviation of 0.81. The minimum quality score is 3, while the maximum is 8. The majority of wines fall within the 5 to 6 quality range, with 75% of the wines having a quality score of 6 or below.

Table 4. Descriptive statistics for the wine quality variable.

Statistic	Value
Count	1599
Mean	5.64
Standard Deviation	0.81
Minimum	3
25th Percentile	5
50th Percentile (Median)	6
75th Percentile	6
Maximum	8

3.5.2. Goodness-of-Fit Tests

To evaluate which distribution best fits the wine quality data, we performed several goodness-of-fit tests, including the Kolmogorov–Smirnov (K-S) test, Anderson–Darling (A-D) test, and Shapiro–Wilk test. These tests were applied to the normal, log-normal, Weibull, gamma, and exponential distributions. The Kolmogorov–Smirnov (K-S) test is a non-parametric statistical method used to determine the goodness of fit between an empirical cumulative distribution function (CDF) and a theoretical distribution or to compare two empirical distributions. It quantifies the maximum distance between the two CDFs, providing a measure of how well the data conforms to the expected distribution [45,46]. Originally developed by Andrey Kolmogorov and Nikolai Smirnov, the K-S test is widely utilized in various fields, including environmental science, finance, and biomedical research, to assess normality and distributional assumptions [47,48].

The K-S test is particularly advantageous due to its applicability to small sample sizes and its robustness against deviations from normality, making it a preferred choice for many researchers [49,50]. Furthermore, its implementation has been enhanced through programming libraries, facilitating its integration into statistical software for broader accessibility [51]. Overall, the K-S test remains a fundamental tool in statistical analysis for evaluating the distributional properties of datasets.

Normal Distribution

The K-S test for the normal distribution yielded a statistic of 0.25, with a p-value of approximately 2.08×10^{-88}, indicating a significant deviation from normality. The Anderson–Darling test further supports this result, with a statistic of 110.63, which exceeds the critical values at all significance levels. The Shapiro–Wilk test produced a statistic of 0.86 and a p-value of 9.52×10^{-36}, providing strong evidence against the normality assumption. The equation for the normal distribution's probability density function is

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \tag{9}$$

where μ is the mean and σ is the standard deviation.

Log-Normal Distribution

The log-normal distribution showed a slightly better fit compared to the normal distribution. The K-S test yielded a statistic of 0.24, with a p-value of 6.68×10^{-81}, still indicating a significant departure from a log-normal distribution. The probability density function of the log-normal distribution is given by

$$f(x) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left(-\frac{(\ln x-\mu)^2}{2\sigma^2}\right) \tag{10}$$

Weibull Distribution

The Weibull distribution also did not fit the wine quality data well. The K-S test returned a statistic of 0.25 and a p-value of 9.32×10^{-87}, suggesting a poor fit. The Weibull distribution's probability density function is

$$f(x) = \frac{k}{\lambda}\left(\frac{x}{\lambda}\right)^{k-1} \exp\left(-\left(\frac{x}{\lambda}\right)^{k}\right) \tag{11}$$

where k is the shape parameter and λ is the scale parameter.

Gamma Distribution

The gamma distribution provided a similar result, with a K-S statistic of 0.25 and a p-value of 1.53×10^{-86}, indicating that this distribution is also not a good fit for the data. The probability density function for the gamma distribution is

$$f(x) = \frac{x^{k-1}e^{-x/\theta}}{\theta^k \Gamma(k)} \tag{12}$$

where k is the shape parameter, θ is the scale parameter, and $\Gamma(k)$ is the gamma function.

Exponential Distribution

The exponential distribution showed the worst fit among all tested distributions. The K-S test yielded a statistic of 0.49, with a p-value of 0, suggesting that the exponential distribution does not describe the data at all. The probability density function for the exponential distribution is

$$f(x) = \lambda e^{-\lambda x} \tag{13}$$

where λ is the rate parameter.

3.5.3. Visualization of Fitted Distributions

We visualized the empirical wine quality data along with the fitted distributions (normal, log-normal, Weibull, gamma, and exponential) in Figure 7. The histogram of the empirical data is overlaid with the fitted distributions, where different colors represent each distribution. It is evident from the plot that the normal, log-normal, Weibull, and gamma distributions follow similar patterns, while the exponential distribution diverges significantly, as confirmed by the goodness-of-fit tests.

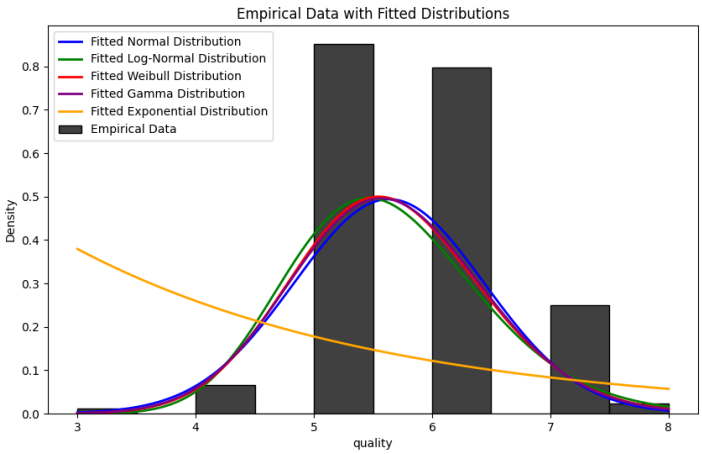


Figure 7. Overlay of empirical wine quality data and fitted distributions (normal, log-normal, Weibull, gamma, and exponential).

3.5.4. Kernel Density Estimation

In addition to fitting parametric distributions, we also computed a kernel density estimate (KDE) to visually inspect the shape of the quality distribution. This technique involves placing a kernel function, which is a smooth and symmetric function, at each data point and summing these contributions to create a continuous estimate of the density. The choice of kernel and the bandwidth parameter significantly influence the quality of the density estimate, with various kernels being proposed to optimize performance under different conditions [52,53].

KDE is widely applicable across various fields, including statistics, machine learning, and signal processing, serving as a fundamental tool for data visualization and analysis [54–56]. Advancements in computational techniques have improved the efficiency and stability of multivariate KDE, enabling the method to handle complex datasets [57,58]. The method's flexibility allows it to adapt to the underlying structure of the data, making it a valuable approach for exploratory data analysis [59–61]. The KDE is a non-parametric way to estimate the probability density function of a variable and is defined as

$$\hat{f}_h(x) = \frac{1}{n\,h} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right) \tag{14}$$

where K is the kernel function, h is the bandwidth, and x_i are the data points. The KDE plot, shown in Figure 8, highlights the multimodal nature of the data, with peaks around quality scores of 5 and 6. This further reinforces the findings from the goodness-of-fit tests, as the parametric distributions struggled to capture this complexity.

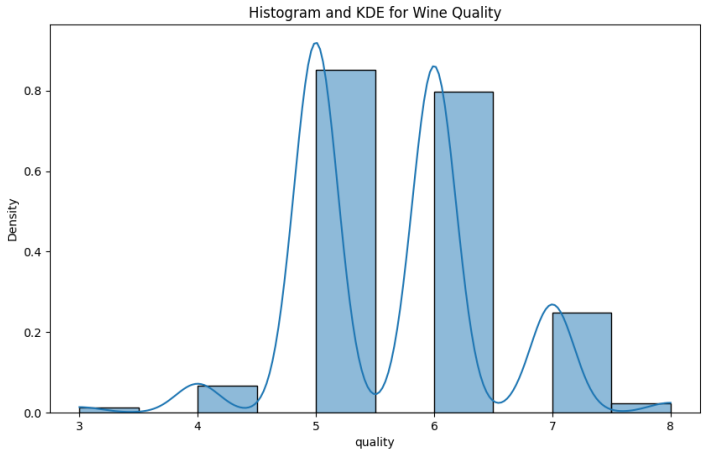


Figure 8. Histogram and kernel density estimate (KDE) of wine quality.

3.6. Limitations of Subjective Quality Scores and Mitigation Strategies

The wine quality dataset utilized in this study assigns quality scores based on evaluations from expert human tasters. These scores, while reflective of sensory attributes such as taste, aroma, and overall balance, are inherently subjective and may introduce potential biases. Such biases arise from differences in individual perceptions, cultural preferences, and contextual factors that could affect the consistency of quality ratings across samples. However, these scores are the industry standard and provide valuable insights into consumer-relevant quality metrics, making them an essential benchmark for wine quality assessment.

To address these limitations, the following considerations were integrated into our methodology:

- Data Data Characteristics: The dataset provides a comprehensive representation of physicochemical properties that are objectively measurable and quantifiable, such as alcohol content, sulfates, residual sugar, and acidity. These features serve as reliable predictors independent of subjective interpretations.
- Emphasis on Objective Features: Our analysis prioritizes the physicochemical attributes as primary predictors of wine quality, aiming to minimize the influence of subjective bias in the modeling process.
- Potential for Cross-Validation with Alternative Datasets: While this study is limited to the UCI dataset, future research could incorporate external datasets that include

alternative scoring systems or combine subjective scores with chemical and physical measurements for enhanced reliability.

- Robust Predictive Modeling: The methodology focuses on predictive modeling that explores potential wine compositions and their predicted quality scores. This predictive approach shifts the emphasis from subjective quality ratings to the underlying relationships between measurable features and quality outcomes.

While recognizing the inherent subjectivity of the dataset, this study demonstrates that the proposed model can provide valuable insights into wine quality prediction through objective feature analysis. Future studies should consider employing multiple datasets with diverse scoring systems or incorporating sensory panel validation to further mitigate the impact of subjective biases.

3.7. Feature Selection and Limitations

The 11 physicochemical properties included in the dataset are objectively measurable and have been extensively validated in the literature as key predictors of wine quality. These features, such as alcohol content, acidity levels, and sulfates, are directly linked to wine's taste, aroma, and structural characteristics. Moreover, their inclusion reflects industry standards for physicochemical analysis.

While the dataset does not include sensory attributes or geographic data, which may provide additional insights, these omissions were due to the dataset's structure and availability. Future studies could incorporate additional features, such as

- Sensory Attributes: Measurable aspects like aroma profile and color intensity, which can be quantified using gas chromatography–mass spectrometry (GC-MS) or spectrophotometric methods, respectively.
- Geographic and Climatic Data: Information about vineyard location, soil type, and climatic conditions could capture regional influences on wine quality.

By considering these additional features in future research, the scope and predictive power of wine quality models can be expanded, enabling more comprehensive analyses.

3.8. Future Directions: Dataset Expansion and Data Augmentation

The current study relies on a single publicly available dataset from the UCI Machine Learning Repository [10], which includes quality scores assigned by expert human tasters. While this dataset provides a robust basis for initial modeling, its limited range of quality scores (3 to 8) and the subjective nature of the evaluations present inherent constraints.

To address these limitations, future research could benefit from incorporating additional datasets from diverse wine-producing regions and vintages. Such datasets would allow for the cross-validation of the model across different contexts, enhancing its generalizability and applicability. Furthermore, exploring data augmentation techniques, such as the synthetic generation of feature combinations or bootstrapping methods, could expand the dataset's diversity and mitigate the limitations imposed by the original dataset's fixed range.

By applying these strategies, future studies could improve the robustness of the model and provide a more comprehensive understanding of wine quality prediction across diverse contexts.

3.9. Limitations of Extrapolation and Validation Requirements

One of the key considerations in this study is the scope of applicability of the predictive models. The models were developed and validated using observed data, making them most reliable for interpolation within the range of quality scores included in the dataset. Predictions beyond this range, particularly hypothetical scores exceeding 10, are inherently

exploratory and should be interpreted with caution. Extrapolation relies on the assumption that the relationships captured by the model remain consistent outside the observed data range, which may not hold true in practice.

To ensure practical applicability, it is essential to validate these predictions experimentally. This involves identifying or creating wine samples with physicochemical compositions similar to those predicted by the model and subjecting them to sensory evaluation by expert tasters. Such validation would help confirm whether the predicted scores align with human quality assessments.

Furthermore, certain model predictions may suggest physicochemical compositions that are challenging to achieve due to physical constraints in grape composition or winemaking processes. For example, a low pH combined with reduced sulfur oxide content may lead to high predicted scores but could be physically unattainable or result in undesirable wine characteristics. Additionally, human perception of physicochemical changes is non-linear and influenced by complex sensory interactions, which may not fully align with the patterns captured by the model.

Recognizing these limitations, we emphasize that the model is a tool for hypothesis generation rather than definitive prediction. Its primary utility lies in identifying potential directions for quality improvement and guiding experimental validation efforts.

3.10. Nonlinear Relationships and Data Transformation

Wine quality is influenced by complex interactions among physicochemical parameters, which often exhibit nonlinear relationships with quality scores. For example,

- Alcohol Content: A moderate alcohol level may positively influence wine quality, while excessive levels may detract from the sensory experience.
- Volatile Acidity: High volatile acidity is typically associated with poor quality, but its impact may vary depending on other attributes like sugar content and pH.
- pH and Acidity: These features have a nonlinear impact on quality due to their role in balancing flavor and preserving wine.

To address these nonlinearities, we employed polynomial regression models (2nd and 3rd degree) to capture the complex relationships between features and quality. Polynomial terms allowed the model to account for interactions and diminishing returns observed in certain attributes.

Additionally, skewness in several features, such as alcohol and residual sugar, necessitated preliminary data transformations. Log transformations were applied to stabilize variance and linearize relationships, enabling better model fitting and interpretation. For instance,

- Log-transformed variables reduced the influence of extreme values, making the model more robust to outliers.
- Transformed features exhibited higher correlations with quality scores, enhancing their predictive power.

Incorporating these steps also provides an educational framework for researchers and practitioners working on similar prediction problems in food science and beyond.

4. Statistical Assessment of Transformation Techniques for Wine Quality Features

4.1. Selecting the Right Transformation

Before proceeding with predictive modeling for wine quality, we applied various statistical tests to assess the need for transformations on the following features: Fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, and total

sulfur dioxide. The aim was to reduce skewness, stabilize variance, and improve normality through log transformation. Our analysis was guided by skewness, kurtosis, and normality tests (Shapiro–Wilk test), as well as variance stabilization checks (Levene's test).

4.1.1. Square Root Transformation

The square root transformation is considered for variables exhibiting increasing variance with the mean (heteroscedasticity) [62]. Compared to logarithmic transformation, it is less aggressive and can be applied to zero but not negative values.

Thinh et al. [63] demonstrated its effectiveness in regression modeling, showing improved accuracy when applied to surface roughness data. Similarly, it has been utilized in various contexts, including enhancing model precision [64] and facilitating compositional data analysis [65].

Variance stabilization is a key advantage of this transformation, particularly when underlying model assumptions are violated, as noted in Bland–Altman analysis [66]. The transformation's applicability extends to phylogenetic studies [67] and spectroscopic data preprocessing [68], underscoring its broad utility.

$$\text{transformed} = \sqrt{x} \tag{15}$$

4.1.2. Inverse Transformation

The inverse transformation is considered for cases where large values need compression while small values require amplification. This transformation can significantly alter data distribution and is particularly effective when dealing with left-skewed variables.

Saffari et al. [69] demonstrated the utility of inverse transformation in modeling cognitive outcomes for early-stage neurocognitive disorders. By addressing the skewness in cognitive data, the study highlights the importance of selecting appropriate transformations to enhance model interpretability and accuracy.

$$\text{transformed} = \frac{1}{x} \tag{16}$$

This transformation's impact on wine quality data will be further evaluated in subsequent sections, considering its suitability for improving regression model performance.

4.1.3. Box-Cox Transformation

The Box–Cox transformation provides a flexible family of power transformations, including logarithmic and square root transformations, depending on the parameter λ. It is widely employed to stabilize variance and improve normality in datasets, ensuring more reliable statistical modeling [70].

This transformation has been applied across multiple fields to address skewness and enhance predictive accuracy. In environmental science, it improves dataset homogeneity, reducing variance disparities [71]. In predictive modeling, it has been shown to enhance normality and improve model performance by addressing nonlinearity [72].

Mosquin et al. [73] found that applying the Box–Cox transformation in streamflow evaluation significantly reduced skewness, leading to superior predictive accuracy in linear models. Similarly, in anomaly detection, it plays a crucial role in stabilizing residual distributions [74], while in remote sensing applications, it corrects biases and minimizes errors in data distribution [75].

Beyond data distribution corrections, this transformation has proven valuable in climatological forecasting by normalizing hydroclimatic variables [76] and in trend and seasonality modeling for predictive applications, such as forecasting box office performance [77]. Additionally, in medical and biological studies, it enhances regression modeling by normal-

izing tumor-infiltrating lymphocyte data for prognosis assessments [78] and contributes to efficient parameter estimation in feed efficiency studies for livestock research [79].

Moreover, the Box–Cox transformation is widely utilized in statistical modeling to normalize data distributions and improve parameter estimation [80]. It plays a crucial role in ensuring that assumptions of normality and variance stability hold across different datasets, making it an essential tool in regression and predictive analytics.

$$\text{transformed} = \begin{cases} \frac{x^{\lambda}-1}{\lambda} & \text{if } \lambda \neq 0 \\ \log(x) & \text{if } \lambda = 0 \end{cases} \tag{17}$$

4.1.4. Exponential and Logarithmic Transformation

Exponential and logarithmic transformations are considered for datasets spanning multiple orders of magnitude or following a power-law distribution. These transformations help improve model interpretability and predictive accuracy by stabilizing variance and addressing skewed distributions.

Bosse et al. [81] demonstrated the benefits of transforming count data before applying scoring metrics such as the Continuous Ranked Probability Score (CRPS) or Weighted Interval Score (WIS), leading to more meaningful and interpretable results in epidemiological forecasting. Schmid et al. [82] emphasized the role of exponential decay models in accounting for time dependency in meta-analyses, improving concordance probability estimates.

In ecological modeling, Zhang et al. [83] explored the use of exponential regression in biomass estimation, highlighting its ability to improve the predictive capabilities of regression models. Similarly, Yu et al. [84] demonstrated the effectiveness of logarithmic regression in revealing relationships between variables, reinforcing its utility in statistical modeling.

Rasheed et al. [85] applied both exponential and logarithmic regression techniques in analyzing the properties of octane isomers, illustrating the adaptability of these transformations across scientific domains. The importance of inverse transformations was further demonstrated by Bueno-López et al. [86], who addressed the challenges of de-transforming logarithmic conversions in regression models.

In medical research, logarithmic transformations have proven effective in predictive modeling. Tu et al. [87] applied them to analyze factors influencing prognosis after liver transplantation, while Pizarro et al. [88] utilized natural logarithmic transformations to assess cardiac autonomic dysfunction in adult congenital heart disease patients, demonstrating their relevance in clinical investigations.

Environmental applications also benefit from these transformations. Subi et al. [89] incorporated logarithmic transformations to estimate organic matter content in farmland soil, highlighting their role in environmental modeling. Similarly, Hieu et al. [90] used logarithmic models to estimate chlorophyll-a levels, reinforcing their effectiveness in remote sensing applications for environmental monitoring.

$$\text{transformed} = e^{x} \tag{18}$$

or

$$\text{transformed} = \log(x) \tag{19}$$

4.1.5. Power Transformation

Power transformations are considered when data exhibits nonlinear relationships that do not fit the criteria for logarithmic or root transformations. Raising data to a power (e.g., square or cube) can help improve model fit and stabilize variance.

Li et al. [91] demonstrated the effectiveness of power transformations in enhancing the accuracy of regression-based statistical postprocessing models for short-term precipitation forecasts. Their findings highlight the importance of transformation techniques in refining predictive models.

$$\text{transformed} = x^2 \tag{20}$$

or

$$\text{transformed} = x^3 \tag{21}$$

4.1.6. Log Transformation

Log transformation is evaluated as a method to stabilize variance, improve normality, and mitigate the influence of outliers in the dataset. This transformation is particularly relevant for right-skewed variables, where applying a logarithmic function can enhance model performance.

The general form of a log transformation is

$$y = \log_b(x + c) \tag{22}$$

where

- x is the original data value;
- y is the transformed value;
- b is the logarithm base, commonly natural logarithm (e) or base 10;
- c is a small constant added to prevent undefined values when $x = 0$, typically 0.1 or 1, depending on the dataset.

  Two commonly used forms of log transformation in data preprocessing are Natural Log Transformation:

$$y = \log(x + c) \tag{23}$$

  Base-10 Log Transformation:

$$y = \log_{10}(x + c) \tag{24}$$

The choice of base depends on the scale and interpretability of transformed values. In the context of wine quality modeling, log transformation is considered for variables exhibiting skewness, as discussed in the previous section. The effectiveness of this transformation will be further assessed in subsequent analyses.

4.1.7. Skewness and Kurtosis

Skewness measures the asymmetry of a distribution, while kurtosis quantifies the tails' heaviness. A positive skewness indicates a longer tail on the right, while a negative skewness indicates a longer tail on the left [92,93]. Kurtosis, on the other hand, measures the "tailedness" of the distribution, reflecting the presence of outliers. High kurtosis (leptokurtic) indicates heavy tails and a sharp peak, while low kurtosis (platykurtic) suggests light tails and a flatter peak [94,95]. Both skewness and kurtosis are essential in various fields, including finance and environmental science, as they provide insights into the underlying data distribution, which can influence statistical modeling and hypothesis testing [96,97]. For a normal distribution, the skewness should be close to zero, and the kurtosis should be approximately three. Log transformation often reduces these measures, bringing the data closer to normality. The results of skewness and kurtosis for the original and log-transformed features are summarized in Table 5. Log transformation successfully reduced the skewness and kurtosis of several features. For example, the skewness of fixed acidity

was reduced from 0.98 to 0.45, while the kurtosis decreased from 1.12 to 0.14, indicating a more symmetrical and normally distributed feature. Similarly, volatile acidity saw a reduction in skewness from 0.67 to 0.27.

Table 5. Skewness and kurtosis of original and log-transformed features.

Feature	Skewness (Original)	Skewness (Log)	Kurtosis (Original)	Kurtosis (Log)
Fixed Acidity	0.98	0.45	1.12	0.14
Volatile Acidity	0.67	0.27	1.22	0.18
Citric Acid	0.32	0.09	−0.79	−1.04
Residual Sugar	4.54	2.25	28.52	7.15
Chlorides	5.68	5.07	41.58	33.50
Free Sulfur Dioxide	1.25	−0.10	2.01	−0.66
Total Sulfur Dioxide	1.51	−0.04	3.79	−0.69

4.1.8. Normality Testing (Shapiro-Wilk Test)

We applied the Shapiro–Wilk test to each feature, both in its original and log-transformed states, to assess improvements in normality.

Table 6 presents the test statistics and p-values. Notably, the log-transformed fixed acidity showed a significant improvement in normality, with its Shapiro–Wilk p-value increasing from 1.52×10^{-24} to 2.65×10^{-13}, as shown in Figure 9. Similarly, volatile acidity and free sulfur dioxide demonstrated improved normality following the log transformation (Figures 10 and 11).

Table 6. Shapiro–Wilk test results for normality (original vs. log-transformed).

Feature	Shapiro-Wilk p-Value (Original)	Shapiro-Wilk p-Value (Log)
Fixed Acidity	1.52×10^{-24}	2.65×10^{-13}
Volatile Acidity	2.69×10^{-16}	1.88×10^{-7}
Citric Acid	1.02×10^{-21}	3.36×10^{-21}
Residual Sugar	0.0	1.59×10^{-40}
Chlorides	0.0	0.0
Free Sulfur Dioxide	7.70×10^{-31}	9.29×10^{-12}
Total Sulfur Dioxide	3.57×10^{-34}	3.15×10^{-9}

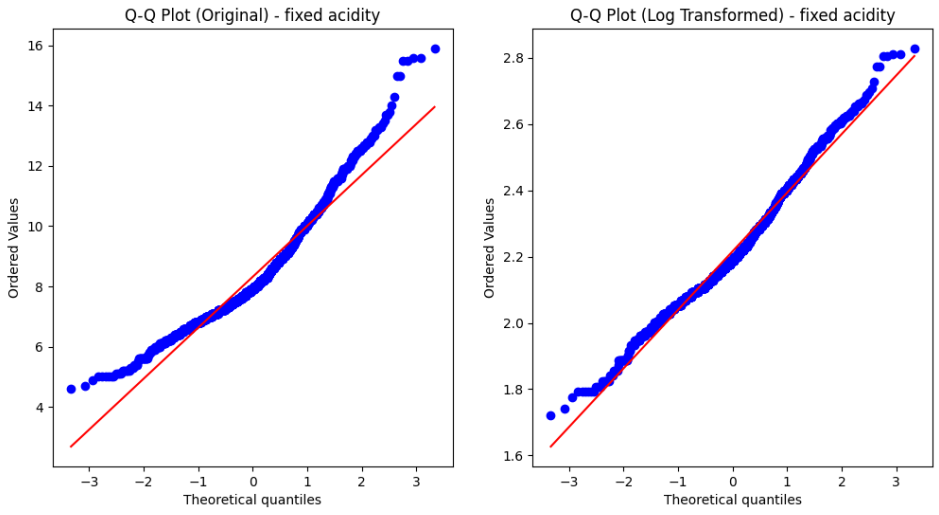


Figure 9. Q-Q plot for fixed acidity (original and log-transformed).

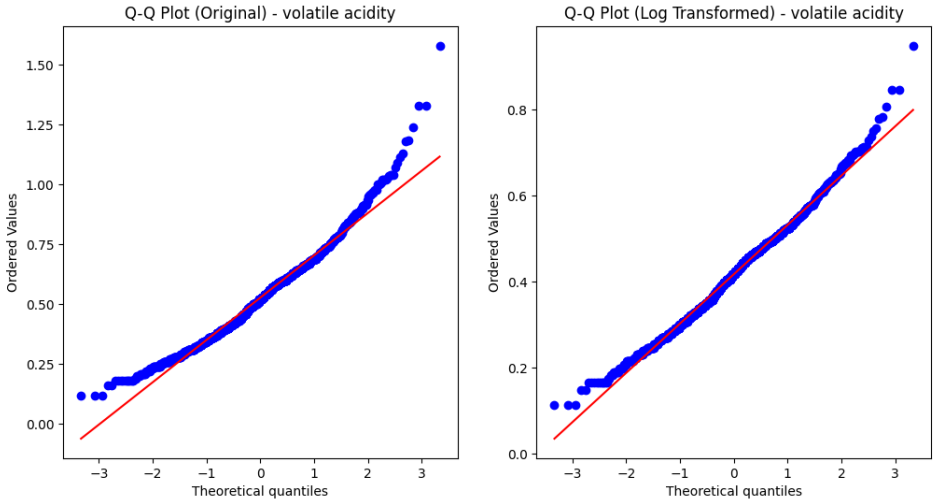


Figure 10. Q-Q plot for volatile acidity (original and log-transformed).

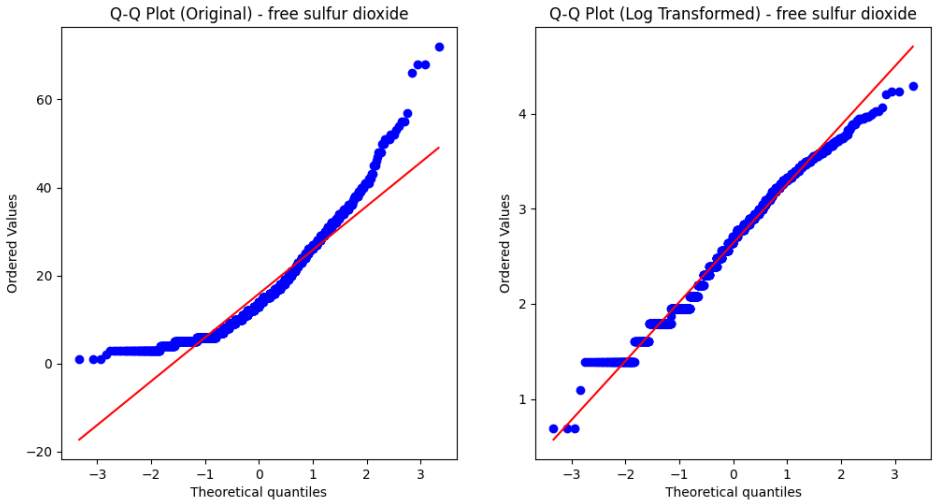


Figure 11. Q-Q plot for free sulfur dioxide (original and log-transformed).

4.1.9. Q-Q Plots

Quantile–Quantile (Q-Q) plots were generated to visually assess the effect of the log transformation on each feature. The Q-Q plot compares the quantiles of the sample data to the theoretical quantiles of a normal distribution. For normally distributed data, the points on the Q-Q plot should fall approximately along the 45-degree line. If the data follows the theoretical distribution closely, the points will lie approximately along a straight line [98,99]. Q-Q plots are particularly useful for assessing normality, as deviations from the line indicate departures from normality [100–102]. They are widely employed in various fields, including statistics, environmental science, and genomics, to evaluate the fit of data to a specified distribution and to identify outliers [103,104]. The visual nature of Q-Q plots (Figures 9–15) allows for a quick assessment of data distribution, making them a staple in statistical analysis [105,106].

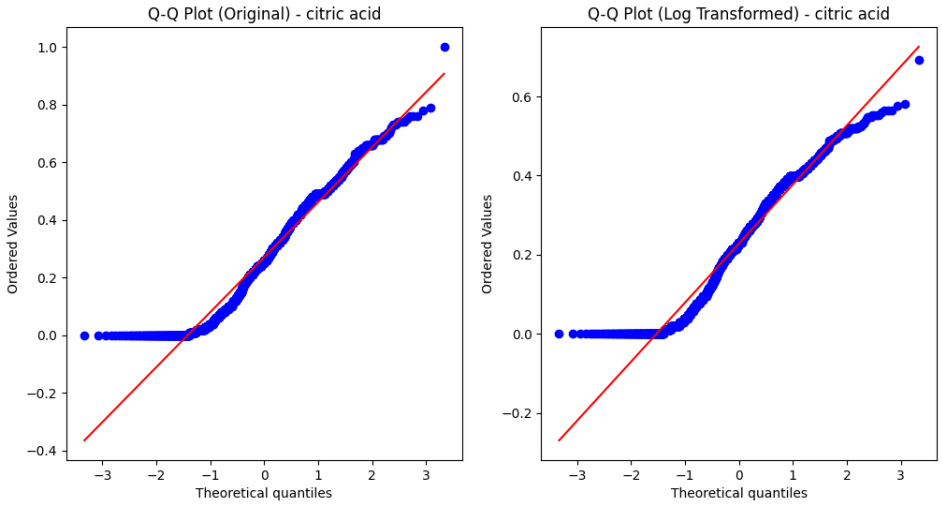


Figure 12. Q-Q plot for citric acid (original and log-transformed).

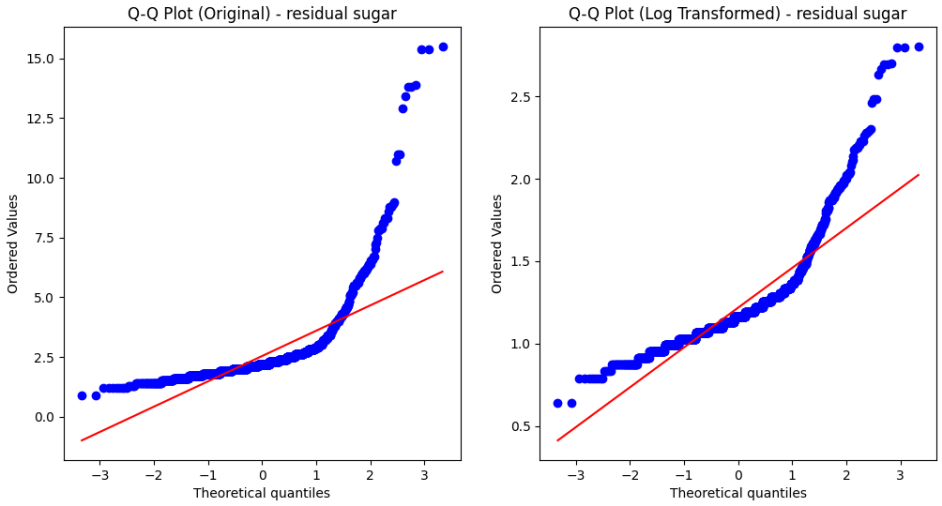


Figure 13. Q-Q plot for residual sugar (original and log-transformed).

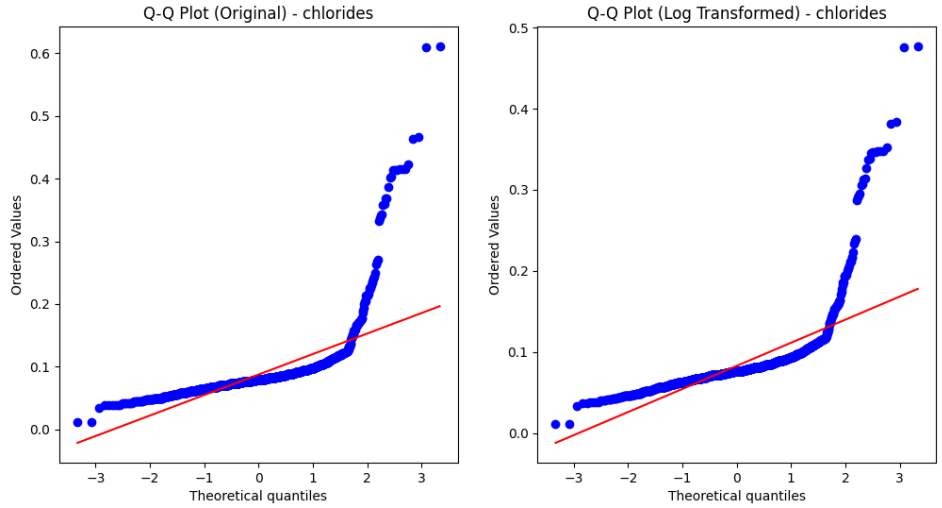


Figure 14. Q-Q plot for chlorides (original and log-transformed).

The Q-Q plots confirm that the log transformation brought several features closer to a normal distribution. For example, as seen in Figure 11, the free sulfur dioxide feature

improved significantly after log transformation. Similar improvements are visible in total sulfur dioxide and volatile acidity, as illustrated in Figures 10 and 15, respectively.

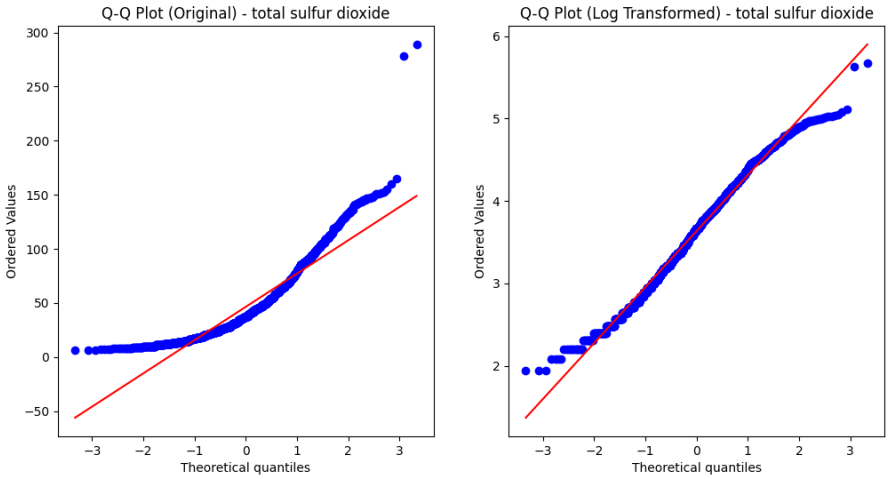


Figure 15. Q-Q plot for total sulfur dioxide (original and log-transformed).

4.1.10. Variance Stabilization (Levene's Test)

We conducted Levene's test to determine whether the log transformation stabilized variance across different groups of the features. Levene's test is a statistical procedure used to assess the homogeneity of variances across different groups. It was introduced by Levene in 1960 and is particularly advantageous because it is robust against violations of the normality assumption, making it suitable for data that may not follow a normal distribution [107,108]. The test evaluates whether the variances of multiple groups are equal, which is a critical assumption in various statistical analyses, including ANOVA [109,110].

Levene's test operates by analyzing the absolute deviations of each observation from its group mean or median, providing a more reliable measure of variance equality compared to traditional methods like Bartlett's test, especially in the presence of non-normal data [111,112]. The results of Levene's test inform researchers whether to proceed with parametric tests that assume equal variances or to adopt alternative methods that do not [113,114]. The Levene's test statistic is calculated as

$$W = \frac{(N - k)}{(k - 1)} \frac{\sum_{i=1}^{k} N_i (Z_{i.} - Z_{..})^2}{\sum_{i=1}^{k} \sum_{j=1}^{N_i} (Z_{ij} - Z_{i.})^2} \quad (25)$$

where N represents the total number of samples, k is the number of groups, Z_{ij} is the transformed data, and $Z_{i.}$ and $Z_{..}$ refer to the group and overall means, respectively. Our test results indicate significant improvements in variance homogeneity for several features, as shown in Table 7. For instance, the p-value for fixed acidity improved from 2.43×10^{-270}, confirming the effectiveness of the log transformation.

Table 7. Levene's test results for variance stabilization.

Feature	Levene's Test Stat	Levene's *p*-Value
Fixed Acidity	1506.44	2.43×10^{-270}
Volatile Acidity	236.01	1.92×10^{-51}
Citric Acid	108.58	5.02×10^{-25}
Residual Sugar	242.66	8.60×10^{-53}
Chlorides	2.46	0.116
Free Sulfur Dioxide	1818.60	5.66×10^{-315}
Total Sulfur Dioxide	1616.09	2.39×10^{-286}

4.1.11. Transformation Considerations for Wine Quality Data

To assess the suitability of transformations for our dataset, histograms of wine quality features (Figure 4) were analyzed. Several features exhibit skewed distributions, suggesting that logarithmic transformations may improve normalization and enhance predictive modeling.

- Fixed Acidity: Displays slight right-skewness. A log transformation could improve symmetry.
- Volatile Acidity: Right-skewed distribution; log transformation may help reduce skewness.
- Citric Acid: Concentration near zero with a long right tail, making log transformation highly suitable.
- Residual Sugar: Highly right-skewed with outliers, benefiting from log transformation to normalize the distribution.
- Chlorides: Exhibits extreme right-skewness, making log transformation a strong candidate.
- Free Sulfur Dioxide and Total Sulfur Dioxide: Both right-skewed, suggesting log transformations could improve normality and aid in linear modeling.
- Density, pH, Sulfates, and Alcohol: Appear relatively symmetrical, so transformation may not be necessary.

Log transformations for fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, and total sulfur dioxide could enhance predictive modeling by stabilizing variance and improving feature distributions. To accommodate zero values, a small constant (e.g., 0.1) is added before applying the transformation where necessary.

4.2. Conclusions

The statistical tests and visualizations (Q-Q plots and Levene's test) indicate that the log transformation effectively improved normality and stabilized variance for most of the selected features. While some features like Chlorides did not show a significant improvement, the majority of features demonstrated better behavior post-transformation, making them more suitable for linear modeling.

5. Regression Models for Wine Quality Prediction

5.1. Linear Regression

Linear regression is a fundamental statistical method used to model the relationship between a dependent variable y and one or more independent variables X. In the context of our wine quality dataset, the goal is to predict the quality of wine based on several chemical properties [115]. It aims to establish a linear equation that best predicts the value of the dependent variable based on the independent variables. This method is widely utilized in various fields, including research statistics, healthcare, finance, and machine learning, due to its simplicity and interpretability [116]. The core principle of linear regression involves fitting a straight line to the data points to minimize the difference between the observed values and the values predicted by the model [117].

In linear regression, the dependent variable is often referred to as the response variable, while the independent variables are known as predictors or explanatory variables [118]. The relationship between these variables is assumed to be linear, meaning that a change in the independent variable leads to a proportional change in the dependent variable [119]. The model estimates the coefficients of the independent variables to quantify their impact on the dependent variable, allowing for predictions and inference based on the established linear relationship [120].

Linear regression is particularly useful for understanding the association between variables, making predictions, and identifying trends in the data [121]. It provides insights into how changes in the independent variables affect the outcome, enabling researchers to make informed decisions based on the model's results [122]. Linear regression also serves as a foundation for more complex regression techniques and machine learning algorithms, making it a crucial tool in data analysis and predictive modeling [123].

One of the key advantages of linear regression is its interpretability, as the coefficients in the model represent the strength and direction of the relationships between variables [124]. This transparency allows researchers to understand the impact of each predictor on the outcome, facilitating the identification of significant factors influencing the dependent variable [125]. Additionally, linear regression can be extended to handle multiple predictors through multiple linear regression, enabling the analysis of more complex relationships in the data [126].

Despite its advantages, linear regression has certain assumptions that need to be met for the model to be valid. These assumptions include linearity, independence of errors, homoscedasticity (constant variance of errors), and normality of residuals [119]. Violations of these assumptions can lead to biased estimates and inaccurate predictions, highlighting the importance of assessing the model's validity before drawing conclusions from the results [127].

In practice, linear regression is applied in a wide range of scenarios, from predicting stock prices and estimating medical costs to analyzing climate data and forecasting economic indicators [128]. Researchers have also explored variations of linear regression, such as quantile regression and Bayesian linear regression, to address specific research questions and improve model performance [129].

5.1.1. The Mathematical Model

Linear regression assumes the following linear relationship between the dependent variable y (wine quality) and the independent variables X [130]:

$$y = \beta_0 + \sum_{i=1}^{n} \beta_i x_i + \epsilon \tag{26}$$

where

- y is the dependent variable representing wine quality.
- x_i are the independent variables representing chemical properties such as fixed acidity, volatile acidity, and so on.
- β_0 is the intercept, the value of y when all $x_i = 0$.
- β_i are the coefficients, which indicate how each independent variable affects the quality of the wine.
- ϵ is the error term, representing the residuals or the difference between the actual and predicted values.

5.1.2. Objective: Minimizing the Error

The goal of linear regression is to find the optimal coefficients (β) that minimize the error between the predicted and actual values of y [131]. This error is typically measured using the sum of squared residuals, also known as the "least squares" method. The residual for each data point i is calculated as

$$r_i = y_i - \hat{y}_i \tag{27}$$

where $\hat{y}_i$ is the predicted value of y_i. The sum of squared residuals is then

$$SS_{res} = \sum_{i=1}^{n} \left(y_i - \left(\beta_0 + \sum_{j=1}^{n} \beta_j x_j \right) \right)^2 \quad (28)$$

5.1.3. Ordinary Least Squares (OLS) [132]

To minimize the sum of squared residuals, the method of Ordinary Least Squares (OLS) is used. OLS calculates the coefficients $\beta_0, \beta_1, \ldots, \beta_n$ that minimize SS_{res}. The optimal coefficients are derived using the following normal equation:

$$b = (X^T X)^{-1} X^T y \quad (29)$$

where

- b is the vector of coefficients $[\beta_0, \beta_1, \ldots, \beta_n]$.
- X is the matrix of input features, augmented with a column of ones for the intercept.
- y is the vector of observed values of y.

5.1.4. Model Evaluation

Once the coefficients are determined, the linear regression model is evaluated using various metrics, including

- R-squared (R^2): Measures the proportion of variance in the dependent variable that is predictable from the independent variable(s). An R^2 close to 1 indicates a good fit.
- Mean Squared Error (MSE): The average of the squared residuals.
- Root Mean Squared Error (RMSE): The square root of the MSE.

5.1.5. Application to the Wine Dataset

In our wine quality dataset, we applied linear regression to model the relationship between the wine's chemical properties and its quality rating. We derived a linear equation:

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_n x_n + \epsilon \quad (30)$$

where y is the wine quality, x_i represents the chemical properties, and β_i are the coefficients that define the influence of each property on the wine quality. By minimizing the error, we ensured that the equation closely approximated the actual quality ratings given the chemical properties. This method provided a clear and interpretable model for predicting wine quality.

5.2. Linear Regression with Second-Degree Polynomial

Linear regression can be extended to model nonlinear relationships by introducing polynomial features. This is known as polynomial regression, and it still maintains the properties of a linear model, but allows for more complex relationships between the variables. In the context of our wine quality dataset, this approach can capture the effects of interactions and quadratic terms between the chemical properties. The use of second-degree polynomials in regression analysis is particularly advantageous when the data exhibit a curved pattern that cannot be adequately captured by a simple linear model [133].

Researchers in various fields have applied second-degree polynomial regression to address specific challenges and make accurate predictions. For instance, in agricultural research, second-degree polynomial regression has been utilized to predict sugarcane yield, demonstrating its effectiveness in capturing the complex distribution of variables and providing a better fit for the data [134]. In healthcare, researchers have employed second-

degree fractional polynomials in linear regression models to identify factors associated with post-stunting linear growth in under-five children, showcasing the versatility of this approach in understanding growth patterns [133].

The application of polynomial regression extends beyond traditional fields to encompass machine learning and statistical modeling. In machine learning, polynomial regression has been integrated into predictive models to account for non-linear relationships between predictors and responses, offering a more nuanced understanding of complex data patterns [135]. Additionally, in resource management in edge servers, polynomial regression has been utilized to model capacity factors, highlighting its utility in optimizing system performance and resource allocation [136].

In environmental studies, second-degree polynomial regression has been instrumental in water quality index trend analysis, where polynomial regression models of varying degrees were compared to assess water quality trends accurately [137]. Additionally, in the estimation of chlorophyll and nitrogen contents in plants, second-degree polynomials have been employed to establish correlations between non-destructive and destructive methods, emphasizing the role of polynomial regression in bridging different measurement techniques [138].

In mathematical and computational research, polynomial regression has been leveraged to address diverse challenges. For instance, in the study of hyperbolic polynomials with a span less than four, researchers have explored the construction and properties of such polynomials using algebraic operations, showcasing the analytical power of polynomial regression in mathematical investigations [139]. In the context of differential equations with polynomial coefficients, polynomial regression has been used to analyze the behavior of polynomial solutions, providing insights into the characteristics of these equations [140].

5.2.1. The Mathematical Model

For second-degree polynomial regression, the model takes the form

$$\begin{aligned} y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n \\ + \beta_{11} x_1^2 + \ldots + \beta_{nn} x_n^2 + \beta_{12} x_1 x_2 + \ldots + \epsilon \end{aligned} \tag{31}$$

where

- y is the dependent variable.
- x_i are the independent variables.
- β_0 is the intercept.
- β_i are the coefficients for the linear terms.
- β_{ii} are the coefficients for the quadratic terms.
- β_{ij} are the coefficients for the interaction terms.
- ϵ is the error term.

5.2.2. Objective: Minimizing the Error

Like linear regression, the objective of polynomial regression is to minimize the sum of squared residuals:

$$SS_{res} = \sum_{i=1}^{n} \left(y_i - \left(\beta_0 + \sum_{j=1}^{n} \beta_j x_j + \sum_{k=1}^{n} \beta_{kk} x_k^2 + \sum_{j<k} \beta_{jk} x_j x_k \right) \right)^2 \tag{32}$$

The coefficients can be calculated using the Ordinary Least Squares (OLS) method:

$$b = (X^T X)^{-1} X^T y \tag{33}$$

where

- b is the vector of coefficients $[\beta_0, \beta_1, \ldots, \beta_{nn}]$.
- X is the matrix of input features, augmented with a column of ones for the intercept.
- y is the vector of observed values of y.

5.2.3. Application to the Wine Dataset

In our wine quality dataset, we applied second-degree polynomial regression to model the relationship between the wine's chemical properties and its quality rating. This allowed us to capture more complex relationships between the variables. We derived an equation of the form

$$\begin{aligned} y = \beta_0 + \beta_1 x_1 + \ldots + \beta_n x_n \\ + \beta_{11} x_1^2 + \ldots + \beta_{nn} x_n^2 + \beta_{12} x_1 x_2 + \ldots + \epsilon \end{aligned} \tag{34}$$

where y is the wine quality, x_i represents the chemical properties, and β_i are the coefficients that define the influence of each property on the wine quality. By incorporating polynomial terms, this model provided a flexible and effective way to predict wine quality.

5.3. Ridge Regression

Ridge regression is a type of linear regression that addresses some of the limitations of ordinary linear regression, particularly when dealing with multicollinearity or overfitting. It is also known as Tikhonov regularization [141]. Ridge regression, a modification of the Ordinary Least Squares (OLS) method, is a regularization technique used in linear regression to address multicollinearity and improve the stability of coefficient estimates [142]. By penalizing the L2-norm of the coefficients, ridge regression introduces a degree of bias that helps regulate the estimated coefficients, making them more robust in the presence of high-dimensional regressors [143]. This method is particularly valuable when dealing with data exhibiting multicollinearity, where independent variables are highly correlated, as ridge regression provides a solution to the issue of unstable coefficient estimates [144].

The theoretical foundation and practical applications of ridge regression have been extensively developed, establishing it as a well-recognized technique in regression analysis [145]. By adding a small positive quantity to the regression coefficients, ridge regression acts as a biased estimation method that offers more stable results compared to traditional linear regression, especially in scenarios with multicollinearity in the independent variables [146]. The introduction of a biasing parameter in ridge regression mitigates the effects of multicollinearity and leads to more reliable estimation of coefficients [147].

In model selection and forecasting, ridge regression has proven to be an efficient tool for handling multicollinearity and outliers in datasets, making it a preferred choice when dealing with non-normal residuals and data anomalies [148]. Researchers have also explored the use of ridge regression in genetic studies, demonstrating its versatility across different fields and its ability to provide robust parameter estimation and prediction in the presence of outliers and multicollinearity [149]. Furthermore, the application of ridge regression in logistic regression models has led to the development of improved estimators to overcome the challenges posed by multicollinearity in logistic regression analysis [150].

A key advantage of ridge regression is its ability to produce consistent estimates of predictor variables even in the presence of multicollinearity, a common issue in regression analysis [151]. By introducing a regularization term that shrinks the coefficients towards zero, ridge regression helps prevent overfitting and improves the generalization capabilities of the model [152]. Ridge regression has also been compared to other regularization approaches like LASSO and Elastic Net, showcasing its effectiveness in handling multicollinearity and improving the stability of coefficient estimates [142].

The optimization of ridge parameters in regression analysis is crucial for achieving accurate and reliable results. Researchers have proposed various methods for determining the optimal ridge parameter, such as using the ridge trace and selecting a value that stabilizes the regression coefficients [153]. The choice of the ridge parameter significantly impacts the performance of ridge regression models, as it directly influences the bias–variance trade-off and the overall predictive accuracy of the model [154].

5.3.1. The Mathematical Model

Ridge regression introduces a regularization term to the linear regression equation to penalize large coefficients [155]. The equation is similar to that of linear regression, but with an additional term:

$$y = \beta_0 + \sum_{i=1}^{n} \beta_i x_i + \epsilon \tag{35}$$

However, the objective function that ridge regression minimizes is

$$\text{Minimize} \quad \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \tag{36}$$

where

- y is the dependent variable.
- x_i are the independent variables.
- β_j are the coefficients.
- λ is the regularization parameter.

The regularization parameter λ controls the amount of shrinkage. If λ is 0, ridge regression is identical to linear regression. As λ increases, the coefficients β_j shrink towards zero, helping to prevent overfitting.

5.3.2. Objective: Minimizing the Error with Regularization

The goal of ridge regression is to minimize the sum of squared residuals, similar to linear regression, but with an additional penalty on the size of the coefficients. This regularization helps prevent the model from fitting the noise in the data.

The coefficients are calculated using the following formula:

$$b = (X^T X + \lambda I)^{-1} X^T y \tag{37}$$

where

- b is the vector of coefficients $[\beta_0, \beta_1, \ldots, \beta_n]$.
- X is the matrix of input features, augmented with a column of ones for the intercept.
- y is the vector of observed values of y.
- I is the identity matrix.
- λ is the regularization parameter.

5.3.3. Comparison with Linear Regression

Ridge regression differs from ordinary linear regression by adding a penalty term to the loss function. While linear regression seeks to minimize the sum of squared residuals, ridge regression also minimizes the size of the coefficients. This makes ridge regression more robust when the data has multicollinearity or when the model is prone to overfitting.

5.3.4. Model Evaluation

Like linear regression, ridge regression can be evaluated using metrics such as

- R-squared (R^2): Measures the proportion of variance in the dependent variable that is predictable from the independent variable(s). An R^2 close to 1 indicates a good fit.
- Mean Squared Error (MSE): The average of the squared residuals.
- Root Mean Squared Error (RMSE): The square root of the MSE.

5.3.5. Application to the Wine Dataset

In our wine quality dataset, we applied ridge regression to model the relationship between the wine's chemical properties and its quality rating. The regularization helped prevent overfitting, and we derived the following equation:

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_n x_n + \epsilon \quad (38)$$

where y is the wine quality, x_i represents the chemical properties, and β_i are the coefficients that define the influence of each property on the wine quality. By incorporating regularization, ridge regression provided a more robust model for predicting wine quality.

5.4. Ridge Regression with Second-Degree Polynomial

Ridge regression is a type of linear regression that includes a regularization term to prevent overfitting, particularly useful when dealing with complex relationships and multicollinearity. Ridge regression can also be extended to handle polynomial features, allowing it to capture nonlinear relationships effectively. In the context of our wine quality dataset, ridge regression with second-degree polynomials can model interactions and quadratic effects between the chemical properties. Ridge regression is employed to estimate the coefficients of multiple regression models, particularly when addressing highly correlated independent variables. This technique introduces a regularization term to prevent overfitting by shrinking the regression coefficients towards zero [136]. Incorporating a second-degree polynomial into the model allows researchers to capture nonlinear relationships between variables, enabling more nuanced and accurate predictions [156].

The integration of a second-degree polynomial within the ridge regression framework provides a robust approach to modeling data exhibiting curvature or nonlinearity. This hybrid model utilizes the regularization properties of ridge regression to address multicollinearity and prevent model overfitting while leveraging the flexibility of polynomial functions to capture intricate patterns in the data [157]. By combining these techniques, researchers can strike a balance between bias and variance, leading to more robust and reliable regression models [158].

In practical applications, ridge regression with a second-degree polynomial has demonstrated versatility and effectiveness across various fields. For instance, in predicting the progression of COVID-19 outbreaks, a hybrid polynomial–Bayesian ridge regression model was developed to forecast the spread of the virus, demonstrating the utility of this approach in epidemiological modeling [156]. Similarly, in fault detection for ship systems operations, a multiple polynomial ridge regression model accurately detected developing faults in engine parameters, highlighting the applicability of this technique in predictive maintenance [157].

By creating a ridge regression estimator within a polynomial framework, researchers can effectively handle scenarios with complex interdependencies among variables, ensuring robust and stable model performance [158]. This approach has also been applied in the analysis of monthly and annual rainfall variability, demonstrating its efficacy in handling geographical variables and predicting climatological phenomena [159].

Furthermore, the use of second-degree polynomial regression within the ridge regression framework has been pivotal in fields such as agriculture and environmental science. For instance, in predicting sugarcane yield, researchers utilized second-degree polyno-

mial regression to achieve the best fit for the distribution of variables, emphasizing the importance of polynomial functions in capturing the nonlinear relationships inherent in agricultural data [134]. Similarly, in water quality trend analysis, higher-degree polynomial models, including second-degree polynomials, provided a good fit to the data, highlighting the utility of polynomial regression in environmental research [137].

5.4.1. The Mathematical Model

For ridge regression with second-degree polynomials, the model takes the form

$$\begin{aligned} y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n \\ + \beta_{11} x_1^2 + \ldots + \beta_{nn} x_n^2 + \beta_{12} x_1 x_2 + \ldots + \epsilon \end{aligned} \tag{39}$$

However, the objective function that ridge regression minimizes is

$$\text{Minimize} \quad \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \tag{40}$$

where

- y is the dependent variable.
- x_i are the independent variables.
- β_j are the coefficients.
- λ is the regularization parameter.

The regularization parameter λ controls the amount of shrinkage. As λ increases, the coefficients β_j shrink towards zero, helping to prevent overfitting.

5.4.2. Application to the Wine Dataset

In our wine quality dataset, we applied ridge regression with second-degree polynomials to model the relationship between the wine's chemical properties and its quality rating. The regularization helped prevent overfitting, and we derived the following equation:

$$\begin{aligned} y = \beta_0 + \beta_1 x_1 + \ldots + \beta_n x_n \\ + \beta_{11} x_1^2 + \ldots + \beta_{nn} x_n^2 \\ + \beta_{12} x_1 x_2 + \ldots + \epsilon \end{aligned} \tag{41}$$

where y is the wine quality, x_i represents the chemical properties, and β_i are the coefficients that define the influence of each property on the wine quality. By incorporating regularization, ridge regression provided a more robust model for predicting wine quality, particularly when the relationships were complex or when overfitting was a concern.

6. Applying Regression Models and Obtaining Equations

6.1. General Information for All Equations

6.1.1. Log Transformation

The equations are based on the log-transformed features for certain attributes, which means that for a given feature x, the transformed value x' is computed as follows:

$$x' = \log(x + 1) \tag{42}$$

The log transformation is applied to the following features:

- Fixed acidity;
- Volatile acidity;

- Citric acid;
- Residual sugar;
- Chlorides;
- Free sulfur dioxide;
- Total sulfur dioxide.

6.1.2. How to Use the Equations

To use these equations for predicting the wine quality based on new data, the relevant features should be log-transformed as specified above.

- For a new wine sample with attributes $x_1, x_2, \ldots, x_{11}$, where
  - $x_1 =$ fixed acidity;
  - $x_2 =$ volatile acidity;
  - $x_3 =$ citric acid;
  - $x_4 =$ residual sugar;
  - $x_5 =$ chlorides;
  - $x_6 =$ free sulfur dioxide;
  - $x_7 =$ total sulfur dioxide;
  - $x_8 =$ density;
  - $x_9 =$ pH;
  - $x_{10} =$ sulphates;
  - $x_{11} =$ alcohol.
- The log-transformed values should be

$$x_i' = \log(x_i + 1) \quad \text{for } i = 1, 2, \ldots, 7 \tag{43}$$

- For features x_8, x_9, x_{10}, x_{11}, the original values are used directly.
- The equations can then be used directly with these transformed values.

6.1.3. Accuracy Calculation

$$\text{Accuracy} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}\left(\lfloor y_i + 0.5 \rfloor = y_i^{\text{true}}\right) \tag{44}$$

where

- n is the number of samples,
- y_i is the predicted quality for the ith sample,
- y_i^{true} is the actual quality for the ith sample,
- $\mathbf{1}(\cdot)$ is the indicator function.

This formula computes the fraction of samples where the rounded predicted quality ($\lfloor y_i + 0.5 \rfloor$) matches the actual quality.

Given an input y_i,

(i) The value $y_i + 0.5$ is computed first, which simply adds 0.5 to the value of y_i.
(ii) The floor function $\lfloor \cdot \rfloor$ is then applied to $y_i + 0.5$. This rounds y_i to the nearest integer.

To explain this with an example, suppose $y_i = 3.4$:

(i) $y_i + 0.5 = 3.4 + 0.5 = 3.9$
(ii) $\lfloor 3.9 \rfloor = 3$, because 3 is the largest integer less than or equal to 3.9.

Now, if $y_i = 3.6$,

(i) $y_i + 0.5 = 3.6 + 0.5 = 4.1$

(ii) $\lfloor 4.1 \rfloor = 4$, because 4 is the largest integer less than or equal to 4.1.

Thus, the expression $\lfloor y_i + 0.5 \rfloor$ is a common way to round a real number y_i to the nearest integer. The floor function $\lfloor \cdot \rfloor$ rounds down to the largest integer smaller than or equal to the given value. The trick of adding 0.5 before applying the floor function effectively rounds y_i to the nearest integer. And the indicator function, often denoted as $\mathbf{1}(\cdot)$, is a mathematical function used to indicate whether a certain condition is true or false.

These results indicate that both models performed similarly, with the linear regression model being slightly more accurate.

6.1.4. Repeatability of Experiments

To ensure the repeatability of our experiments and to facilitate a fair comparison between different algorithms, we used a consistent random seed and a fixed test size for our data splits. Specifically, we set `random_state = 42` and `test_size = 0.2`. This standardization ensures that if others follow the same procedures using Python, they will obtain identical results.

For logarithmic transformations, we used the natural logarithm with a base e. The base e is approximately equal to 2.71828. For example, e to the power of 2 is approximately 7.3891, and the natural logarithm of 7.3891 is 2.

When using Python, it is important to note the distinction between `np.log(x)` and `np.log1p(x)`. The function `np.log(x)` computes the natural logarithm of x. In contrast, `np.log1p(x)` computes the natural logarithm of $1 + x$, which is equivalent to $\ln(1 + x)$. The `np.log1p(x)` function is preferred because it helps avoid precision issues when x is close to zero.

For those using other statistical tools such as MATLAB, SAS, or SPSS, the equivalent function to use is $\ln(1 + x)$ or $\log_e(1 + x)$. This consistency ensures that results across different platforms remain comparable.

6.2. Linear Regression Equation

Given the wine dataset, we first predicted the quality of the wine (y) using linear regression; the equation is in Table 8.

Table 8. Linear regression equation.

$y = 16.5244 + 0.4318 \cdot \log(\text{fixed acidity} + 1) - 1.6438 \cdot \log(\text{volatile acidity} + 1)$
$-0.3281 \cdot \log(\text{citric acid} + 1) + 0.0358 \cdot \log(\text{residual sugar} + 1)$
$-1.9195 \cdot \log(\text{chlorides} + 1) + 0.1374 \cdot \log(\text{free sulfur dioxide} + 1)$
$-0.1773 \cdot \log(\text{total sulfur dioxide} + 1) - 13.3040 \cdot \text{density}$
$-0.2703 \cdot \text{pH} + 0.8345 \cdot \text{sulphates} + 0.2794 \cdot \text{alcohol}$

Model Performance

The linear regression model yielded the following performance metrics:

- RMSE: 0.6731
- Accuracy: 0.5844

6.3. Ridge Regression Equation

Given the wine dataset, we then predicted the quality of the wine (y) using ridge regression; the equation is in Table 9.

Table 9. Ridge regression equation.

$$y = 3.3837 + 0.3218 \cdot \log(\text{fixed acidity} + 1) - 1.5670 \cdot \log(\text{volatile acidity} + 1)$$
$$-0.2715 \cdot \log(\text{citric acid} + 1) - 0.0099 \cdot \log(\text{residual sugar} + 1)$$
$$-1.2091 \cdot \log(\text{chlorides} + 1) + 0.1447 \cdot \log(\text{free sulfur dioxide} + 1)$$
$$-0.1817 \cdot \log(\text{total sulfur dioxide} + 1) - 0.0101 \cdot \text{density} - 0.2875 \cdot \text{pH}$$
$$+0.7380 \cdot \text{sulphates} + 0.2988 \cdot \text{alcohol}$$

Model Performance

The ridge regression model yielded the following performance metrics:

- RMSE: 0.6801
- Accuracy: 0.5750

6.4. Second-Degree Linear Regression Equation

Given the wine dataset, we then predicted the quality of the wine (y) using linear regression with second-degree polynomial features; the equation is in Table 10.

Table 10. Second-degree linear regression equation.

$$y = -19532.8969 - 828.2545 \cdot \log(\text{fixed acidity} + 1) - 178.4674 \cdot \log(\text{volatile acidity} + 1)$$
$$-469.2032 \cdot \log(\text{citric acid} + 1) - 87.6453 \cdot \log(\text{residual sugar} + 1)$$
$$-390.8942 \cdot \log(\text{chlorides} + 1) - 48.6337 \cdot \log(\text{free sulfur dioxide} + 1)$$
$$+30.7540 \cdot \log(\text{total sulfur dioxide} + 1) + 43721.3870 \cdot \text{density} - 948.3317 \cdot \text{pH}$$
$$+51.7720 \cdot \text{sulphates} + 76.8861 \cdot \text{alcohol} - 6.5101 \cdot \log(\text{fixed acidity} + 1)^2$$
$$-3.2325 \cdot \log(\text{fixed acidity} + 1) \cdot \log(\text{volatile acidity} + 1)$$
$$-4.8311 \cdot \log(\text{fixed acidity} + 1) \cdot \log(\text{citric acid} + 1)$$
$$-1.3614 \cdot \log(\text{fixed acidity} + 1) \cdot \log(\text{residual sugar} + 1)$$
$$-11.3986 \cdot \log(\text{fixed acidity} + 1) \cdot \log(\text{chlorides} + 1)$$
$$-0.3382 \cdot \log(\text{fixed acidity} + 1) \cdot \log(\text{free sulfur dioxide} + 1)$$
$$-0.1987 \cdot \log(\text{fixed acidity} + 1) \cdot \log(\text{total sulfur dioxide} + 1)$$
$$+885.0942 \cdot \log(\text{fixed acidity} + 1) \cdot \text{density} - 8.1624 \cdot \log(\text{fixed acidity} + 1) \cdot \text{pH}$$
$$+0.8050 \cdot \log(\text{fixed acidity} + 1) \cdot \text{sulphates} + 0.8007 \cdot \log(\text{fixed acidity} + 1) \cdot \text{alcohol}$$
$$-1.5159 \cdot \log(\text{volatile acidity} + 1)^2 + 0.7501 \cdot \log(\text{volatile acidity} + 1) \cdot \log(\text{citric acid} + 1)$$
$$-1.0388 \cdot \log(\text{volatile acidity} + 1) \cdot \log(\text{residual sugar} + 1)$$
$$+8.6115 \cdot \log(\text{volatile acidity} + 1) \cdot \log(\text{chlorides} + 1)$$
$$-1.2481 \cdot \log(\text{volatile acidity} + 1) \cdot \log(\text{free sulfur dioxide} + 1)$$
$$+1.9331 \cdot \log(\text{volatile acidity} + 1) \cdot \log(\text{total sulfur dioxide} + 1)$$
$$+187.3941 \cdot \log(\text{volatile acidity} + 1) \cdot \text{density}$$
$$-3.0621 \cdot \log(\text{volatile acidity} + 1) \cdot \text{pH}$$
$$-3.0658 \cdot \log(\text{volatile acidity} + 1) \cdot \text{sulphates}$$
$$+0.7150 \cdot \log(\text{volatile acidity} + 1) \cdot \text{alcohol}$$

Model Performance

The linear regression with second-degree polynomial model yielded the following performance metrics:

- RMSE: 0.6778
- Accuracy: 0.5875

6.5. Second-Degree Ridge Regression Equation

Given the wine dataset, we then predicted the quality of the wine (y) using ridge regression with second-degree polynomial features; the equation is in Table 11.

Table 11. Second-degree ridge regression equation.

$$
\begin{aligned}
y = {} & 1.4010 + 0.1332 \cdot \log(\text{fixed acidity} + 1) - 0.1449 \cdot \log(\text{volatile acidity} + 1) \\
& + 0.0107 \cdot \log(\text{citric acid} + 1) + 0.0477 \cdot \log(\text{residual sugar} + 1) \\
& + 0.0158 \cdot \log(\text{chlorides} + 1) + 0.0512 \cdot \log(\text{free sulfur dioxide} + 1) \\
& + 0.1154 \cdot \log(\text{total sulfur dioxide} + 1) - 0.0095 \cdot \text{density} - 0.0778 \cdot \text{pH} \\
& + 0.1245 \cdot \text{sulphates} + 0.3976 \cdot \text{alcohol} - 0.3019 \cdot \log(\text{fixed acidity} + 1)^2 \\
& - 0.3202 \cdot \log(\text{fixed acidity} + 1) \cdot \log(\text{volatile acidity} + 1) \\
& + 0.2113 \cdot \log(\text{fixed acidity} + 1) \cdot \log(\text{citric acid} + 1) \\
& + 0.1456 \cdot \log(\text{fixed acidity} + 1) \cdot \log(\text{residual sugar} + 1) \\
& - 0.0295 \cdot \log(\text{fixed acidity} + 1) \cdot \log(\text{chlorides} + 1) \\
& - 0.1328 \cdot \log(\text{fixed acidity} + 1) \cdot \log(\text{free sulfur dioxide} + 1) \\
& + 0.0663 \cdot \log(\text{fixed acidity} + 1) \cdot \log(\text{total sulfur dioxide} + 1) \\
& + 0.1104 \cdot \log(\text{fixed acidity} + 1) \cdot \text{density} + 0.4407 \cdot \log(\text{fixed acidity} + 1) \cdot \text{pH} \\
& + 0.1533 \cdot \log(\text{fixed acidity} + 1) \cdot \text{sulphates} - 0.0449 \cdot \log(\text{fixed acidity} + 1) \cdot \text{alcohol} \\
& - 0.4413 \cdot \log(\text{volatile acidity} + 1)^2 + 0.1403 \cdot \log(\text{volatile acidity} + 1) \cdot \log(\text{citric acid} + 1) \\
& - 0.2926 \cdot \log(\text{volatile acidity} + 1) \cdot \log(\text{residual sugar} + 1) \\
& + 0.0417 \cdot \log(\text{volatile acidity} + 1) \cdot \log(\text{chlorides} + 1) \\
& - 0.4266 \cdot \log(\text{volatile acidity} + 1) \cdot \log(\text{free sulfur dioxide} + 1) \\
& + 0.7298 \cdot \log(\text{volatile acidity} + 1) \cdot \log(\text{total sulfur dioxide} + 1) \\
& - 0.1480 \cdot \log(\text{volatile acidity} + 1) \cdot \text{density} \\
& - 0.3012 \cdot \log(\text{volatile acidity} + 1) \cdot \text{pH} \\
& + 0.0879 \cdot \log(\text{volatile acidity} + 1) \cdot \text{sulphates} - 0.0234 \cdot \log(\text{volatile acidity} + 1) \cdot \text{alcohol}
\end{aligned}
$$

Model Performance

The ridge regression with second-degree polynomial model yielded the following performance metrics:

- RMSE: 0.6614
- Accuracy: 0.6000

6.6. Equations with Third-Degree Polynomial Features

When we derived the third-degree equations for predicting wine quality, our results were as follows: 'Linear Regression': 'RMSE': 0.8023, 'Accuracy': 0.5594, 'Ridge Regression': 'RMSE': 0.6778, 'Accuracy': 0.5875. This means that our accuracies decreased and RMSE increased. Therefore, we can conclude that, up to now, the second-degree ridge regression formula is the best among all for predicting wine quality.

6.7. Best Second-Degree Ridge Regression Equation

In our study, we systematically explore all 2047 possible combinations of 11 wine characteristics to determine their predictive power using a second-degree polynomial ridge regression model. This analysis is facilitated by the following combination formula:

$$C(n, x) = \frac{n!}{x!(n-x)!} \tag{45}$$

where

- $n!$ is the factorial of n, representing the product of all integers up to n.
- $x!$ is the factorial of x.
- $(n-x)!$ is the factorial of the difference between n and x.

This formula calculates the number of different ways to select x features from a set of n features, irrespective of order. For our dataset comprising 11 features, we calculate combinations for each subset from 1 to 11, totaling $2^{11} - 1 = 2047$ non-empty subsets.

Each subset undergoes evaluation to determine the effectiveness of its corresponding second-degree polynomial ridge regression model in predicting wine quality. We contin-

uously refine our search until identifying the subset that produces the highest accuracy, thereby optimizing our predictive model.

Given the following 11 features: fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulfates, and alcohol, we considered all possible non-empty combinations of these features. For each combination, we fit a second-degree polynomial ridge regression model and evaluated its performance based on Root Mean Square Error (RMSE) and accuracy. The optimal combination, yielding the highest accuracy and lowest RMSE, was found to be ***the optimized best second-degree ridge regression equation***; the equation is in Table 12.

Table 12. Best second-degree ridge regression equation.

$$
\begin{aligned}
y = {} & -0.0591 + 0.0897 \cdot \log(\text{fixed acidity} + 1) + 0.1011 \cdot \log(\text{residual sugar} + 1) \\
& -0.0101 \cdot \log(\text{free sulfur dioxide} + 1) + 0.0710 \cdot \log(\text{total sulfur dioxide} + 1) \\
& -0.0080 \cdot \text{pH} - 0.0168 \cdot \text{sulphates} + 0.8735 \cdot \text{alcohol} \\
& -0.1538 \cdot \log(\text{fixed acidity} + 1)^2 + 0.1669 \cdot \log(\text{fixed acidity} + 1) \cdot \log(\text{residual sugar} + 1) \\
& -0.0790 \cdot \log(\text{fixed acidity} + 1) \cdot \log(\text{free sulfur dioxide} + 1) \\
& -0.0678 \cdot \log(\text{fixed acidity} + 1) \cdot \log(\text{total sulfur dioxide} + 1) \\
& +0.3872 \cdot \log(\text{fixed acidity} + 1) \cdot \text{pH} - 0.0335 \cdot \log(\text{fixed acidity} + 1) \cdot \text{sulphates} \\
& -0.0356 \cdot \log(\text{fixed acidity} + 1) \cdot \text{alcohol} + 0.1188 \cdot \log(\text{residual sugar} + 1)^2 \\
& +0.2192 \cdot \log(\text{residual sugar} + 1) \cdot \log(\text{free sulfur dioxide} + 1) \\
& -0.1412 \cdot \log(\text{residual sugar} + 1) \cdot \log(\text{total sulfur dioxide} + 1) \\
& -0.0235 \cdot \log(\text{residual sugar} + 1) \cdot \text{pH} + 0.0508 \cdot \log(\text{residual sugar} + 1) \cdot \text{sulphates} \\
& -0.0894 \cdot \log(\text{residual sugar} + 1) \cdot \text{alcohol} - 0.0580 \cdot \log(\text{free sulfur dioxide} + 1)^2 \\
& +0.0005 \cdot \log(\text{free sulfur dioxide} + 1) \cdot \log(\text{total sulfur dioxide} + 1) \\
& -0.0591 \cdot \log(\text{free sulfur dioxide} + 1) \cdot \text{pH} - 0.4979 \cdot \log(\text{free sulfur dioxide} + 1) \cdot \text{sulphates} \\
& +0.0865 \cdot \log(\text{free sulfur dioxide} + 1) \cdot \text{alcohol} - 0.0502 \cdot \log(\text{total sulfur dioxide} + 1)^2 \\
& +0.4780 \cdot \log(\text{total sulfur dioxide} + 1) \cdot \text{pH} - 0.3883 \cdot \log(\text{total sulfur dioxide} + 1) \cdot \text{sulphates} \\
& -0.0898 \cdot \log(\text{total sulfur dioxide} + 1) \cdot \text{alcohol} - 0.3737 \cdot \text{pH}^2 \\
& +0.7016 \cdot \text{pH sulphates} - 0.0998 \cdot \text{pH alcohol} - 0.6216 \cdot \text{sulphates}^2 \\
& +0.2689 \cdot \text{sulphates alcohol} - 0.0046 \cdot \text{alcohol}^2
\end{aligned}
$$

Model Performance

The best second-degree ridge regression model with seven variables yielded the following performance metrics:

- RMSE: 0.6801
- Accuracy: 0.6031

This improved accuracy compared to the previous best model, indicating that the selected feature combination and polynomial degree yielded a slightly better formula for predicting wine quality.

We choose the equation with seven variables as the best performing because models with fewer features are less likely to overfit the training data, which means they are more likely to generalize well to unseen data. This model with seven variables is simpler, less prone to overfitting, easier to interpret, and computationally more efficient, which are all very desirable characteristics.

Overfitting occurs when a model learns not only the underlying pattern in the training data but also the noise or random fluctuations. This happens when the model is too complex relative to the amount of data available. As a result, while the model performs exceptionally well on the training data, it often performs poorly on unseen data.

Our best-performing variables in the optimized equation are as follows: fixed acidity, residual sugar, free sulfur dioxide, total sulfur dioxide, pH, sulfates, alcohol. There are seven variables in total in our best performing second-degree ridge regression equation.

6.8. Summary of Mathematical Equations for Wine Quality Prediction

In our analysis, we aimed to predict the quality of wine using various mathematical models. Throughout the process, we derived different mathematical equations to predict wine quality. Our objective was to maximize accuracy while minimizing the Root Mean Square Error (RMSE).

6.8.1. Results from Different Models

As seen in Table 13, the highest accuracy we achieved was 0.6031 using an optimized second-degree ridge regression model. The equation for this model, highlighted in yellow below, had an RMSE of 0.6801.

Table 13. Eqaution comparison for wine quality prediction.

Rank	Model	RMSE	Accuracy
1	Best 2nd-Degree Ridge Regression	0.6801	0.6031
2	2nd-Degree Ridge Regression	0.6614	0.6000
3	Linear Regression	0.6731	0.5844
4	2nd-Degree Linear Regression	0.6778	0.5875
5	3rd-Degree Ridge Regression	0.6778	0.5875
6	Ridge Regression	0.6801	0.5750
7	3rd-Degree Linear Regression	0.8023	0.5594

6.8.2. Conclusions

In conclusion, the best second-degree ridge regression equation from Table 12 performed the best in terms of accuracy, indicating that it is a reliable model for predicting wine quality. The other equations performed relatively well, but none surpassed the optimized ridge regression equation in accuracy.

7. Model Comparison for Wine Quality Prediction

In this section, we aimed to compare the accuracy of our optimized second-degree ridge regression equation with different machine learning models. The objective was to determine which model best predicted wine quality using a dataset with several chemical properties as features. We tested various models including linear regression, Lasso Regression, logistic regression, Elastic Net, K-Nearest Neighbors (KNN), neural networks, Naive Bayes, support vector machines (SVM), Principal Component Regression (PCR), partial least squares regression (PLSR), Random Forest, and extreme gradient boosting (XGBoost).

7.1. Data Preparation

We started by loading the dataset, which contains 11 chemical features and the target variable quality. To handle skewness in the data, we applied a logarithmic transformation to certain features for our best second-degree ridge model, such as fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, and total sulfur dioxide. For other models, no transformation was applied, and scaling was only applied when necessary.

7.2. Optimal Feature Selection for Ridge Regression

To identify the best performing model, we used ridge regression with second-degree polynomial features. We evaluated different combinations of the 11 features to find the best subset. We split the data into training and testing sets using an 80–20 split with a fixed *ran-*

dom_state for consistency. Our best subset of features included fixed acidity, residual sugar, free sulfur dioxide, total sulfur dioxide, pH, sulfates, and alcohol. This subset, which consisted of seven features, gave an accuracy of 0.6031 and an RMSE of 0.6801.

7.3. Model Evaluation

We then tested all the other models using the same subset of features identified by our ridge regression for our model, while the other models were tested using all 11 features. Each model was evaluated using accuracy and RMSE as metrics. The results are shown in Table 14:

Table 14. Model comparison for wine quality prediction.

#	Model	RMSE	Acc.	Vars
1	XGBoost	0.6098	0.6750	11
2	Random Forest Classifier	0.6614	0.6469	11
3	Neural Network Classifier	0.6801	0.6406	11
4	Best 2nd Degree Polynomial Ridge	0.6801	0.6031	7
5	Support Vector Machine Classifier	0.6778	0.5969	11
6	KNN Regressor	0.7159	0.5875	11
7	Logistic Regression	0.7071	0.5844	11
8	Linear Regression	0.6892	0.5813	11
9	Principal Component Regression	0.6892	0.5813	11
10	Partial Least Squares Regression	0.6801	0.5750	11
11	KNN Classifier	0.8062	0.5563	11
12	Lasso	0.8678	0.4125	11
13	Elastic Net	0.8678	0.4125	11
14	Naive Bayes	1.3852	0.3344	11

7.4. Conclusions

From Table 14, it is clear that our ridge regression model with second-degree polynomial features performed exceptionally well, ranking fourth out of the fourteen models tested. The Random Forest Regressor and XGBoost models performed better, likely due to their ability to handle complex, nonlinear relationships effectively. Our mathematical model's strong performance demonstrates the effectiveness of selecting the right subset of features and using polynomial transformations for prediction.

With a quality formula, we can understand the marginal effects of changes and determine the best direction to improve quality. This understanding empowers farmers and producers to make informed decisions, leading to higher quality wines. This project showcases how a well-crafted mathematical formula can outperform complex simulations, providing clear and actionable insights into improving wine quality.

Impact of Taster Variability on Model Evaluation

The quality scores in the dataset are inherently subjective, as they are assigned by human tasters based on sensory evaluation. Such assessments are influenced by individual preferences, cultural backgrounds, and environmental factors, resulting in variability in the dependent variable ("y"). This variability establishes a baseline level of noise in the dataset that cannot be eliminated by any predictive model.

When comparing different models, the differences in their errors, such as RMSE or accuracy, must be interpreted in the context of this inherent noise. For example, even if a model achieves a lower RMSE than another, the improvement might not hold significant practical value if the reduction in error is smaller than the natural variability in the quality scores. This underscores the importance of considering the limits of prediction accuracy imposed by the subjectivity of the target variable.

In this study, while models such as XGBoost and Random Forests demonstrated higher predictive accuracy, the ridge regression model was chosen for its interpretability and simplicity. These qualities allow stakeholders in wine production to understand the contribution of individual physicochemical properties to wine quality, facilitating actionable insights. The study's emphasis on interpretability addresses practical concerns in the wine industry, where the ability to analyze and apply model results is as important as achieving high predictive accuracy.

Future work could explore ways to reduce the impact of subjective variability by incorporating additional objective measures or validating the predictions with sensory panels under controlled conditions.

7.5. Practical Applications of Predictive Modeling in Winemaking

The results from this study can provide winemakers with actionable strategies to enhance wine quality. The ridge regression model, with its interpretable coefficients, enables producers to identify key physicochemical attributes that contribute to higher quality scores and make informed adjustments to production processes.

7.5.1. Adjusting Alcohol Content for Optimal Balance

Our model indicates that alcohol has the most significant positive impact on wine quality, with a coefficient of $+0.8735$. This suggests that maintaining an optimal alcohol level enhances sensory attributes and overall balance. Winemakers can control alcohol content by

- Adjusting fermentation temperature to influence yeast activity.
- Selecting yeast strains with different fermentation efficiencies to control ethanol production.
- Implementing partial de-alcoholization techniques if necessary.

7.5.2. Managing Residual Sugar to Optimize Perception

The model assigns a positive coefficient to residual sugar ($+0.1011$), indicating that a moderate amount enhances quality perception. However, excessive sugar combined with sulfur dioxide can be detrimental. Practical steps for winemakers include

- Controlling fermentation duration to leave the desired amount of residual sugar.
- Using selective yeast strains that naturally metabolize sugars while preserving aroma.
- Monitoring sulfur dioxide additions to avoid overpreservation effects that might negatively impact flavor.

7.5.3. Balancing Sulfur Dioxide Levels for Stability

Total sulfur dioxide positively influences quality ($+0.0710$), but free sulfur dioxide (-0.0101) has a slightly negative impact, highlighting the importance of precise sulfur dioxide management. Winemakers can

- Adjust SO_2 additions based on pH and microbial stability requirements.
- Use alternative preservation techniques (e.g., micro-oxygenation) to reduce SO_2 reliance.
- Regularly monitor free and bound SO_2 to maintain balance.

7.5.4. Controlling Acidity and Sulfates for Sensory Harmony

Sulfates (-0.0168) and pH (-0.0080) show minor negative influences, suggesting that extreme values should be avoided to prevent off-flavors. Strategies include

- Adjusting acid content through malolactic fermentation or tartaric acid additions.
- Fine-tuning sulfate concentrations to enhance structure without increasing astringency.
- Blending different batches to achieve a balanced acidity profile.

8. Predicting Out-of-Sample Wine Quality Values

In this section, we delve into the mathematical equation we developed for predicting wine quality. The formula used for our best second-degree polynomial ridge regression was specifically designed to capture the interplay between multiple physicochemical features, including

$$y = -0.0591 + 0.0897 \cdot \log(1 + x_1) + 0.1011 \cdot \log(1 + x_2) + \ldots \tag{46}$$

where the variables x_1 to x_7 represent the relevant features: fixed acidity, residual sugar, free sulfur dioxide, total sulfur dioxide, pH, sulfates, and alcohol. The model captures both linear and quadratic interactions, as well as cross-products between these variables.

Given the formula, we aimed to identify a composition with a predicted quality score just above 10. To achieve this, we evaluated the equation across a dense grid of possible values for each variable. These values span the ranges observed in the training set after applying the log transformation:

$$x_i \in [\min(x_i), \max(x_i)], \text{ for } i = 1, \ldots, 7 \tag{47}$$

For each variable, we divided the range into 10 intervals, creating a grid with 10^7 possible combinations. We then evaluated the equation for each combination, aiming to identify the minimum value just above 10:

$$\text{min quality above 10} = \min\{y \mid y > 10\} \tag{48}$$

Through this exploration, we identified the composition of a hypothetical wine with the minimum predicted quality above 10 as follows:

The minimum predicted quality above 10 is 10.000000

Features for the minimum predicted quality above 10:

- Fixed Acidity: 2.8273
- Residual Sugar: 1.3624
- Free Sulfur Dioxide: 3.4472
- Total Sulfur Dioxide: 2.3597
- pH: 3.0222
- Sulfates: 1.0944
- Alcohol: 14.9000

In Figure 16, we present 3D surface plots of wine quality predictions using our optimal second-degree polynomial ridge regression model. As observed, the model predicts quality values exceeding 10, which surpass the typical upper limit of human taste perception. These graphs aim to enhance understanding of the quality predictions. By varying combinations of two parameters, we examine their impact on predicted quality values. These visualizations help in identifying how changes in these variables influence the quality outcome predicted by the model.

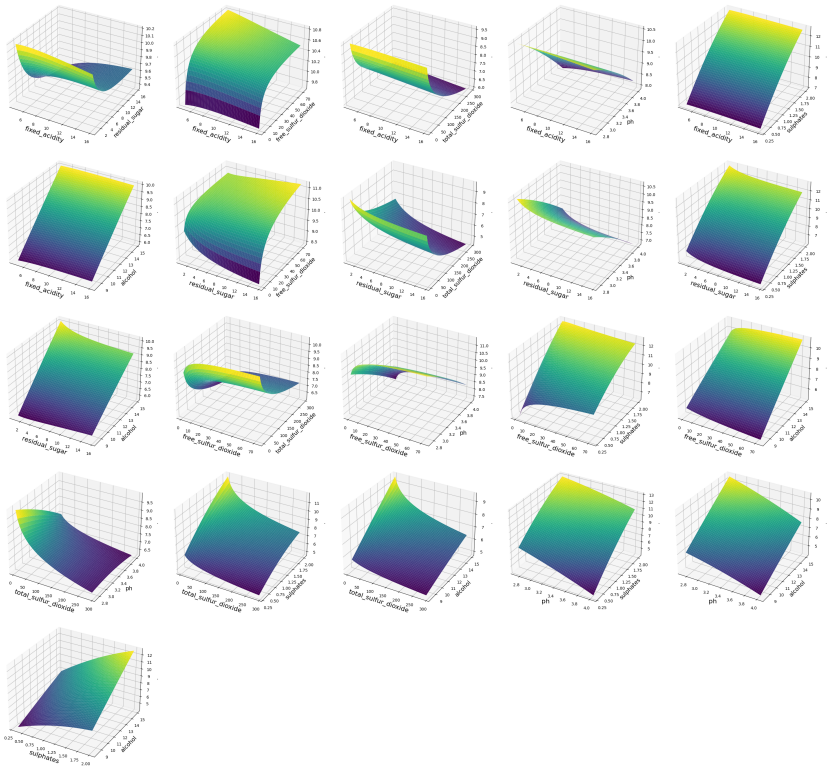


Figure 16. 3D surface plots of wine quality predictions using our best second-degree polynomial ridge regression model.

For each graph, we display the quality value along with two variables. Given the number of combinations of two variables from a set of seven, the total number of unique pair combinations is calculated using the combination formula $\binom{7}{2}$, which yields twenty-one combinations.

When varying the values of these two variables to draw the quality graph, we hold the other five variables fixed at levels corresponding to the previously calculated minimum predicted quality value that exceeds 10.

These findings offer insights into the potential composition of a hypothetical wine just beyond the threshold of human perception.

The predictions of hypothetical quality scores exceeding 10 were designed to explore the potential of the model to identify optimal physicochemical compositions associated with the highest possible quality ratings. These scores, while hypothetical, provide a theoretical framework for guiding future experimental studies. It is important to emphasize that these predictions are not definitive but rather serve as a starting point for identifying promising wine compositions.

Scientifically, these predictions are interpreted as an extrapolation of the relationship between physicochemical properties and quality scores within the observed data range. While the model captures patterns in the dataset, its validity beyond the observed range (quality scores of 3 to 8) is inherently uncertain. Future validation using experimental and sensory evaluations will be necessary to assess the practical applicability of these predictions.

9. Conclusions

This study explored the use of mathematical models for predicting wine quality, aiming to provide a framework for extending predictions beyond traditional human rating scales. By analyzing 11 physicochemical variables of wine and applying various regression

techniques, including linear regression, ridge regression, and polynomial models up to the third degree, we sought to develop an accurate and reliable predictive model.

Through a detailed evaluation, the second-degree polynomial ridge regression model emerged as the most effective approach, balancing predictive performance and simplicity. To optimize the model, we evaluated all possible subsets of the 11 variables, ultimately identifying a reduced set of seven variables that provided the best balance of accuracy and model interpretability. The optimized model achieved an accuracy of 0.6031 and demonstrated robustness by outperforming several other machine learning models.

While ensemble methods such as XGBoost and Random Forests achieved higher accuracy and lower RMSE in this study, they often lack the interpretability necessary for understanding the individual contributions of physicochemical features. Neural networks, while capable of capturing complex, nonlinear relationships, also suffer from a lack of transparency, making it challenging to derive actionable insights.

In contrast, our optimized second-degree polynomial ridge regression model strikes a balance between predictive performance and interpretability. This is particularly valuable in the wine-making context, where understanding the marginal effects of individual features is critical for guiding production decisions and optimizing quality. The ridge regression model's ability to provide a clear mathematical formula enables wine producers to identify and adjust specific attributes to improve wine quality effectively.

Future research could explore the use of ensemble methods or neural networks to benchmark accuracy further, while maintaining a focus on enhancing interpretability through techniques such as feature importance analysis or explainable AI (XAI) approaches.

One of the key contributions of this work is the exploration of hypothetical wine compositions with predicted quality scores extending beyond the conventional upper limit of 10. By leveraging the developed model, we demonstrated the potential to predict and analyze wine quality metrics that surpass human perceptual thresholds. While these predictions are hypothetical, they provide a novel perspective on how predictive modeling can extend beyond traditional boundaries. While this approach offers a novel perspective, it is essential to acknowledge the limitations of extrapolating beyond the observed data range. These predictions are not definitive but rather serve as a theoretical framework to guide future experimental validation. By highlighting potential physicochemical compositions, this study aims to inspire further research and innovation in optimizing wine quality.

This methodology has broader implications for optimizing and exploring the quality of other food products, such as bread, by providing insights into hypothetical compositions that may enhance quality metrics. For instance, in the case of bread, measurable chemical and physical features such as moisture content, protein content, crumb structure, crust color, and elasticity could be associated with a quality score assigned by food experts. By applying a similar modeling approach, it would be possible to predict the optimal composition of these features to achieve higher quality scores than currently observed.

Although the quality parameters of bread differ significantly from those of wine, this study illustrates the versatility of the proposed methodology for predicting and enhancing quality metrics across diverse food products. Future work will prioritize validation within the wine production industry while exploring potential adaptations of the methodology to other food domains.

The study demonstrates that a carefully optimized polynomial ridge regression model can achieve competitive performance while maintaining interpretability, making it suitable for practical applications in wine quality prediction. Comparisons with advanced machine learning models were conducted to validate the robustness of the proposed approach and highlight the trade-offs between accuracy and interpretability. By focusing on actionable insights and exploring hypothetical compositions, this work contributes to advancing

the understanding of wine quality and its determinants, rather than serving as a general tutorial on data processing methods.

9.1. Analysis of Ridge Regression for Predicting Wine Quality

The ridge regression model applied to our wine quality dataset yielded the results shown in Figures 17 and 18, which present a residuals vs. predicted values plot and a Q-Q plot of the residuals, respectively. This subsection outlines the strengths and limitations of the model, based on key performance metrics and visual diagnostic tests.

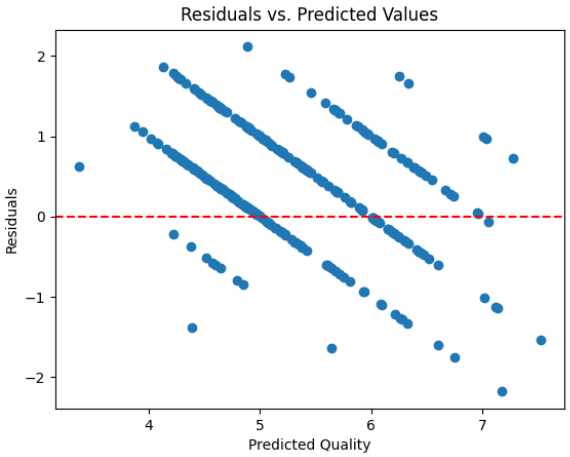


Figure 17. Residuals vs. predicted values.

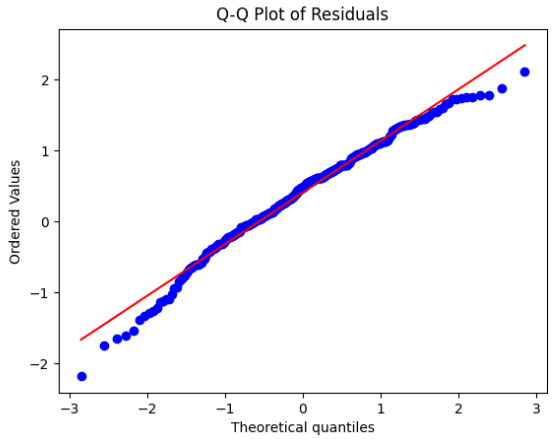


Figure 18. Q-Q plot of residuals.

9.1.1. Model Performance Metrics

The performance metrics, as provided in the statistical analysis, are as follows:

- R^2 on Training Set: -0.1397
- R^2 on Test Set: -0.0636
- Adjusted R^2 on Training Set: -0.1460
- Cross-Validation MSE: 0.7415
- Predicted quality after 5% perturbation: 10.3362
- 95% Confidence Interval for predicted quality: (4.1488, 6.9951)

The R^2 values indicate how much of the variance in wine quality the model explains. The negative R^2 values on both the training and test sets suggest that the model is currently underperforming. However, cross-validation results show an acceptable mean squared error (MSE) of 0.7415, indicating that the model's predictive power could be enhanced with further refinement.

Furthermore, the perturbation analysis demonstrates that the model can respond to small variations in input features. A 5% increase in the features resulted in a predicted quality score of 10.3362, showing that the model is capable of reflecting changes in the input data.

9.1.2. Residual Analysis

Residual analysis is crucial for assessing the model's accuracy. The residuals are defined as

$$e_i = y_i - \hat{y}_i \tag{49}$$

where e_i is the residual, y_i is the observed wine quality, and $\hat{y}_i$ is the predicted quality. Figure 17 illustrates the relationship between the predicted values and residuals. The plot shows a relatively even distribution of residuals around the zero line, which suggests that the model does not systematically overpredict or underpredict across the range of predicted wine quality scores. This even distribution is key to avoiding model bias.

Figure 18 presents the Q-Q plot of residuals, which compares the quantiles of the residuals with the quantiles of a standard normal distribution. The majority of residuals align with the theoretical quantile line, indicating that the residuals are approximately normally distributed. This supports the validity of the ridge regression model, as it fulfills the assumption of normally distributed residuals.

9.1.3. Predictive Confidence Intervals

We calculated the 95% confidence interval for the predicted quality scores. Confidence intervals provide a range of values within which the true prediction is likely to fall, and they are computed as follows:

$$CI = \hat{y} \pm Z_{\alpha/2} \cdot \sigma_{\hat{y}} \tag{50}$$

where $Z_{\alpha/2}$ is the critical value from the standard normal distribution and $\sigma_{\hat{y}}$ is the standard error of the prediction. The calculated 95% confidence interval for predicted wine quality is between 4.1488 and 6.9951. This range suggests that the model can predict moderate-quality wines with confidence, though it struggles to accurately predict very high- or very low-quality wines.

In conclusion, although the ridge regression model currently shows some limitations in terms of explaining the full variability in wine quality (as indicated by the negative R^2 values), its residual analysis, cross-validation results, and predictive confidence intervals demonstrate that it can still provide valuable predictions. The model is responsive to small changes in input features and maintains an approximately normal distribution of residuals. With further refinement and feature selection, its predictive performance can be improved.

9.2. Comparison with Related Studies and Suggestions for Improvement

The achieved accuracy of 0.6031 for the optimized second-degree polynomial ridge regression model demonstrates a balance between predictive performance and interpretability. Ridge regression was selected for its simplicity and ability to reveal the relationships between physicochemical properties and wine quality. This interpretability is particularly valuable in winemaking, where understanding feature contributions is essential for practical applications.

The choice of ridge regression aligns with the study's objectives of balancing performance with interpretability and extending predictions beyond the observed data range. Although models such as Neural Network Classifier, Random Forest Classifier, and XGBoost achieved higher accuracies within the dataset's range, their predictions are often limited to the patterns present in the training data. In contrast, ridge regression, with its

assumptions of linearity and regularization, provides a framework that balances predictive accuracy with extrapolation capabilities.

Moreover, ridge regression offers direct insights into feature contributions, which are crucial for applications such as optimizing wine quality based on individual physicochemical properties. For this study's goal of identifying potential wine compositions that achieve high-quality scores, ridge regression provides a practical and interpretable approach. Future research could enhance this framework by integrating ridge regression with advanced models, enabling the cross-validation of predictions and improving the robustness of extrapolated results.

There are limitations of relying solely on a single dataset, as it restricts the diversity of samples and may limit the model's generalizability. Future studies could benefit from the following:

- Cross-Validation with Diverse Datasets: Incorporating datasets from different wine-producing regions or vintages could help evaluate the robustness and applicability of the model across varied conditions. Such datasets could capture unique patterns in physicochemical properties and quality scores, thereby enhancing the model's predictive power.
- Data Augmentation Techniques: Synthetic generation of feature combinations could expand the dataset's diversity, enabling the exploration of a broader range of physicochemical properties. For instance, Gaussian noise addition or feature interpolation methods could simulate new data points while maintaining the relationships observed in the original dataset.

By integrating diverse datasets and leveraging data augmentation techniques, future work can address the current limitations and further improve the robustness and applicability of predictive models for wine quality and other food quality assessments.

We also acknowledge that advanced machine learning models, such as neural networks and ensemble methods like Random Forests, could potentially achieve higher predictive accuracy. Neural networks, for instance, excel in capturing complex, non-linear relationships and may provide better performance within the observed data range. However, their extrapolation capabilities are limited and require careful training with augmented or synthetic data to predict values beyond the training range. Random Forests, on the other hand, inherently lack the ability to predict out-of-range values due to their reliance on decision-tree splits, which restrict predictions to the observed range of the training data.

Future studies could benefit from benchmarking the ridge regression model against these advanced techniques to evaluate trade-offs in accuracy, complexity, and interpretability. Additionally, research could explore strategies for enhancing the out-of-range prediction capabilities of advanced models through synthetic data augmentation or domain-specific feature engineering. This comprehensive comparison would provide valuable insights into the strengths and limitations of each approach, guiding the selection of models based on specific research and practical objectives.

For the present study, ridge regression aligns well with our emphasis on understanding physicochemical contributions to wine quality and exploring hypothetical compositions for optimization. By maintaining a focus on interpretability, this study aims to provide actionable insights for both researchers and practitioners in the wine-making industry.

The achieved accuracy for the optimized second-degree polynomial ridge regression model was evaluated in the context of similar studies that applied machine learning techniques for wine quality prediction. To the best of our knowledge, there is a lack of studies specifically focusing on continuous mathematical modeling for wine quality prediction. Most existing research employs machine learning models aimed at the classification or categorization of wine quality into discrete classes. In contrast, this study uniquely emphasizes

continuous prediction, aiming to quantify quality scores beyond the observed range of the current dataset. This predictive approach provides novel insights into potential wine compositions and quality metrics that surpass traditional categorizations. The prediction of quality scores exceeding the observed range (e.g., 10) is exploratory and serves to identify theoretical compositions that could inform experimental designs. However, these predictions must be interpreted cautiously, as they extrapolate beyond the dataset's range. Future studies should validate such predictions through experimental and sensory evaluations to ensure their scientific and practical relevance.

While the current model shows promise, there are opportunities to improve its predictive performance in future research. Specifically, incorporating additional objective and measurable variables not currently included in the dataset could provide more comprehensive insights. These variables include

- Volatile Compounds: Measurable using gas chromatography–mass spectrometry (GC-MS), these compounds are key contributors to the aroma profile of wine.
- Phenolic Content: Total phenolic compounds, including tannins, can be quantified using spectrophotometry and are linked to the astringency and structure of the wine.
- Color Intensity and Hue: Quantifiable using spectrophotometric methods, these parameters provide insights into the wine's appearance and aging potential.
- Glycerol Content: Quantifiable using enzymatic assays, glycerol influences the wine's body and mouthfeel.

In addition to incorporating these measurable variables, the following strategies could enhance model performance:

- Expanding the Dataset: Including wines with a broader range of physicochemical properties and quality scores could improve the model's generalizability and robustness. A more diverse dataset may capture underrepresented patterns, thereby enhancing accuracy.
- Validation with Independent Datasets: Testing the model on independent datasets from different wine-producing regions could ensure robustness and applicability across diverse contexts.

By integrating additional objective variables and applying these strategies, future studies can address the current limitations and further enhance the predictive power and applicability of mathematical models for wine quality and other food quality assessments.

9.3. Interpretation of Ridge Regression Coefficients and Their Implications

The best second-degree ridge regression equation derived in this study provides valuable insights into the influence of various physicochemical properties on wine quality. The regression coefficients reveal the significance of individual features and their interactions, highlighting key factors that affect wine composition and quality scores.

9.3.1. Key Contributors to Wine Quality

The regression results indicate that certain physicochemical attributes play a dominant role in determining wine quality:

- Alcohol (+0.8735) is the most significant positive contributor to wine quality. This aligns with enological research, where higher alcohol content is often associated with better sensory attributes and structural balance in wine.
- Residual sugar (+0.1011) has a positive impact, suggesting that wines with moderate residual sugar levels are perceived as higher quality. However, excessive sugar, particularly when combined with high sulfur dioxide levels, can be detrimental.

- Total sulfur dioxide (+0.0710) contributes positively, indicating that controlled levels of sulfur dioxide help preserve wine quality. However, its interaction effects with free sulfur dioxide suggest that excessive levels can be undesirable.
- Free sulfur dioxide (−0.0101) exhibits a negative coefficient, reinforcing the importance of maintaining balanced sulfur dioxide levels to avoid off-flavors and undesirable oxidative effects.
- Sulfates (−0.0168) have a slight negative influence on quality, supporting the need for careful sulfate adjustments to prevent excessive astringency.
- pH (−0.0080) also exhibits a minor negative effect, underscoring the importance of maintaining optimal acidity levels.

9.3.2. Non-Linear Effects and Interaction Terms

The quadratic and interaction terms further highlight the complexity of wine chemistry:

- Fixed acidity ($-0.1538 \cdot \log(\text{fixed acidity} + 1)^2$) follows a non-linear pattern, where moderate levels enhance quality, but excessive acidity negatively affects perception.
- Residual sugar ($+0.1188 \cdot \log(\text{residual sugar} + 1)^2$) demonstrates that slight increases improve quality, but excessive sugar can reduce consumer preference.
- Fixed acidity and residual sugar interaction (+0.1669) indicates that a well-balanced combination of these attributes enhances wine quality.
- Free sulfur dioxide and sulfates interaction (−0.4979) suggests that excessive use of sulfur dioxide and sulfates together has a strong negative effect, reinforcing the importance of precise sulfite management.
- Residual sugar and alcohol interaction (−0.0894) highlights that too much sugar combined with high alcohol content can reduce overall quality.

9.3.3. Practical Implications for Winemaking

These findings provide actionable insights for wine producers:

- Maintaining alcohol levels within optimal ranges is crucial for achieving high-quality wines.
- Controlling residual sugar ensures a balance between sweetness and acidity, preventing undesirable sensory effects.
- Managing sulfur dioxide levels appropriately minimizes oxidation while avoiding excessive sulfite-related defects.
- Considering the interactions between acidity, sulfur dioxide, and other components can enhance consistency and quality control.

By leveraging these insights, winemakers can optimize production processes and improve wine quality predictions through data-driven decision-making. Future research could further refine these findings by incorporating sensory evaluation metrics and testing various winemaking conditions.

9.4. Implications for Predictive Modeling and Winemaking Applications

The statistical transformations and modeling approaches applied in this study provide crucial insights into the relationship between wine composition and quality. The results highlight the effectiveness of transformation techniques in stabilizing variance and improving the normality of key wine features, ultimately enhancing predictive model performance.

One key finding is that applying log transformations to variables such as fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, and total sulfur dioxide significantly improved normality and reduced skewness. These transformations directly impact model accuracy, as demonstrated in prior studies [63,73]. The

improved model predictions enable winemakers to make data-driven decisions in optimizing wine composition for higher quality scores.

Furthermore, the application of polynomial ridge regression provides a more refined understanding of how different chemical components interact to influence wine quality. The results indicate that non-linear relationships play a critical role, particularly for features like residual sugar and sulfites, which exhibit non-Gaussian distributions prior to transformation. These findings suggest that conventional linear modeling may be insufficient for capturing the complexities of wine chemistry, reinforcing the necessity of advanced machine learning techniques in oenology research.

From a practical standpoint, the results can aid winemakers in identifying key variables that require stricter control during the production process. For example, the transformation results indicate that minimizing variations in sulfur dioxide content and balancing acidity levels can lead to more consistent quality scores. Additionally, winemakers can leverage these insights to refine blending strategies, ensuring that optimal feature compositions align with high-quality wines.

Overall, the integration of statistical transformations and advanced modeling contributes to more accurate quality assessments and predictive frameworks for wine production. Future research can extend these methodologies by incorporating sensory evaluation data and real-time monitoring systems, further bridging the gap between data-driven analytics and practical winemaking applications.

9.5. Validity and Practical Application of Ridge Regression in Wine Quality Prediction

The application of ridge regression in this study provides an interpretable and robust approach to predicting wine quality. While the model achieves an accuracy of approximately 0.6, it remains valuable for practical applications in winemaking. Similar predictive models with moderate accuracy have been successfully used in various fields, including food science, healthcare, and engineering, demonstrating their ability to provide actionable insights despite statistical limitations.

9.5.1. Regression Models in Wine and Food Quality Assessment

Regression-based predictive modeling has been widely applied in wine and food quality research. Kasimati et al. investigated multiple regression techniques, including Ordinary Least Squares (OLS) and decision trees, for predicting total soluble solids in wine grapes, demonstrating that such models can achieve comparable levels of accuracy [160]. Similarly, Beh and Farver highlighted the advantages of using regression models to analyze wine characteristics, emphasizing their effectiveness in extracting meaningful patterns from physicochemical data [161].

Multivariate regression models have also proven effective in wine quality assessment. Yin et al. applied partial least squares regression (PLSR) to correlate grape physicochemical indices with wine quality, illustrating the predictive efficiency of these models in real-world applications [162]. Similarly, Aleixandre-Tudó et al. demonstrated the potential of multivariate regression methods in predicting sensory attributes in red wines, reinforcing the value of statistical models for practical winemaking decisions [163]. Furthermore, recent advancements in machine learning have improved the predictive power of regression-based approaches, with studies such as Zeng et al. demonstrating the benefits of ensemble learning for wine quality evaluation [164].

9.5.2. Predictive Models in Practical Applications

Predictive models with moderate statistical performance have been successfully employed in various domains, including food safety, healthcare, and engineering, where their predictions guide important decisions.

In food science, predictive microbiology models forecast microbial growth in food products to ensure quality and safety. Despite their moderate predictive accuracy, these models play a significant role in optimizing food production and extending shelf life [165,166]. Similarly, healthcare applications leverage machine learning models with accuracy levels around 0.6 to predict disease progression and patient outcomes, offering valuable insights for clinical decision-making [167–169]. Engineering applications also utilize predictive models for optimizing design and operational processes, where even models with moderate accuracy provide useful recommendations [170–172].

These examples highlight that predictive models do not need perfect accuracy to generate meaningful insights. Instead, they provide a structured approach to decision-making, allowing domain experts to integrate model-based recommendations with empirical knowledge.

9.5.3. Interpreting Model Performance in Wine Quality Prediction

While the model exhibits negative determination coefficients in certain cases, this does not necessarily render it unsuitable for practical application. Negative R^2 values typically indicate instances where the linear regression terms struggle to fully capture the variance in quality evaluations, particularly due to the complexity of sensory attributes and human perception in winemaking. However, the presence of non-linear interaction terms in the model compensates for these limitations, as demonstrated by the improved feature relationships captured through polynomial ridge regression.

In practice, many quality assessment models function effectively despite moderate accuracy. For example, regression-based models have been used in the food industry to evaluate sensory attributes such as texture and flavor in processed foods, including cheese, sausage, and confectionery products [173]. These studies demonstrate that, even with statistical limitations, predictive models can contribute to process optimization and quality control.

9.5.4. Potential for Experimental Validation

To further validate the practical application of this model, future research can explore experimental testing by evaluating wines that align with the predicted physicochemical properties. This validation approach would involve

- Selecting wine samples with chemical compositions matching the model's optimal predictions.
- Conducting controlled sensory evaluations by trained wine tasters to assess whether predicted quality aligns with perceived quality.
- Comparing experimental results with model outputs to refine predictive accuracy and enhance practical usability.

Such an approach would provide empirical confirmation of the model's effectiveness and bridge the gap between statistical modeling and real-world winemaking applications.

Despite inherent variability in wine quality assessments, ridge regression offers a structured and interpretable approach to predicting key quality factors. The application of regression-based models in food science and winemaking has been well documented, with numerous studies demonstrating their value in optimizing production processes. While model accuracy remains moderate, predictive analytics continues to provide actionable insights for wine producers, guiding quality enhancement strategies. Future work integrating experimental validation with model-based predictions will further strengthen its applicability in winemaking.

Author Contributions: Conceptualization, C.E.Y., J.K. and Y.J.; methodology, C.E.Y., J.K., L.C., C.K. and Y.J.; software, J.K. and C.K.; validation, J.K. and Y.J.; formal analysis, J.K.; investigation, J.K. and Y.J.; resources, L.C.; writing—original draft preparation, C.E.Y. and L.C.; writing—review and editing, C.E.Y., L.C., C.K. and Y.J.; visualization, C.K.; supervision, L.C.; project administration, C.K. All authors have read and agreed to the published version of the manuscript.

Funding: This research was partially funded by the National Science Foundation (NSF), grant number 2321939.

Data Availability Statement: The dataset analyzed in this study is publicly available from the UCI Machine Learning Repository at https://archive.ics.uci.edu/dataset/186/wine+quality, DOI: https://doi.org/10.24432/C56S3T [11].

Acknowledgments: We acknowledge Georgia Southern University for providing resources and support during the course of this research.

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

1. Cemil Emre Yavas, J.K.; Chen, L. Exploring Flavors Through AI: The Future of Culinary Taste Prediction. In Proceedings of the 2024 IEEE/ACIS 22nd International Conference on Software Engineering Research, Management and Applications (SERA), Honolulu, HI, USA, 30 May–1 June 2024; pp. 139–147. [CrossRef]
2. Yu, P.; Low, M.Y.; Zhou, W. Development of a Partial Least Squares-Artificial Neural Network (PLS-ANN) Hybrid Model for the Prediction of Consumer Liking Scores of Ready-to-Drink Green Tea Beverages. *Food Res. Int.* **2018**, *103*, 68–75. [CrossRef] [PubMed]
3. Sudha, A.; Devi, K.S.P.; Sangeetha, V.; Sangeetha, A. Development of Fermented Millet Sprout Milk Beverage Based on Physicochemical Property Studies and Consumer Acceptability Data. *J. Sci. Ind. Res. (JSIR)* **2016**, *75*, 239–243.
4. Dahal, K.R.; Dahal, J.N.; Banjade, H.; Gaire, S. Prediction of Wine Quality Using Machine Learning Algorithms. *Open J. Stat.* **2021**, *11*, 278–289. [CrossRef]
5. Yavas, C.E.; Kim, J.; Chen, L. Mastering Precision in Pivotal Variables Defining Wine Quality via Incremental Analysis of Baseline Accuracy. *IEEE Access* **2024**, *12*, 105429–105459. [CrossRef]
6. Hui-Ye Chiu, T.; Wu, C.; Chen, C.H. A Generalized Wine Quality Prediction Framework by Evolutionary Algorithms. *Int. J. Interact. Multimed. Artif. Intell.* **2021**, *6*, 60. [CrossRef]
7. Gupta, Y. Selection of important features and predicting wine quality using machine learning techniques. *Procedia Comput. Sci.* **2018**, *125*, 305–312. [CrossRef]
8. Bhardwaj, P.; Tiwari, P.; Olejar, K.; Parr, W.; Kulasiri, D. A machine learning application in wine quality prediction. *Mach. Learn. Appl.* **2022**, *8*, 100261. [CrossRef]
9. Luque, A.; Mazzoleni, M.; Zamora-Polo, F.; Ferramosca, A.; Lama, J.R.; Previdi, F. Determining the Importance of Physicochemical Properties in the Perceived Quality of Wines. *IEEE Access* **2023**, *11*, 115430–115449. [CrossRef]
10. Kelly, M.; Longjohn, R.; Nottingham, K. The UCI Machine Learning Repository. 2024. Available online: https://archive.ics.uci.edu (accessed on 1 June 2024).
11. Paulo Cortez, A.C. Wine Quality. 2009. Available online: https://doi.org/10.24432/C56S3T (accessed on 23 February 2025).
12. Hoerl, A.E.; Kennard, R.W. Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics* **1970**, *12*, 55–67. [CrossRef]
13. Rumelhart, D.E.; Hinton, G.E.; Williams, R.J. Learning Representations by Back-Propagating Errors. *Nature* **1986**, *323*, 533–536. [CrossRef]
14. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
15. Pearson, K. LIII. *On Lines and Planes of Closest Fit to Systems of Points in Space*. *Lond. Edinb. Dublin Philos. Mag. J. Sci.* **1901**, *2*, 559–572. [CrossRef]
16. O'Shea, L.E.; Dickens, G.L. Predictive Validity of the Short-Term Assessment of Risk and Treatability (START) for Aggression and Self-Harm in a Secure Mental Health Service: Gender Differences. *Int. J. Forensic Ment. Health* **2015**, *14*, 132–146. [CrossRef]
17. Spurk, D.; Keller, A.C.; Hirschi, A. Do Bad Guys Get Ahead or Fall Behind? Relationships of the Dark Triad of Personality with Objective and Subjective Career Success. *Soc. Psychol. Personal. Sci.* **2016**, *7*, 113–121. [CrossRef]

18. Nogueira, T.F.; Clausen, T.H.; Corbett, A.C. Does Practice Make Perfect? Assessing the Formation of Expertise amongst New Venture Founders. *Int. J. Entrep. Behav. Res.* **2022**, *28*, 1851–1867. [CrossRef]
19. Creaven, A.M.; Howard, S.; Hughes, B.M. Social Support and Trait Personality Are Independently Associated with Resting Cardiovascular Function in Women. *Br. J. Health Psychol.* **2013**, *18*, 556–573. [CrossRef] [PubMed]
20. Bower, J.; Acar, S.; Kursuncu, U. Measuring Creativity in Academic Writing: An Analysis of Essays in Advanced Placement Language and Composition. *J. Adv. Acad.* **2023**, *34*, 183–214. [CrossRef]
21. Akinwande, M.O.; Dikko, H.G.; Samson, A. Variance Inflation Factor: As a Condition for the Inclusion of Suppressor Variable(s) in Regression Analysis. *Open J. Stat.* **2015**, *05*, 754–767. [CrossRef]
22. Tummala-Narra, P.; Sathasivam-Rueckert, N. Perceived Support from Adults, Interactions with Police, and Adolescents' Depressive Symptomology: An Examination of Sex, Race, and Social Class. *J. Adolesc.* **2013**, *36*, 209–219. [CrossRef]
23. Yin, X.; Wei, X.; Irfan, M.; Yasin, S. Revitalizing Organizational Efficiency: Unpacking the Relationship between CEO Turnover, Research and Development, and Pay-Performance Sensitivities in the Financial Sector of Pakistan. *Sustainability* **2023**, *15*, 10578. [CrossRef]
24. Vatcheva, K.P.; Lee, M. Multicollinearity in Regression Analyses Conducted in Epidemiologic Studies. *Epidemiol. Open Access* **2016**, *6*, 227. [CrossRef] [PubMed]
25. Warmenhoven, J.; Bargary, N.; Liebl, D.; Harrison, A.; Robinson, M.A.; Gunning, E.; Hooker, G. PCA of Waveforms and Functional PCA: A Primer for Biomechanics. *J. Biomech.* **2021**, *116*, 110106. [CrossRef] [PubMed]
26. Marukatat, S. Tutorial on PCA and Approximate PCA and Approximate Kernel PCA. *Artif. Intell. Rev.* **2023**, *56*, 5445–5477. [CrossRef]
27. Madhavan, J.; Porkumaran, K. Performance Comparison of PCA, DWT-PCA And LWT-PCA for Face Image Retrieval. *Comput. Sci. Eng. Int. J.* **2012**, *2*, 41–50. [CrossRef]
28. Wijaya, I.G.P.S. Face Recognition Using Holistic Features and Within Class Scatter-Based PCA. *GSTF Int. J. Comput. (JoC)* **2013**, *3*, 2. [CrossRef]
29. KumarBansal, A.; Chawla, P. Performance Evaluation of Face Recognition Using PCA and N-PCA. *Int. J. Comput. Appl.* **2013**, *76*, 14–20. [CrossRef]
30. Mirin, S.N.S.; Abdul Wahab, N. Fault Detection and Monitoring Using Multiscale Principal Component Analysis at a Sewage Treatment Plant. *J. Teknol.* **2014**, *70*, 3469. [CrossRef]
31. Subiyanto, S.; Priliyana, D.; Riyadani, M.E.; Iksan, N.; Wibawanto, H. Face Recognition System with PCA-GA Algorithm for Smart Home Door Security Using Rasberry Pi. *J. Teknol. Dan Sist. Komput.* **2020**, *8*, 210–216. [CrossRef]
32. Mendes, M.; Pala, A. Type I Error Rate and Power of Three Normality Tests. *Inf. Technol. J.* **2003**, *2*, 135–139. [CrossRef]
33. Okwe, E.I.; Olanrewaju, O.I.; Heckman, M.; Chileshe, N. Barriers to Building Information Modelling and Facility Management Practices Integration in Nigeria. *J. Facil. Manag.* **2023**, *21*, 845–865. [CrossRef]
34. Carper, C.; Olguin, S.; Brown, J.; Charlton, C.; Borowczak, M. Challenging Assumptions of Normality in AES S-Box Configurations under Side-Channel Analysis. *J. Cybersecur. Priv.* **2023**, *3*, 844–857. [CrossRef]
35. Guo, J.; Alemayehu, D.; Shao, Y. Tests for Normality Based on Entropy Divergences. *Stat. Biopharm. Res.* **2010**, *2*, 408–418. [CrossRef]
36. Mbah, A.K.; Paothong, A. Shapiro–Francia Test Compared to Other Normality Test Using Expected *p*-Value. *J. Stat. Comput. Simul.* **2015**, *85*, 3002–3016. [CrossRef]
37. Evans, R. Verifying Model Assumptions and Testing Normality. *Vet. Surg.* **2024**, *53*, 17. [CrossRef]
38. Sasmita, N.R.; Khairul, M.; Sofyan, H.; Kruba, R.; Mardalena, S.; Dahlawy, A.; Apriliansyah, F.; Muliadi, M.; Saputra, D.C.E.; Noviandy, T.R.; et al. Statistical Clustering Approach: Mapping Population Indicators Through Probabilistic Analysis in Aceh Province, Indonesia. *Infolitika J. Data Sci.* **2023**, *1*, 63–72. [CrossRef]
39. Coin, D. Testing Normality in the Presence of Outliers. *Stat. Methods Appl.* **2008**, *17*, 3–12. [CrossRef]
40. Karakullukçu, A.; Çakto, P. Analysis of Attitudes of Students Attending Secondary Education Institutions towards Physical Education and Sports According to Various Variables. *Rev. Gestão Secr. (Manag. Adm. Prof. Rev.)* **2023**, *14*, 14532–14544. [CrossRef]
41. Kusuma, H.A.; Oktavia, D.; Nugaraha, S.; Suhendra, T.; Refly, S. Sensor BMP280 Statistical Analysis for Barometric Pressure Acquisition. *IOP Conf. Ser. Earth Environ. Sci.* **2023**, *1148*, 012008. [CrossRef]
42. Ostertagová, E.; Ostertag, O.; Kováč, J. Methodology and Application of the Kruskal-Wallis Test. *Appl. Mech. Mater.* **2014**, *611*, 115–120. [CrossRef]
43. Yang, X.; Yin, X.; Huang, W.; Duan, Z.; Wu, H.; Li, Y.; Li, M.; Zhang, T.; Zhou, C.; Xu, H. Distribution of Cardiorespiratory Fitness in Children and Adolescents at Different Latitudes. *Am. J. Hum. Biol.* **2023**, *35*, e23908. [CrossRef] [PubMed]
44. Nowakowski, M. The ANOVA Method as a Popular Research Tool. *Stud. Pr. WNEiZ* **2019**, *55*, 67–77. [CrossRef]
45. Xie, X.; Xu, Z. Blind Watermark Detection Based on K-S Test in Radio-frequency Signals. *Electron. Lett.* **2020**, *56*, 30–32. [CrossRef]
46. Aslam, M. Introducing Kolmogorov–Smirnov Tests under Uncertainty: An Application to Radioactive Data. *ACS Omega* **2020**, *5*, 914–917. [CrossRef] [PubMed]

47. Sundari, R.; Hasibuan, R.; Abdul Malik, F. Metals Distribution (Cu, Zn, Pb, Mn and Ni) in Campus Wastewater: K-S Test and Friedman ANOVA. *Adv. Mater. Res.* **2013**, *864–867*, 1755–1758. [CrossRef]
48. Biu, E.O.; Nwakuya, M.T.; Wonu, N. Detection of Non-Normality in Data Sets and Comparison between Different Normality Tests. *Asian J. Probab. Stat.* **2020**, *5*, 1–20. [CrossRef]
49. Arsham, H. Adaptive K-S Tests for White Noise in the Frequency Domain. *Int. J. Pure Apllied Math.* **2013**, *82*. [CrossRef]
50. Wee, S.; Choi, C.; Jeong, J. Blind Interleaver Parameters Estimation Using Kolmogorov–Smirnov Test. *Sensors* **2021**, *21*, 3458. [CrossRef]
51. Olusegun, O.J.; Toyin, E.; Aderemi, A.A. Programming Development of Kolmogorov-Smirnov Goodness-of-Fit Testing of Data Normality as a Microsoft Excel® Library Function. *J. Softw. Syst. Dev.* **2015**, *2015*, 1–15. [CrossRef]
52. Afere, B.A.E. On the Fourth-Order Hybrid Beta Polynomial Kernels in Kernel Density Estimation. *J. Niger. Soc. Phys. Sci.* **2024**, *6*, 1631. [CrossRef]
53. Peel, S.; Wilson, L.J. Modeling the Distribution of Precipitation Forecasts from the Canadian Ensemble Prediction System Using Kernel Density Estimation. *Weather Forecast.* **2008**, *23*, 575–595. [CrossRef]
54. Chacón, J.E.; Duong, T. Multivariate Plug-in Bandwidth Selection with Unconstrained Pilot Bandwidth Matrices. *TEST* **2010**, *19*, 375–398. [CrossRef]
55. Zhang, J.; Shi, H.; Dong, Z. Real-Time Remaining Useful Life Prediction Based on Adaptive Kernel Window Width Density. *Meas. Sci. Technol.* **2022**, *33*, 105122. [CrossRef]
56. Erdogmus, D.; Jenssen, R.; Rao, Y.N.; Principe, J.C. Gaussianization: An Efficient Multivariate Density Estimation Technique for Statistical Signal Processing. *J. VLSI Signal Process. Syst. Signal Image Video Technol.* **2006**, *45*, 67–83. [CrossRef]
57. Langrené, N.; Warin, X. Fast and Stable Multivariate Kernel Density Estimation by Fast Sum Updating. *J. Comput. Graph. Stat.* **2019**, *28*, 596–608. [CrossRef]
58. Lopez-Novoa, U.; Mendiburu, A.; Miguel-Alonso, J. Kernel Density Estimation in Accelerators: Implementation and Performance Evaluation. *J. Supercomput.* **2016**, *72*, 545–566. [CrossRef]
59. Duong, T. ks: Kernel Density Estimation and Kernel Discriminant Analysis for Multivariate Data in *R. J. Stat. Softw.* **2007**, *21*. [CrossRef]
60. Shadrokh, A.; Salehi, R. The Bias Reduction in Density Estimation Using a Geometric Extrapolated Kernel Estimator. *Hacet. J. Math. Stat.* **2016**, *47*, 1003–1021. [CrossRef]
61. Hazelton, M.L. Variable Kernel Density Estimation. *Aust. N. Z. J. Stat.* **2003**, *45*, 271–284. [CrossRef]
62. Cai, C.; Poor, H.V.; Chen, Y. Uncertainty Quantification for Nonconvex Tensor Completion: Confidence Intervals, Heteroscedasticity and Optimality. *IEEE Trans. Inf. Theory* **2023**, *69*, 407–452. [CrossRef]
63. Thinh, H.X.; Dua, T.V. Optimal Surface Grinding Regression Model Determination with the SRP Method. *Eng. Technol. Appl. Sci. Res.* **2024**, *14*, 14713–14718. [CrossRef]
64. Thinh, H.X.; Khiem, V.V.; Giang, N.T. Towards enhanced surface roughness modeling in machining: An analysis of data transformation techniques. *EUREKA Phys. Eng.* **2024**, *2*, 149–156. [CrossRef]
65. Lee, J.; Kim, B. Compositional data analysis by the square-root transformation: Application to NBA USG% data. *Commun. Stat. Appl. Methods* **2024**, *31*, 349–363. [CrossRef]
66. Gerke, O.; Möller, S. Modeling Bland–Altman Limits of Agreement with Fractional Polynomials—An Example with the Agatston Score for Coronary Calcification. *Axioms* **2023**, *12*, 884. [CrossRef]
67. Langguth, J.R.; Zadworny, M.; Andraczek, K.; Lo, M.; Tran, N.; Patrick, K.; Mucha, J.; Mueller, K.E.; McCormack, M.L. Gymnosperms demonstrate patterns of fine-root trait coordination consistent with the global root economics space. *J. Ecol.* **2024**, *112*, 1425–1439. [CrossRef]
68. Khanolkar, A.; Pawale, P.; Thorat, V.; Patil, B.; Samanta, G. Near infrared spectroscopy for determination of moisture content in lyophilized formulation. *J. Near Infrared Spectrosc.* **2024**, *32*, 18–28. [CrossRef]
69. Saffari, S.E.; Soo, S.A.; Mohammadi, R.; Ng, K.P.; Greene, W.; Kandiah, N. Modelling the Distribution of Cognitive Outcomes for Early-Stage Neurocognitive Disorders: A Model Comparison Approach. *Biomedicines* **2024**, *12*, 393. [CrossRef] [PubMed]
70. Li, X.; Sun, Y.; Chen, X.; Li, Y.; Jiang, T.; Liang, Z. Saline-Sodic Soil EC Retrieval Based on Box-Cox Transformation and Machine Learning. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2022**, *15*, 1692–1700. [CrossRef]
71. Loaiza, J.G.; Rangel-Peraza, J.G.; Monjardín-Armenta, S.A.; Bustos-Terrones, Y.A.; Bandala, E.R.; Sanhouse-García, A.J.; Rentería-Guevara, S.A. Surface Water Quality Assessment through Remote Sensing Based on the Box–Cox Transformation and Linear Regression. *Water* **2023**, *15*, 2606. [CrossRef]
72. Qi, Z.; Gao, K.; Obracaj, D.; Liu, Y.; Yuan, K. Forecasting the Friction Factor of a Mine Airway Using an Improvement Stacking Ensemble Learning Approach. *IEEE Access* **2024**, *12*, 24813–24830. [CrossRef]
73. Mosquin, P.L.; Aldworth, J.; Chen, W.; Grant, S. Evaluation of streamflow as a covariate in models for predicting daily pesticide concentrations. *JAWRA J. Am. Water Resour. Assoc.* **2023**, *59*, 1459–1476. [CrossRef]

74. Zhang, Z.; Pan, S.; Gao, W.; Wang, C.; Tao, X.; Liu, H. Anomalous Ambiguity Detection Between Reference Stations Based on Box-Cox Transformation of Tropospheric Residual Estimation. *GPS Solut.* **2024**, *28*, 193. [CrossRef]
75. Quevedo-Castro, A.; Monjardín-Armenta, S.A.; Plata-Rocha, W.; Rangel-Peraza, J.G. Implementation of remote sensing algorithms to estimate TOC, Chl-a and TDS in a tropical water body; Sanalona reservoir, Sinaloa, Mexico. *Environ. Monit. Assess.* **2023**, *196*, 175. [CrossRef]
76. Huang, Z.; Zhao, T.; Tian, Y.; Chen, X.; Duan, Q.; Wang, H. Reliability of Ensemble Climatological Forecasts. *Water Resour. Res.* **2023**, *59*, e2023WR034942. [CrossRef]
77. Kuang, X. Research on the prediction method of first-day box office considering the holiday factor. *Appl. Comput. Eng.* **2024**, *50*, 293–306. [CrossRef]
78. Janssen, L.M.; Vries, B.B.L.P.; Janse, M.H.A.; Wall, E.; Elias, S.G.; Salgado, R.; Diest, P.J.; Gilhuijs, K.G.A. Tumor infiltrating lymphocytes and change in tumor load on MRI to assess response and prognosis after neoadjuvant chemotherapy in breast cancer. *Breast Cancer Res. Treat.* **2024**, *209*, 167–175. [CrossRef] [PubMed]
79. Oliveira, T.S.D.; Silva, I.N.; Aniceto, E.S.; Júnior, J.R.M. Application of the multivariate and univariate analyses to estimate the feed efficiency in beef cattle. *Biosci. J.* **2023**, *39*, e39059. [CrossRef]
80. Habibzadeh, F. Data Distribution: Normal or Abnormal? *J. Korean Med. Sci.* **2024**, *39*, e35. [CrossRef] [PubMed]
81. Bosse, N.I.; Abbott, S.; Cori, A.; Van Leeuwen, E.; Bracher, J.; Funk, S. Scoring epidemiological forecasts on transformed scales. *PLoS Comput. Biol.* **2023**, *19*, e1011393. [CrossRef]
82. Schmid, M.; Friede, T.; Klein, N.; Weinhold, L. Accounting for time dependency in meta-analyses of concordance probability estimates. *Res. Synth. Methods* **2023**, *14*, 807–823. [CrossRef]
83. Zhang, R.P.; Zhou, J.H.; Guo, J.; Miao, Y.H.; Zhang, L.L. Inversion models of aboveground grassland biomass in Xinjiang based on multisource data. *Front. Plant Sci.* **2023**, *14*, 1152432. [CrossRef] [PubMed]
84. Yu, G.; Siddiqui, M.K.; Hussain, M.; Hussain, N.; Saddique, Z.; Petros, F.B. On topological indices and entropy measures of beryllonitrene network via logarithmic regression model. *Sci. Rep.* **2024**, *14*, 7187. [CrossRef]
85. Rasheed, M.W.; Mahboob, A.; Hanif, I. Investigating the properties of octane isomers by novel neighborhood product degree-based topological indices. *Front. Phys.* **2024**, *12*, 1369939. [CrossRef]
86. Bueno-López, S.W.; Caraballo-Rojas, L.R.; Torres-Herrera, J.G. Evaluation of Different Modeling Approaches to Estimating Total Bole Volume for Pinus occidentalis, Swartz in Different Ecological Zones. *Preprints* **2024**. [CrossRef]
87. Tu, Z.; Jin, P.; Chen, D.; Chen, Z.; Li, Z.; Zhang, B.; Zhang, L.; Zhang, W.; Bai, X.; Liang, T. Infection evaluation in the early period after liver transplantation: A single-center exploration. *Transpl. Infect. Dis.* **2023**, *25*, e14002. [CrossRef]
88. Pizarro, C.; Bosse, F.L.; Begrich, C.; Reznakova, B.; Beiert, T.; Schrickel, J.W.; Nickenig, G.; Skowasch, D.; Momcilovic, D. Cardiac autonomic dysfunction in adult congenital heart disease. *BMC Cardiovasc. Disord.* **2023**, *23*, 513. [CrossRef]
89. Subi, X.; Eziz, M.; Zhong, Q. Hyperspectral Estimation Model of Organic Matter Content in Farmland Soil in the Arid Zone. *Sustainability* **2023**, *15*, 13719. [CrossRef]
90. Hieu, N.T.D.; Tri, N.Q.; Huan, N.H.; Dien, T.D.; Tran, N.D.H.; Lien, N.P.; Van, T.T. Estimating the Chlorophyll-a in the Nha Trang Bay using Landsat-8 OLI data. *IOP Conf. Ser. Earth Environ. Sci.* **2023**, *1226*, 012010. [CrossRef]
91. Li, W.; Duan, Q.; Wang, Q.J. Factors Influencing the Performance of Regression-Based Statistical Postprocessing Models for Short-Term Precipitation Forecasts. *Weather Forecast.* **2019**, *34*, 2067–2084. [CrossRef]
92. Wright, D.B.; Herrington, J.A. Problematic Standard Errors and Confidence Intervals for Skewness and Kurtosis. *Behav. Res. Methods* **2011**, *43*, 8–17. [CrossRef] [PubMed]
93. Hong, J.S. Statistics of 6-Hour Forecast Errors Derived from Global Data Assimilation System at the Central Weather Bureau in Taiwan. *Terr. Atmos. Ocean. Sci.* **2001**, *12*, 635. [CrossRef] [PubMed]
94. Westfall, P.H. Kurtosis as Peakedness, 1905–2014. *R.I.P. Am. Stat.* **2014**, *68*, 191–195. [CrossRef] [PubMed]
95. Blanca, M.J.; Arnau, J.; López-Montiel, D.; Bono, R.; Bendayan, R. Skewness and Kurtosis in Real Data Samples. *Methodology* **2013**, *9*, 78–84. [CrossRef]
96. Fičura, M. Forecasting Cross-Section of Stock Returns with Realised Moments. *Eur. Financ. Account. J.* **2019**, *14*, 71–84. [CrossRef]
97. Vidovic, J. Performance of Risk Measures in Portfolio Construction on Central and South-East European Emerging Markets. *Am. J. Oper. Res.* **2011**, *01*, 236–242. [CrossRef]
98. Silva De Souza, R.; Borges, E.M. Teaching Descriptive Statistics and Hypothesis Tests Measuring Water Density. *J. Chem. Educ.* **2023**, *100*, 4438–4448. [CrossRef]
99. Hwu, T.J.; Han, C.P.; Rogers, K. The Combination Test for Multivariate Normality. *J. Stat. Comput. Simul.* **2002**, *72*, 379–390. [CrossRef]
100. Alam, M.J.B.; Aggarwal, M. Probability Distribution of Soil Suction of Engineered Turf Cover and Compacted Clay Cover. *E3S Web Conf.* **2023**, *382*, 24002. [CrossRef]
101. Galwey, N.W. A Q-Q Plot Aids Interpretation of the False Discovery Rate. *Biom. J.* **2023**, *65*, 2100309. [CrossRef] [PubMed]
102. Loy, A.; Follett, L.; Hofmann, H. Variations of *Q–Q* Plots: The Power of Our Eyes! *Am. Stat.* **2016**, *70*, 202–214. [CrossRef]

103. Turner, S.D. Qqman: An R Package for Visualizing GWAS Results Using Q-Q and Manhattan Plots. *J. Open Source Softw.* **2018**, *3*, 731. [CrossRef]
104. Weine, E.; McPeek, M.S.; Abney, M. Application of Equal Local Levels to Improve Q-Q Plot Testing Bands with *R* Package Qqconf. *J. Stat. Softw.* **2023**, *106*, 1–31. [CrossRef]
105. Cox, N.J. Stata Tip 47: Quantile–Quantile Plots without Programming. *Stata J. Promot. Commun. Stat. Stata* **2007**, *7*, 275–279. [CrossRef]
106. Valeinis, J.; Cers, E.; Cielens, J. Two-Sample Problems in Statistical Data Modelling. *Math. Model. Anal.* **2010**, *15*, 137–151. [CrossRef]
107. Deng, W.Q.; Asma, S.; Paré, G. Meta-Analysis of SNPs Involved in Variance Heterogeneity Using Levene's Test for Equal Variances. *Eur. J. Hum. Genet.* **2014**, *22*, 427–430. [CrossRef]
108. Gastwirth, J.L.; Gel, Y.R.; Miao, W. The Impact of Levene's Test of Equality of Variances on Statistical Theory and Practice. *Stat. Sci.* **2009**, *24*, 343–360. [CrossRef]
109. Abdslam, A.; Farag, M.; Mahfouz Omer, S. "Evaluation of the Effect of Two Different Concentration of Arginine on Fluoride Uptake by Demineralized Enamel Surfaces" in Vitro Study. *Dent. Sci. Updat.* **2022**, *3*, 199–208. [CrossRef]
110. Aldababseh, A.; Temimi, M. Analysis of the Long-Term Variability of Poor Visibility Events in the UAE and the Link with Climate Dynamics. *Atmosphere* **2017**, *8*, 242. [CrossRef]
111. Lu, T.; Forgetta, V.; Richards, J.B.; Greenwood, C.M.T. Genetic Determinants of Polygenic Prediction Accuracy within a Population. *Genetics* **2022**, *222*, iyac158. [CrossRef]
112. Tukiyo, T.; Sumekto, D.R.; Suryani, Y.E. Identifying Students' Religiosity and Character Strengths in a Multiculturalism Life Consequence. *AL-ISHLAH J. Pendidik.* **2023**, *15*, 2013–2024. [CrossRef]
113. Ul Islam, T. Preliminary Tests of Homogeneity-Type I Error Rates under Non-Normality. *Biostat. Biom. Open Access J.* **2018**, *6*. [CrossRef]
114. Kim, Y.J.; Cribbie, R.A. ANOVA and the Variance Homogeneity Assumption: Exploring a Better Gatekeeper. *Br. J. Math. Stat. Psychol.* **2018**, *71*, 1–12. [CrossRef]
115. El-Melegy, M.T.; Kamal, A.T. Linear Regression Classification in the Quaternion and Reduced Biquaternion Domains. *IEEE Signal Process. Lett.* **2022**, *29*, 469–473. [CrossRef]
116. Dombrowsky, T. Linear regression: A beginner's guide for nursing research. *Nursing* **2023**, *53*, 56–60. [CrossRef] [PubMed]
117. Suryani, I.; Harafani, H. Predicting the Bitcoin Price Using Linear Regression Optimized with Exponential Smoothing. *J. Ris. Inform.* **2021**, *3*, 277–282. [CrossRef]
118. Božić, D.; Runje, B.; Razumić, A. Risk Assessment for Linear Regression Models in Metrology. *Appl. Sci.* **2024**, *14*, 2605. [CrossRef]
119. Rios-Avila, F.; Maroto, M.L. Moving Beyond Linear Regression: Implementing and Interpreting Quantile Regression Models With Fixed Effects. *Sociol. Methods Res.* **2024**, *53*, 639–682. [CrossRef]
120. Hua, K.F.; Jing, B.Y.; Wu, Y.H. The Application of the Insulin to C-Peptide Molar Ratio (ICPR) in Primary Screening for Insulin Antibodies in Type 2 Diabetes Mellitus Patients: A Further Quantitative Study on the Relationship Between ICPR and Insulin Antibodies. *Diabetes Metab. Syndr. Obes.* **2023**, *16*, 1121–1132. [CrossRef]
121. Rady, E.H.A.; El-Sheikh, A.A.; Department of Applied Statistics and Econometrics, Institute of Statistical Studies and Research, Cairo University, Egypt. The Distribution of the Coefficient of determination in Linear Regression Model: A Review. *J. Univ. Shanghai Sci. Technol.* **2021**, *23*, 126–127. [CrossRef]
122. Cao, C. Prediction Of Medical Insurance Cost Through Linear Regression Model. *Highlights Sci. Eng. Technol.* **2023**, *61*, 152–158. [CrossRef]
123. Ruf, J.; Wang, W. Hedging with Linear Regressions and Neural Networks. *J. Bus. Econ. Stat.* **2022**, *40*, 1442–1454. [CrossRef]
124. Anandhi, P.; Nathiya, D.E. Application of linear regression with their advantages, disadvantages, assumption and limitations. *Int. J. Stat. Appl. Math.* **2023**, *8*, 133–137. [CrossRef]
125. Sugito, N.T.; Abidin, H.Z.; Soemarto, I.; Hendriatiningsih, S. Integration of linear and non-linear regression for estimating land value. *IOP Conf. Ser. Earth Environ. Sci.* **2022**, *1089*, 012032. [CrossRef]
126. Zhou, Y. Stock Forecasting Based on Linear Regression Model and Nonlinear Machine Learning Regression Model. *Adv. Econ. Manag. Political Sci.* **2024**, *57*, 7–13. [CrossRef]
127. Chu, Y.; Yin, Z.; Yu, K. Bayesian scale mixtures of normals linear regression and Bayesian quantile regression with big data and variable selection. *J. Comput. Appl. Math.* **2023**, *428*, 115192. [CrossRef]
128. Satayarak, N.; Benjangkaprasert, C. On the Study of Thai Music Emotion Recognition Based on Western Music Model. *J. Phys. Conf. Ser.* **2022**, *2261*, 012018. [CrossRef]
129. Oktian, S.H.; Adinugroho, W.C.; Anen, N.; Setyaningsih, L. Aboveground Carbon Stock Estimation Model Using Sentinel-2A Imagery in Mbeliling Lanscape in Nusa Tenggara Timur, Indonesia. *KnE Life Sci.* **2022**, *2022*, 368–381. [CrossRef]

130. Sunori, S.K.; Kant, S.; Agarwal, P.; Juneja, P. Development of Rainfall Prediction Models using Linear and Non-linear Regression Techniques. In Proceedings of the 2023 4th IEEE Global Conference for Advancement in Technology (GCAT), Bangalore, India, 6–8 October 2023; pp. 1–5. [CrossRef]
131. Chen, Y.; Wu, C.; Qi, J. Data-driven Power Flow Method Based on Exact Linear Regression Equations. *J. Mod. Power Syst. Clean Energy* **2022**, *10*, 800–804. [CrossRef]
132. Kaur, J.; Goyal, A.; Handa, P.; Goel, N. Solar power forecasting using ordinary least square based regression algorithms. In Proceedings of the 2022 IEEE Delhi Section Conference (DELCON), New Delhi, India, 11–13 February 2022; pp. 1–6. [CrossRef]
133. Faye, C.M.; Fonn, S.; Levin, J. Factors associated with recovery from stunting among under-five children in two Nairobi informal settlements. *PLoS ONE* **2019**, *14*, e0215488. [CrossRef]
134. Jha, S.K.; Patil, V.C.; Rekha, B.U.; Virnodkar, S.S.; Bartalev, S.A.; Plotnikov, D.; Elkina, E.; Patel, N. Sugarcane Yield Prediction Using Vegetation Indices in Northern Karnataka, India. *Univers. J. Agric. Res.* **2022**, *10*, 699–721. [CrossRef]
135. Alam, F.; Usman, M.; Alkhammash, H.I.; Wajid, M. Improved Direction-of-Arrival Estimation of an Acoustic Source Using Support Vector Regression and Signal Correlation. *Sensors* **2021**, *21*, 2692. [CrossRef] [PubMed]
136. Surya, K.; Rajam, V.M.A. Novel Approaches for Resource Management Across Edge Servers. *Int. J. Networked Distrib. Comput.* **2023**, *11*, 20–30. [CrossRef]
137. Zawawi, I.S.M.; Haniffah, M.R.M.; Aris, H. Trend Analysis on Water Quality Index Using the Least Squares Regression Models. *Environ. Ecol. Res.* **2022**, *10*, 561–571. [CrossRef]
138. Güiza-Castillo, L.L.; Pinzón-Sandoval, E.H.; Serrano-Reyes, P.A.; Cely-Reyes, G.E.; Serrano-Agudelo, P.C. Estimation and correlation of chlorophyll and nitrogen contents in *Psidium guajava* L. with destructive and non-destructive methods. *Rev. Colomb. Cienc. Hortícolas* **2020**, *14*, 26–31. [CrossRef]
139. Capparelli, S.; Fra, A.D.; Vietri, A. Searching for Hyperbolic Polynomials with Span Less than 4. *Exp. Math.* **2022**, *31*, 830–842. [CrossRef]
140. Sheremeta, M.; Trukhan, Y. Close-to-convexity of polynomial solutions of a differential equation of the second order with polynomial coefficients of the second degree. *Visnyk Lviv. Universytetu Seriya Mekhaniko-Mat.* **2020**, *90*, 92–104. [CrossRef]
141. Li, Z.; Qiao, B.; Wen, B.; Wang, Y.; Chen, X. Acoustic Mode Measuring Approach Developed on Generalized Minimax-Concave Regularization and Tikhonov Regularization. *IEEE Trans. Instrum. Meas.* **2022**, *71*, 6500411. [CrossRef]
142. Nur, A.R.; Jaya, A.K.; Siswanto, S. Comparative Analysis of Ridge, LASSO, and Elastic Net Regularization Approaches in Handling Multicollinearity for Infant Mortality Data in South Sulawesi. *J. Mat. Stat. Dan Komputasi* **2023**, *20*, 311–319. [CrossRef]
143. Emura, T.; Matsumoto, K.; Uozumi, R.; Michimae, H. g.ridge: An R Package for Generalized Ridge Regression for Sparse and High-Dimensional Linear Models. *Symmetry* **2024**, *16*, 223. [CrossRef]
144. Huang, S.; Wu, Y.; Wang, Q.; Liu, J.; Han, Q.; Wang, J. Estimation of chlorophyll content in pepper leaves using spectral transmittance red-edge parameters. *Int. J. Agric. Biol. Eng.* **2022**, *15*, 85–90. [CrossRef]
145. Hoerl, R.W. Ridge Regression: A Historical Context. *Technometrics* **2020**, *62*, 420–425. [CrossRef]
146. Yu, Y.; Yang, C.; Jiang, G. Ridge regression energy levels calculation of neutral ytterbium (Z = 70). *Chin. Phys. B* **2023**, *32*, 033101. [CrossRef]
147. Safi, S.K.; Alsheryani, M.; Alrashdi, M.; Suleiman, R.; Awwad, D.; Abdalla, Z.N. Optimizing Linear Regression Models with Lasso and Ridge Regression: A Study on UAE Financial Behavior during COVID-19. *Migr. Lett.* **2023**, *20*, 139–153. [CrossRef]
148. Lim, H.Y.; Fam, P.S.; Javaid, A.; Ali, M.K.M. Ridge Regression as Efficient Model Selection and Forecasting of Fish Drying Using V-Groove Hybrid Solar Drier. *Pertanika J. Sci. Technol.* **2020**, *28*, 4. [CrossRef]
149. Arashi, M.; Roozbeh, M.; Hamzah, N.A.; Gasparini, M. Ridge regression and its applications in genetic studies. *PLoS ONE* **2021**, *16*, e0245376. [CrossRef] [PubMed]
150. Varathan, N. An improved ridge type estimator for logistic regression. *Stat. Transit. New Ser.* **2022**, *23*, 113–126. [CrossRef]
151. Kumar, B.; Sharma, M.; Bhat, A.; Kumar, P. An analysis of Indian agricultural workers: A ridge regression approach. *Agric. Econ. Res. Rev.* **2021**, *34*, 121–127. [CrossRef]
152. Yuan, M. Comment: From Ridge Regression to Methods of Regularization. *Technometrics* **2020**, *62*, 447–450. [CrossRef]
153. Irandoukht, A. Optimum Ridge Regression Parameter Using R-Squared of Prediction as a Criterion for Regression Analysis. *J. Stat. Theory Appl.* **2021**, *20*, 242. [CrossRef]
154. Mustafa, S.; Amin, M.; Akram, M.N.; Afzal, N. On the performance of link functions in the beta ridge regression model: Simulation and application. *Concurr. Comput. Pract. Exp.* **2022**, *34*, e7005. [CrossRef]
155. Huo, W.; Tuo, X.; Zhang, Y.; Zhang, Y.; Huang, Y. Balanced Tikhonov and Total Variation Deconvolution Approach for Radar Forward-Looking Super-Resolution Imaging. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 1–5. [CrossRef]
156. Saqib, M. Forecasting COVID-19 outbreak progression using hybrid polynomial-Bayesian ridge regression model. *Appl. Intell.* **2021**, *51*, 2703–2713. [CrossRef] [PubMed]
157. Cheliotis, M.; Lazakis, I.; Theotokatos, G. Machine learning and data-driven fault detection for ship systems operations. *Ocean Eng.* **2020**, *216*, 107968. [CrossRef]

158. Šinkovec, H.; Heinze, G.; Blagus, R.; Geroldinger, A. To tune or not to tune, a case study of ridge logistic regression in small or sparse datasets. *BMC Med. Res. Methodol.* **2021**, *21*, 199. [CrossRef] [PubMed]
159. Abreu, M.C.; De Souza, A.; Lyra, G.B.; Pobocikova, I.; Cecílio, R.A. Analysis of monthly and annual rainfall variability using linear models in the state of Mato Grosso do Sul, Midwest of Brazil. *Int. J. Climatol.* **2021**, *41*, E2445–E2461. [CrossRef]
160. Kasimati, A.; Espejo-Garcia, B.; Vali, E.; Malounas, I.; Fountas, S. Investigating a Selection of Methods for the Prediction of Total Soluble Solids Among Wine Grape Quality Characteristics Using Normalized Difference Vegetation Index Data From Proximal and Remote Sensing. *Front. Plant Sci.* **2021**, *12*, 683078. [CrossRef] [PubMed]
161. Beh, E.J.; Farver, T.B. A Numerical Evaluation of the Classification of Portuguese Red Wine. *Curr. Anal. Chem.* **2012**, *8*, 218–223. [CrossRef]
162. Yin, S.; Zhu, X.; Karimi, H.R. Quality Evaluation Based on Multivariate Statistical Methods. *Math. Probl. Eng.* **2013**, *2013*, 639652. [CrossRef]
163. Aleixandre-Tudó, J.L.; Alvarez, I.; García, M.J.; Lizama, V.; Aleixandre, J.L. Application of Multivariate Regression Methods to Predict Sensory Quality of Red Wines. *Czech J. Food Sci.* **2015**, *33*, 217–227. [CrossRef]
164. Zeng, Q. Prediction of Wine Quality Using Ensemble Learning Approach of Machine Learning. In Proceedings of the 2022 International Conference on Mathematical Statistics and Economic Analysis (MSEA 2022), Dalian, China, 27–29 May 2022; Vilas Bhau, G., Shvets, Y., Mallick, H., Eds.; Atlantis Press International BV: Dordrecht, The Netherlands, 2023; Volume 101, pp. 770–774. [CrossRef]
165. Tarlak, F. The Use of Predictive Microbiology for the Prediction of the Shelf Life of Food Products. *Foods* **2023**, *12*, 4461. [CrossRef] [PubMed]
166. Fujikawa, H. Prediction of Competitive Microbial Growth. *Biocontrol Sci.* **2016**, *21*, 215–223. [CrossRef]
167. Edoh, N.L.; Chigboh, V.M.; Zouo, S.J.C. The Role of Data Analytics in Reducing Healthcare Disparities: A Review of Predictive Models for Health Equity. *Int. J. Manag. Entrep. Res.* **2024**, *6*, 3819–3829. [CrossRef]
168. Yelne, S.; Chaudhary, M.; Dod, K.; Sayyad, A.; Sharma, R. Harnessing the Power of AI: A Comprehensive Review of Its Impact and Challenges in Nursing Science and Healthcare. *Cureus* **2023**, *15*, e49252. [CrossRef] [PubMed]
169. Mansourvar, M.; Wiil, U.K.; Nøhr, C. Big Data Analytics in Healthcare: A Review of Opportunities and Challenges. In *Emerging Technologies in Computing*; Miraz, M.H., Excell, P.S., Ware, A., Soomro, S., Ali, M., Eds.; Springer International Publishing: Cham, Switzerland, 2020; Volume 332, pp. 126–141. [CrossRef]
170. CI&DETS/ESAV, Polytechnic Institute of Viseu, Department of Food Industry, Viseu, Portugal; Guiné, R.P.F. The Use of Artificial Neural Networks (ANN) in Food Process Engineering. *ETP Int. J. Food Eng.* **2019**, *5*, 15–21. [CrossRef]
171. Goyal, S. Artificial Neural Networks (ANNs) in Food Science—A Review. *Int. J. Sci. World* **2013**, *1*, 19–28. [CrossRef]
172. Panerati, J.; Schnellmann, M.A.; Patience, C.; Beltrame, G.; Patience, G.S. Experimental Methods in Chemical Engineering: Artificial Neural Networks–ANNs. *Can. J. Chem. Eng.* **2019**, *97*, 2372–2382. [CrossRef]
173. AI-Driven Automated Feature Engineering to Enhance Performance of Predictive Models in Data Science. *Int. J. Control Autom.* **2020**, *13*, 1558–1571. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.