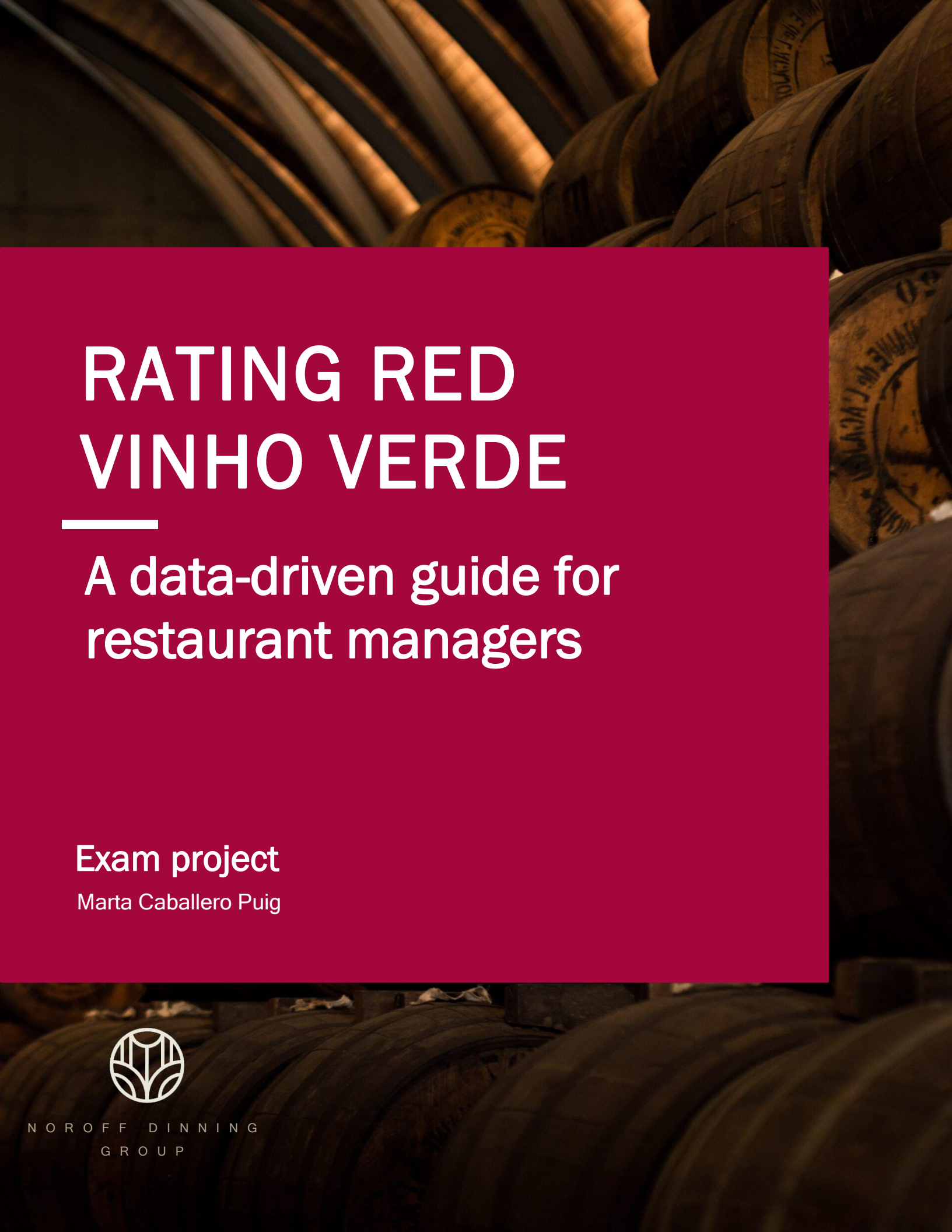

RATING RED VINHO VERDE

A data-driven guide for restaurant managers

Exam project

Marta Caballero Puig

NOROFF DINNING GROUP

TABLE OF CONTENTS

EXECUTIVE SUMMARY

This report analyzes the relationship between different physicochemical properties of red wine samples of Portuguese Vinho Verde and their quality classifications provided through sensory tests. The study uses different techniques and approaches in order to establish which properties and what techniques to use to predict the quality with the smallest space for errors.

Different statistical approaches were explored during this study, including logistic regression, multinomial logistic regression, One-vs-Rest (OvR), segmented regression, and polynomial regression. However, the most accurate model was a multinomial logistic regression approach when looking at binned quality, achieving high accuracy in predicting wine quality, while similar results were attained with a polynomial regression for continuous quality. A detailed comparison of the models is provided in the methodology section.

Key Findings

Significant Chemical Properties:

- Sulphates and Alcohol concentration were identified as the most significant predictors of wine quality, with strong correlations to higher quality classifications.
- Chlorides showed varying but notable effects on quality depending on the classification.
- High-quality wines have distinct chemical characteristics, including optimal levels of alcohol, balanced pH, and lower volatile acidity compared to medium and low quality wines.

Predictive Model Performance:

- The multinomial logistic regression model achieved robust predictive accuracy, with a significant Chi-Square value (154.11, $p<0.05p<0.05$), confirming the model's ability to distinguish between the three quality classes.
- The confusion matrix highlighted a high level of correct predictions for "Medium" and "High" quality wines, with minimal misclassifications.

Class-Specific Insights:

- Low Quality: This category is strongly characterized by lower pH and higher alcohol-sulphate interactions.

- Medium Quality: This category is characterized by balanced levels of Volatile Acidity and moderate Sulphates[2].
- High Quality: This category is characterized by higher pH, elevated Sulphates[2], and well-controlled chlorides.

Conclusions and Implications

The quality of red Vinho Verde wines is strongly tied to their chemical composition, reinforcing the idea that with a proper model and sufficient wine data from the wineries, we can determine the quality of the wine. The predictive model can serve as a reliable tool for restaurant managers to evaluate new wine samples and consistently select higher-quality options for the restaurant chain.

Using the insights from this analysis, Noroff Dining Group's managers can confidently identify and prioritize high-quality wines that meet customer expectations while cutting costs. At the same time, the model provides clear, quantifiable benchmarks that can be used to communicate and negotiate with the suppliers while expecting high-quality standards. In addition, offering wines with scientifically validated quality ensures consistency and enhances the customers' experience, supporting and building the chain's reputation for excellence.

INTRODUCTION

BACKGROUND AND OBJECTIVES

The quality of a wine is influenced by many factors, from the seed, the soil, the weather, the geographic location, the production and aging process, etc., to the chemical composition, which reflects the intricate balance of its components. During winemaking, certain parameters can be considered key to determining the quality of the wine; at the same time, certain benchmarks need to be achieved to pass the wine certification and the quality evaluation.

For restaurant chains aiming to offer the finest wines, understanding and predicting wine quality through scientific methods is essential. This report analyzes red variants of Portuguese Vinho Verde and their physicochemical properties to establish a scientific method for assessing the quality. By using data-driven techniques, this report analyzes key variables such as acidity, sulfates, alcohol concentration, and more, to build robust models capable of distinguishing different levels of wine quality.

Building and choosing the quality prediction model is based on testing, observation and decision-making. The final results are built from a process that is important to mention to understand the final choice. The most noteworthy data mining approaches that were applied are the following:

- Polynomial Regression. Used to capture different non-linear effects of chemical properties observed as correlated and statistically significant. Trial and error were needed to identify subtle, non-linear patterns which led to a highly accurate model able to predict numeric wine quality (1 to 9), but lost accuracy on the highest and lowest ranks.
- Segmented Regression. In order to capture non-linear relationships, and to simplify the quality ranks, the dataset was segmented and binned into high, medium and low as thresholds, and fed to different regressions. This brought attention to changes in variables in each category.
- Logistic Regression. Making use of the binned quality data, this model is used to predict binary outcomes and understand if a wine is of low quality or not, medium quality or not, high quality or not.
- OvR (one-vs-rest): given the multi-class nature of our quality division, this regression dives deeper in the differentiation between the classifications than the logistic regression while

maintaining an easy interpretation. Three separated regression models are trained for each class: low vs medium and high, medium vs high and low, high vs medium and low.

- Multinomial Logistic Regression. After the previous observations and results, this regression can handle a three-classification problem without needing to separate the data into binary models, but instead does it simultaneously. It reduces redundancy and inconsistencies in multiple binary models (like OvR and logistic regression) and can build on light boundaries between quality classes. This proved the most accurate approach to classify wines in categories based on their chemical profiles.

The purpose of this report is to provide actionable insights for Noroff Dining Group management to identify and select high-quality wine, that aligns with customer expectations while saving costs in the process as it skips the need for sensory tests. In the report, the complex analysis will be translated into user-friendly ratings and documentation, empowering an easy decision-making process of the best elements of what the Portuguese Vinho Verde wineries' portfolio can offer.

INITIAL ASSUMPTIONS AND HYPOTHESES

It is a matter of fact that a combination of physicochemical properties and production techniques influences the quality and sensory attributes of wines. This report builds upon already existing research based on the use of data mining and statistical modeling to understand the relationships between certain chemical properties and wine quality. In this study, the focus is on red Vinho Verde wines from Portugal's Minho region, with quality assessed through a combination of objective chemical analyses and subjective sensory evaluations (professional wine tasting).

Assumptions

1. The physicochemical properties of wine significantly influence the perceived wine quality. These properties can be measured and provide insights into sensory characteristics such as aroma, taste, and overall balance.

2. Regression and classification models can effectively predict wine quality grades based on their chemical composition. The relationships between physicochemical variables and overall wine quality are complex but somehow structured, making it possible, through the correct modeling techniques, to make a prediction.

3. Analyzing wine quality in discrete classes (Low, Medium, High) simplifies and improves the understanding of specific chemical thresholds critical for quality differentiation, as

opposed to using detailed grades (1-10) as in the original dataset, which can be more impacted by subjective variability between adjacent grades.

4. An accurate quality prediction model can be built and trained to identify high-quality wine without the need for sensory evaluation.

Hypotheses

1. Chemical properties such as alcohol content, sulphates, and pH have an important role in the classification of wine quality. These properties show nonlinear effects, therefore the use of more advanced modeling techniques like multinomial logistic regression and polynomial regression is needed.

2. Nonlinear regression methods, including polynomial regression and interaction terms, outperform simpler linear models in wine quality prediction.

3. Segmented models (classifying wine quality from grades to Low, Medium, or High) provide better predictions by isolating distinct quality thresholds.

4. Quality classes bring up the importance of specific chemical properties as high or low wine qualities may be characterized by different levels of certain chemical properties.

5. Models based on physicochemical data can mimic or even improve human sensory evaluations, reducing dependence on subjective methods and lowering monetary costs by automating and simplifying the process.

Relevance to the Current Study

The hypotheses are evaluated using a dataset of 1,599 red Vinho Verde wine samples, analyzed through a multinomial logistic regression model. This approach builds on previous studies that highlight the value of data-driven methods in improving quality prediction.

DATA OVERVIEW AND SUMMARY

DATASET DESCRIPTION

The dataset[1] contains 1599 wine observations and thirteen variables, representing twelve feature continuous physicochemical properties of the red Vinho Verde wines, and one target with the numeric (integer) quality classification. The dataset organization is as follows:

Input variables based on physicochemical tests:

1. Fixed acidity
2. Volatile acidity
3. Citric acid
4. Residual sugar
5. Chlorides
6. Free sulfur dioxide
7. Total sulfur dioxide
8. Density
9. pH
10. Sulphates
11. Alcohol

Output variable based on sensory data:

12. Quality (score between 1 and 10)

KEY STATISTICS, DISTRIBUTION AND CHALLENGES

For transparency, these are the most important summary statistics of the variables:

fixed acidity	
Mean	8.31964
Standard Error	0.04354
Median	7.9
Mode	7.2
Standard Deviation	1.74110
Sample Variance	3.03142
Kurtosis	1.13214
Skewness	0.98275
Range	11.3
Minimum	4.6
Maximum	15.9

volatile acidity	
Mean	0.52782
Standard Error	0.00448
Median	0.52
Mode	0.6
Standard Deviation	0.17906
Sample Variance	0.03206
Kurtosis	1.22554
Skewness	0.67159
Range	1.46
Minimum	0.12
Maximum	1.58

citric acid	
Mean	0.27098
Standard Error	0.00487
Median	0.26
Mode	0
Standard Deviation	0.19480
Sample Variance	0.03795
Kurtosis	-0.78900
Skewness	0.31834
Range	1
Minimum	0
Maximum	1

[1] Cortez, P., Cerdeira, A., Almeida, F., Matos, T. and Reis, J. (2009). 'Modeling wine preferences by data mining from physicochemical properties', Decision Support Systems, 47(4), pp. 547-553. Available at: https://doi.org/10.1016/j.dss.2009.05.016 (Accessed: 05 December 2024).

residual sugar	
Mean	2.53881
Standard Error	0.03526
Median	2.2
Mode	2
Standard Deviation	1.40993
Sample Variance	1.98790
Kurtosis	28.61760
Skewness	4.54066
Range	14.6
Minimum	0.9
Maximum	15.5

chlorides	
Mean	0.08747
Standard Error	0.00118
Median	0.079
Mode	0.08
Standard Deviation	0.04707
Sample Variance	0.00222
Kurtosis	41.71579
Skewness	5.68035
Range	0.599
Minimum	0.012
Maximum	0.611

free sulfur dioxide	
Mean	15.87492
Standard Error	0.26159
Median	14
Mode	6
Standard Deviation	10.46016
Sample Variance	109.41488
Kurtosis	2.02356
Skewness	1.25057
Range	71
Minimum	1
Maximum	72

total sulfur dioxide	
Mean	46.46779
Standard Error	0.82264
Median	38
Mode	28
Standard Deviation	32.89532
Sample Variance	1082.10237
Kurtosis	3.80982
Skewness	1.51553
Range	283
Minimum	6
Maximum	289

density	
Mean	0.99675
Standard Error	0.00005
Median	0.997
Mode	0.997
Standard Deviation	0.00189
Sample Variance	0.00000
Kurtosis	0.93408
Skewness	0.07129
Range	0.01362
Minimum	0.99007
Maximum	1.00369

pH	
Mean	3.31111
Standard Error	0.00386
Median	3.31
Mode	3.3
Standard Deviation	0.15439
Sample Variance	0.02384
Kurtosis	0.80694
Skewness	0.19368
Range	1.27
Minimum	2.74
Maximum	4.01

sulphates	
Mean	0.65815
Standard Error	0.00424
Median	0.62
Mode	0.6
Standard Deviation	0.16951
Sample Variance	0.02873
Kurtosis	11.72025
Skewness	2.42867
Range	1.67
Minimum	0.33
Maximum	2

alcohol	
Mean	10.42298
Standard Error	0.02665
Median	10.2
Mode	9.5
Standard Deviation	1.06567
Sample Variance	1.13565
Kurtosis	0.20003
Skewness	0.86083
Range	6.5
Minimum	8.4
Maximum	14.9

quality	
Mean	5.63602
Standard Error	0.02020
Median	6
Mode	5
Standard Deviation	0.80757
Sample Variance	0.65217
Kurtosis	0.29671
Skewness	0.21780
Range	5
Minimum	3
Maximum	8

Table 1. Descriptive statistics of physicochemical variables

No data is missing in any of the observations and, as seasons, climate and location can affect the chemical composition of the samples, and taking into consideration how particular wines can be, no observations were removed as outliers.

Most variables have their means and medians relatively close, suggesting symmetric distributions. However, chlorides (5.68), residual sugar (4.54) and sulphates (2.42) show higher skewness, indicating potential asymmetry. Similarly, the kurtosis for chlorides (41.72) and residual sugar (28.61) is very high, indicating heavy tails.

The ranges for residual sugar (0 to 15.5) and alcohol (8.4 to 14.9) indicate high variability, while density has a much narrower range (0.99 to 1), suggesting limited variability.

What is highly taken into consideration is that the quality grades, which range from 1 to 10, and with a mean score of 5.64, are only represented from 3 to 8 in the dataset. Moreover, the observations are conglomerated in the intermediate range of quality.

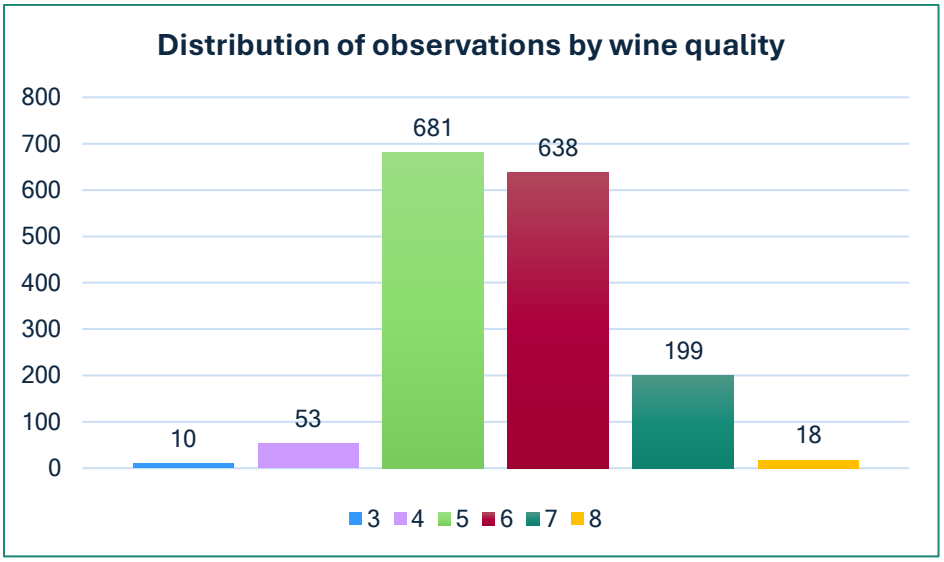

Figure 1. Bar chart distribution of observations by continuous quality

To simplify the categories, ranging the quality from 1 to 10, and having in mind the objective of targeting medium to high quality wines, it was decided to classify the grades as follows:

Quality	Nr. of observations	% of observations
Low (1 to 4)	63	3.93%
Medium (5 to 7)	1518	94.93%
High (8 to 10)	18	1.12%

Table 2. Binned quality classification

The Medium category dominates the dataset, accounting for 94.93% of all observations, while the Low and High categories are underrepresented, with only 3.93% and 1.12% of the observations, respectively. The Low category, as opposed to the other two, is composed of four quality rating categories (1 to 4) due to the model's focus being to prioritize higher ratings while grouping the lower-quality wines together.

The class imbalance in the data distribution poses a challenge for the prediction model, as it may struggle to accurately identify wines in the Low and High quality categories.

To address the class imbalance in the dataset, resampling methods were tested, including oversampling the minority classes (Low and High) and under sampling the majority class (Medium). However, these techniques did not significantly improve the model's accuracy for predicting Low and High quality wines.

EXPLORATION

Strong positive correlations

- Alcohol and Quality (0.476): higher alcohol content is associated with higher wine quality.
- Citric Acid and Fixed Acidity (0.672): closely related variables.

Strong negative correlations

- Density and Alcohol (-0.688): denser wines tend to have lower alcohol content, therefore could lead to lower quality.
- pH and Fixed Acidity (-0.683) : higher acidity is associated with lower pH.

Most variables, especially Residual Sugar, show weak correlations with quality.

	Fixed Acidity	Volatile Acidity	Citric Acid	Residual Sugar	Chlorides	Free Sulfur Dioxide	Total Sulfur Dioxide	Density	pH	Sulphates	Alcohol	Quality
Fixed Acidity	1											
Volatile Acidity	-0.256	1										
Citric Acid	0.672	-0.552	1									
Residual Sugar	0.115	0.002	0.144	1								
Chlorides	0.094	0.061	0.204	0.056	1							
Free Sulfur Dioxide	-0.154	-0.011	-0.061	0.187	0.006	1						
Total Sulfur Dioxide	-0.113	0.076	0.036	0.203	0.047	0.668	1					
Density	0.668	0.022	0.365	0.355	0.201	-0.022	0.071	1				
pH	-0.683	0.235	-0.542	-0.086	-0.265	0.070	-0.066	-0.342	1			
Sulphates	0.183	-0.261	0.313	0.006	0.371	0.052	0.043	0.149	-0.197	1		
Alcohol	-0.062	-0.202	0.110	0.042	-0.221	-0.069	-0.206	-0.496	0.206	0.094	1	
Quality	0.124	-0.391	0.226	0.014	-0.129	-0.051	-0.185	-0.175	-0.058	0.251	0.476	1

Figure 2. Correlation between variables

Strong covariance

- Total Sulfur Dioxide and Free Sulfur Dioxide (109.35): these variables increase together, consistent with the strong positive correlation pointed earlier.
- Total Sulfur Dioxide and Density (1081.43): wines with higher sulfur dioxide levels tend to have higher density, therefore may affect the alcohol content, henceforth the quality.

Strong negative covariance

- Density and Alcohol (-7.72): align with the strong negative correlation and reinforce that denser wines are associated with lower alcohol content.
- Total Sulfur Dioxide and Quality (-4.91): higher sulfur dioxide levels may slightly decrease the perceived wine quality.

Most variables show a weak covariance, near zero, indicating minimal shared variability. Residual Sugar and Quality, like with the weak correlation, show very little covariance. A positive covariance between Alcohol and Quality (0.65) confirms that higher alcoholic content positively contributes to the wine quality.

	Fixed Acidity	Volatile Acidity	Citric Acid	Residual Sugar	Chlorides	Free Sulfur Dioxide	Total Sulfur Dioxide	Density	pH	Sulphates	Alcohol	Quality
Fixed Acidity	3.03											
Volatile Acidity	-0.08	0.03										
Citric Acid	0.23	-0.02	0.04									
Residual Sugar	0.28	0.00	0.04	1.99								
Chlorides	0.01	0.00	0.00	0.00	0.00							
Free Sulfur Dioxide	-2.80	-0.02	-0.12	2.76	0.00	109.35						
Total Sulfur Dioxide	-6.48	0.45	0.23	9.41	0.07	229.59	1081.43					
Density	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00				
pH	-0.18	0.01	-0.02	-0.02	0.00	0.11	-0.34	0.00	0.02			
Sulphates	0.05	-0.01	0.01	0.00	0.00	0.09	0.24	0.00	-0.01	0.03		
Alcohol	-0.11	-0.04	0.02	0.06	-0.01	-0.77	-7.20	0.00	0.03	0.02	1.13	
Quality	0.17	-0.06	0.04	0.02	0.00	-0.43	-4.91	0.00	-0.01	0.03	0.41	0.65

Figure 3. Covariance between variables

Regression

A regression analysis is performed to keep exploring the nuance of the relationships between the physicochemical properties of the wine and its quality. The model is statistically significant, with a high F-statistic and an extremely low p-value. This suggests that the predictor variables explain a substantial portion of the variance in wine quality.

Key Metrics

- Multiple R (0.60): moderate positive correlation between the predictor variables and wine quality.

- R Square (0.36): Approximately 36.1% of the variance in wine quality is explained by the model.
- Adjusted R Square (0.356): About 35.6% of the variance in wine quality is explained after penalizing for any unnecessary complexity, hinting that some predictors might have limited contributions to the model.
- Regression SS (375.75) and Residual SS (666.41): The relatively high Regression (SS) accounts for approximately 36% of the total variation (1042.17) although there is room for improvement.
- Standard Error (0.648): On average, the model's predicted wine quality scores deviate from the actual scores by 0.648 units. It suggests a moderate level of accuracy.

Key Significant Predictors

Volatile acidity, Chlorides, Free Sulfur Dioxide, Total Sulfur Dioxide, pH, Sulphates and Alcohol have a P-value lower than 0.05, hence being statistically significant. On the other hand, The Intercept (Coefficient = 21.97, p = 0.300) and other predictors are not statistically significant

- Alcohol (Coefficient = 0.28, $p < 0.001$): Strongest positive predictor. Indicates that wines with higher alcohol content are rated higher in quality.
- Sulphates (Coefficient = 0.92, $p < 0.001$): Suggests that higher sulphates improve wine quality.
- Volatile Acidity (Coefficient = -1.08, $p < 0.001$): Strong negative effect, indicating that higher volatile acidity is associated with lower wine quality.
- Chlorides (Coefficient = -1.87, $p < 0.001$): Suggests that higher chlorides reduce wine quality.
- Total Sulfur Dioxide (Coefficient = -0.003, $p < 0.001$): Small but significant negative effect indicating that excessive sulfur dioxide lowers the quality.
- pH (Coefficient = -0.41, p = 0.031): Small negative effect, suggests that lower pH (higher acidity), slightly lowers the quality.
- Free Sulfur Dioxide (Coefficient = 0.004, p = 0.045): Minor positive effect.

To further explore the relationships identified in the regression analysis, the following scatterplots illustrate the association between wine quality and the statistically significant predictors. These visualizations help to better understand the direction, strength, and nature of these relationships.

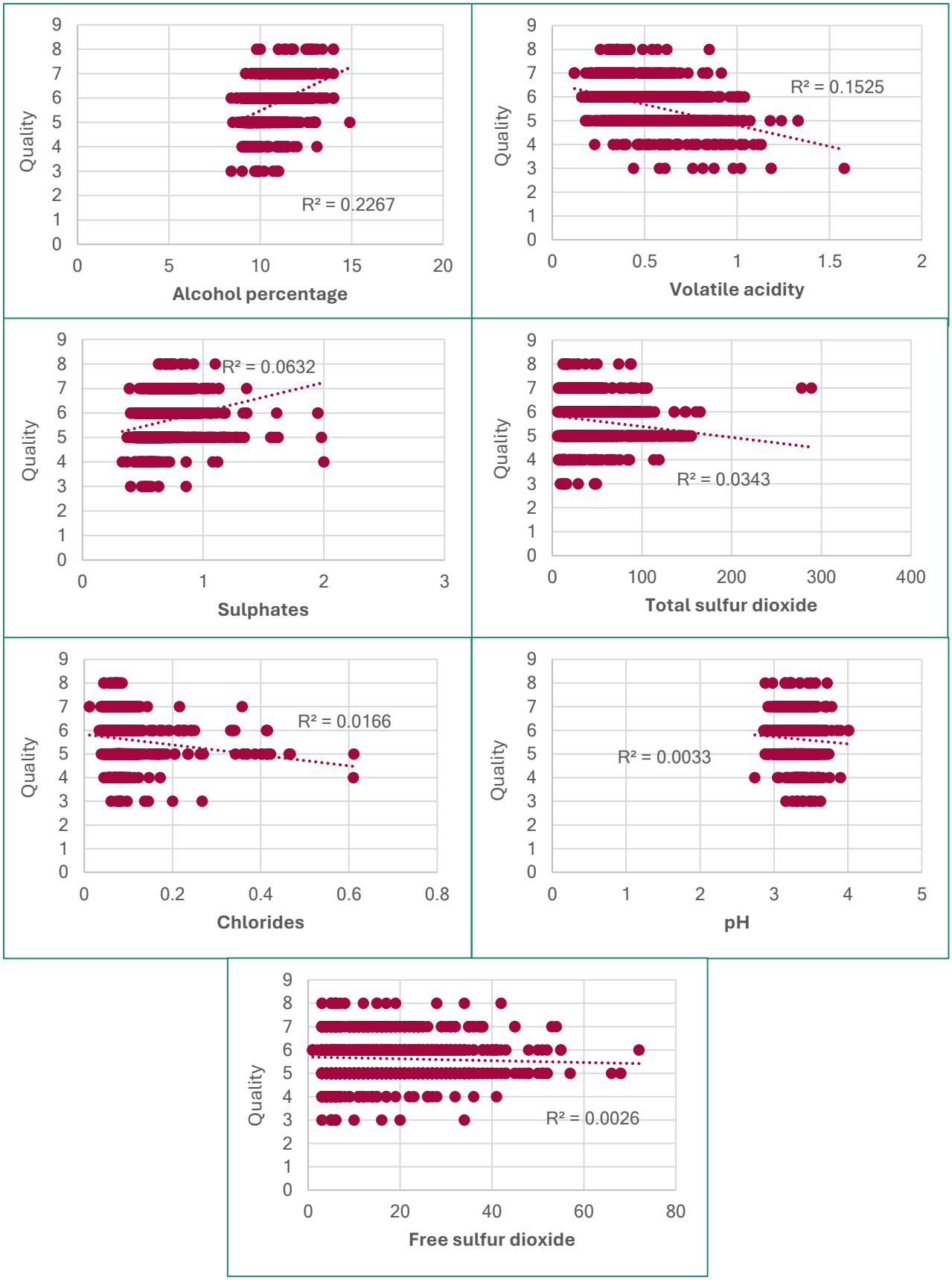

Figure 4. Scatterplots of wine quality and statistically significant variables

As hypothesized, due to the imbalance of the dataset, a basic prediction model solely based on these statistically significant variables is able to capture general trends and relationships in the wine quality, but the performance is limited in predicting extreme quality grades. The scatterplot of the predicted versus actual wine quality values highlights this limitation.

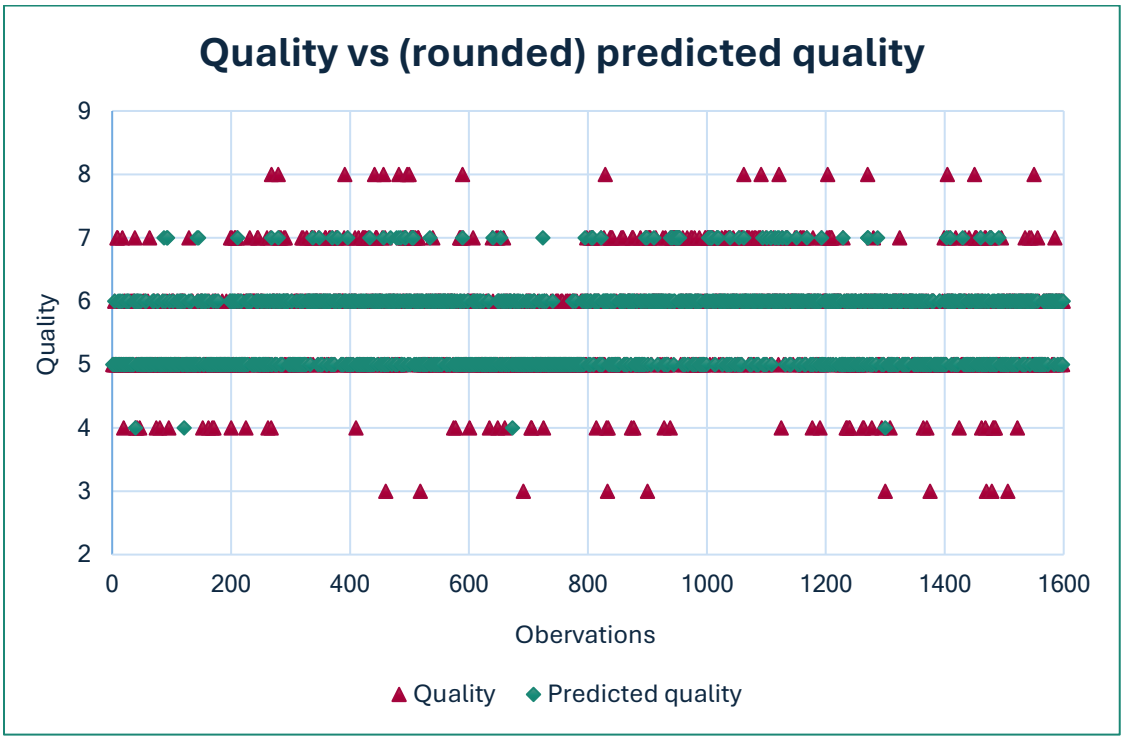

Figure 5. Scatterplot of quality vs quality prediction from regression

This suggests that additional factors or alternative modeling approaches are necessary to improve predictions for these outliers, emphasizing the inherent complexity of wine quality assessment.

MODEL SELECTION AND DEVELOPMENT

Following the Exploratory Data Analysis (EDA), manual investigations using pivot tables and scatterplots were conducted to identify potential patterns and relationships within the dataset. The use of binned quality is introduced (High, Medium, Low). These analyses highlighted the significant influence of alcohol and sulphates on wine quality, particularly when considering non-linear interactions between these variables.

POLYNOMIAL REGRESSION

To address the challenge of class imbalance and capture the nuances in the data, various polynomial regression models were tested with different combinations of chemical properties. Finally, the combination that demonstrated the best performance included the following terms:

- Alcohol × Sulphates
- Sulphates²
- Alcohol²

This model was selected based on its ability to minimize residual errors while maintaining interpretability for practical applications in restaurant management. The following histograms of the residuals are presented below, showing the accuracy of the model in both binned quality (High, Medium and Low) and rounded numeric quality (1 to 10).

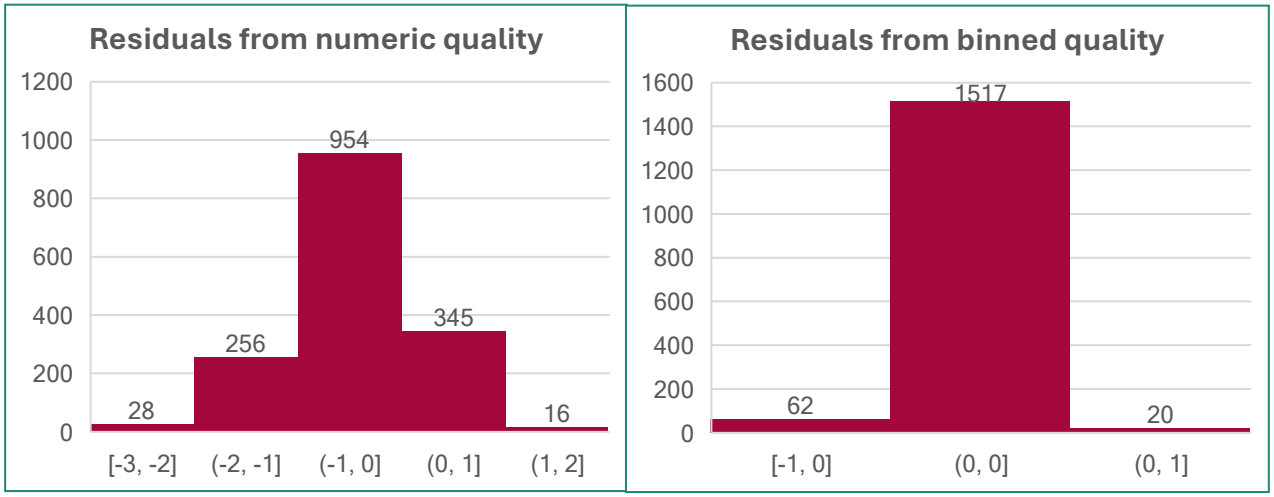

Figure 6. Bar chart of residual results of polynomial logistic regression

Residuals from binned quality: illustrates the residuals when the target variable is categorized into Low, Medium, and High quality. A majority of residuals (94.87% of the predictions) fall within the (0, 0] range, indicating strong performance. The mean absolute error is 0.051, which is indicative of a very strong performance.

Residuals from numeric quality: represent residuals when wine quality is treated as a continuous variable. Most residuals are concentrated near [-1, 0], with fewer extreme deviations, demonstrating the model's effectiveness in capturing overall trends in numeric quality. The mean absolute error is 0.432, which is strong given the range of quality points.

REFINING THE PREDICTION MODEL

After initial experimentation with polynomial regression, several alternative models were explored to improve the accuracy and robustness of wine quality predictions. Building the models from the learning of the polynomial regression proved enriching for the next models. Each of these approaches presented its own strengths and challenges, which helped to decide the best route to take and predictors to be used:

Segmented Regression

The segmented regression was explored to capture potential nonlinear and interaction effects between the physicochemical variables and the wine quality levels. This approach was used to identify thresholds or turning points in relationships that could differentiate the Low, Medium, and High quality categories.

Highlights of the model's insights:

- Alcohol × Sulphates interaction shows a significant positive impact, particularly for Medium-quality wines.
- Quadratic terms for Alcohol and Sulphates reveal nonlinear trends, suggesting that extreme levels of these variables might favor High-quality predictions.
- Consistent negative effects of Volatile Acidity and Free Sulfur Dioxide across the three quality levels, suggesting these variables are generally unfavorable for wine quality.

Coefficient	Low	Medium	High
Intercept	-1.5	-0.2	0.8
Alcohol × Sulphates	0.3	0.6	0.2
Volatile Acidity	-0.5	-0.4	-0.3
Chlorides	0.8	0.7	0.5
Free Sulfur Dioxide	-0.2	-0.1	-0.4
Total Sulfur Dioxide	0.1	0.3	0.5
pH	0.4	0.2	-0.1
$Sulphates^2$	-0.6	-0.3	0.2
$Alcohol^2$	0.7	0.8	0.5

Table 3. Results of segmented regression

Challenges:

Coefficients overlap, especially for the dominant Medium quality, reducing the model's ability to properly distinguish between categories. The model's predictions for the minority classes (Low and High) remain inconsistent due to class imbalance and shared trends among the variables.

Logistic Regression

The logistic regression was tested through Solver as a binary model to predict whether a wine is of High quality or Not High quality. While this approach simplifies the problem, it overlooks the finer distinctions between High, Medium and Low quality wines. This limitation restricts the model's ability to provide further insights.

	Success Observed	Failure Observed	Total
Success Predicted	63	0	63
Failure Predicted	0	1536	1536
Total	63	1536	1599

Table 4. High quality prediction results with logistic regression

Highlights of the model's insights:

- Predicted all 63 High-quality wines correctly.
- Correctly identified all 1536 wines as Not High-quality.
- Achieved 100% accuracy on the training data.
- A classification threshold of 0.5 was used, meaning that probabilities ≥ 0.5 were classified as High quality, and < 0.5 as Not High quality.

Challenges:

- The perfect accuracy raises questions about potential overfitting.
- The class imbalance in the dataset (only 3.93% of wines are High quality) may have influenced the model's performance.

While logistic regression offered a simple baseline, the limitations of a binary model are obvious, and the suspicions of overfitting the data, therefore its inability to handle multi-class classification directly and the dataset's class imbalance led to suboptimal performance, particularly for minority classes. These limitations prompted the exploration of more tailored models, such as One-vs-Rest (OvR) and multinomial logistic regression, for better handling of the task.

One-vs-Rest (OvR)

Due to the challenges that the logistic regression presents, and based on the character of our dataset with three categories to predict, a more suitable model was thought to be OvR. This model can create multiple binary classifiers, each dedicated to distinguishing one class from the rest. Therefore, it allows individual focus on each class, making it easier to handle the imbalance, and it leverages binary classification techniques while enabling multi-class predictions.

Multinomial Logistic Regression

As the next step forward, a multinomial logistic regression was tested, which can handle multi-class problems by modeling the probabilities of each class relative to a reference class simultaneously. This approach directly accounts for the three categories, avoiding the need for binary splits and providing consistent predictions across the entire dataset. Therefore, the categories are divided as follows:

Reference	Quality
2	High
1	Medium
0	Low

Table 5. Quality reference for multinomial logistic regression

Interestingly, both OvR and multinomial logistic regression showed identical residuals and overall performance metrics. This result, while unexpected, can be due to the dataset structure, and given that the predictors have a linear relationship with the target variable and the data is well-separated, or due to the fact that both methods were based on the same maximum likelihood estimation, which can converge to similar outcomes given the same data patterns.

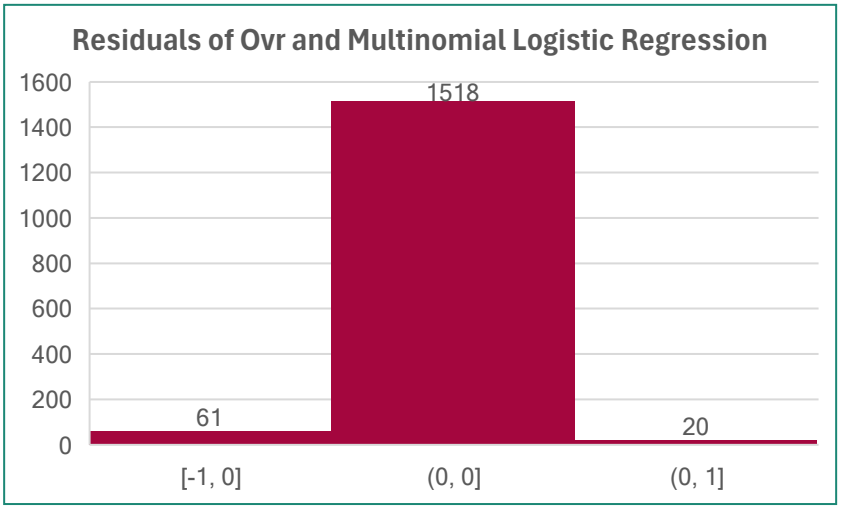

Figure 7. Bar chart of residuals results from OvR and multinomial logistic regression

Highlights of the model's insights:

- The R-squared (Nagelkerke) value is 0.2515, indicating that the model explains 25.15% of the variability in wine quality.
- The model has a Chi-square of 154.11 ($p<0.001p<0.001$), confirming that the predictors significantly improve wine quality classification.
- The Akaike Information Criterion (AIC) is 608.72, indicating a reasonable balance between model complexity and fit.
- While the model fits moderately well, additional predictors or interactions might enhance accuracy.

Key chemical predictors:

We look into the exponentiated coefficient and the p-value to observe the relationship between predictor variables and the odds of the outcome of certain wine quality classifications. The exp(b) value tells us how strongly a chemical property impacts wine quality, while the p-value tells us how confident we are that this relationship is real and not due to random chance.

High Quality (2)

- Alcohol × Sulphates ($exp(b)=9.11$, $p<0.001$): the interaction of alcohol and sulphates has an even stronger positive impact on High quality compared to Medium.
- pH ($exp(b)=0.00048$, $p<0.001$): wines with lower pH values (higher acidity) are highly associated with High quality.
- Sulphates² ($exp(b)=6.06\times10-6$, $p<0.001$): significant nonlinear effect, where very high sulphates levels are negative to High quality.
- Chlorides ($exp(b)=1.99\times10-17$, $p=0.03$): chlorides (saltiness) negatively impact High quality.

Medium Quality (1)

- Alcohol × Sulphates ($exp(b)=2.25$, $p<0.001$): significant positive relationship, meaning that wines with higher alcohol and sulphates interaction contributes to the wine quality.
- Volatile Acidity ($exp(b)=0.0196$, $p<0.001$): strong negative relationship, suggesting that reducing volatile acidity increases the likelihood of Medium quality.
- pH ($exp(b)=0.0403$, $p=0.002$): lower pH (more acidic wines) is associated with higher chances of Medium quality.
- Sulphates² ($exp(b)=0.0163$, $p<0.001$): significant non-linear effect of sulphates on Medium quality.

While this equivalence highlights the consistency of the methods, multinomial logistic regression is more efficient and computationally streamlined for multi-class problems.

- The multinomial logistic regression provides a more direct and interpretable solution for predicting wine quality across all categories, while focusing on targeting Medium to High quality wines.
- OvR can still be a viable alternative, especially in systems where binary classifiers are to be used.

Interpretation guidelines for restaurant managers

- High Alcohol and Sulphates: Wines with a strong interaction between alcohol and sulphates are more likely to be of higher quality. Wines with balanced sulphates and alcohol content should be prioritized when selecting higher-end quality options.
- Lower Volatile Acidity: High-quality wines tend to have lower volatile acidity, contributing to a smoother and less sour taste. This trait can be tailored to the customer's taste.
- Optimal pH: Wines with slightly lower pH (higher acidity) are favored in both Medium and High quality. This trait can be tailored to customer's taste.
- Non-linear effects of Sulphates: While sulphates enhance quality up to a certain point, excessive levels can reduce it. Balance is key.
- Negative impact of Chlorides: Wines with higher chloride content (linked to saltiness) are less likely to achieve High quality, making this an important factor for producers to control.

CONCLUSION & RECOMMENDATIONS

The wine quality prediction model provides restaurant managers with a data-driven tool to optimize wine selection and inventory management. By accurately classifying wine quality into low, medium, or high categories, the model supports better purchasing decisions, ensuring a balance between cost and quality for customer satisfaction. Additionally, insights from the model can help refine wine pairings, enhancing the dining experience and boosting revenue.

The multinomial logistic regression model demonstrated strong reliability for predicting binned wine quality categories, making it particularly effective for practical decision-making scenarios such as selecting Medium to High quality wines for purchase. However, due to the inherent class imbalance in the dataset, the prediction accuracy for minority classes (High and Low quality wines) was slightly less robust compared to the dominant Medium category. On the other hand, the polynomial regression model excelled at predicting continuous wine quality scores, capturing nuanced trends and interactions between physicochemical properties. With a mean absolute error of 0.432, the polynomial regression provides a high level of predictive confidence, ensuring its suitability for tasks requiring precise quality estimation, such as pricing or ranking wines. These findings emphasize the complementary strengths of both models, offering valuable tools for strategic decision-making in wine selection and procurement. As the model further evolves with more data, it has the potential to guide strategic partnerships with suppliers and strengthen the chain quality seal.

The multinomial logistic regression model stands out as the most accurate and suitable choice for predicting wine quality due to its ability to classify wines into distinct quality categories (Low, Medium, and High). The model leverages critical predictors such as alcohol-sulphates interaction, pH, and volatile acidity, which significantly influence wine quality. Its statistically significant results, with a Chi-square value of 154.11 ($p<0.001p<0.001$), and robust predictive metrics, like R-squared (Nagelkerke) of 25.15%, underscore its reliability. By focusing on key chemical properties that affect wine quality, this model empowers restaurant managers to make informed decisions when selecting wines that align with high quality standards and even match certain customer preferences, while delivering consistent quality.

Implementing this model across the restaurant chain provides a data-driven approach to wine purchasing, ensuring managers can objectively assess wine quality before placing orders. This minimizes reliance on subjective evaluations such as sensory tests, reducing cost in the process, lowers purchasing risks, and enhances customer satisfaction by consistently offering high-quality

wines. Moving forward, integrating this model into a user-friendly digital tool or application will enable managers to input chemical property data and instantly receive quality predictions. Additionally, expanding the model by incorporating other relevant variables or validating it on a larger dataset will further enhance its predictive power, ensuring its effectiveness in diverse markets, seasons and wineries, even with the possibility of exploring partnerships with producers. This approach positions Noroff Dinng Group as an industry leader in leveraging data to optimize wine offerings, reducing costs and risks, and overall, elevate the dining experience.

On a personal note

I developed the fictional restaurant chain brand "Noroff Dining Group," envisioning a group of medium to high quality restaurants that offer a fine-dining experience while maintaining the business model of a managed restaurant chain. The branding concept draws inspiration from the “shield-like” logo of Noroff School, which I reimagined with an elegant twist to match the restaurant’s envisioned identity. The design incorporates minimalist lines and a sophisticated color palette: dark red (representing wine), dark green (symbolizing nature and vine leaves), earthy browns (evoking soil and vineyards), turquoise (representing water and adding a touch of optimism), and black (emphasizing elegance and minimalism).

The tone used, the technicality of the wording, charts and numbers presented are meant to match the audience of stakeholders and restaurant managers. These could range from having no technical understanding to being somewhat familiar with some concepts.

Building a predictive model using only Excel has been a challenging experience, and it was tempting to use Python or R –as recommended in the reference study by Cortez et al. (2009). Addressing the quality imbalance in the dataset was particularly difficult and took me on a long learning journey where I explored, read, tried and gained valuable insights through trial and error. However, I recognize that there is room for improvement in my model, such as uncovering additional predictor relationships, working with a richer dataset, and making use of more advanced analytical tools.

BIBLIOGRAPHY

Eunicia, M, Skyszygfrid, R., Vitri, T., Caren, V. (2022). "Modeling Red Wine Quality Based on Physicochemical Tests: A Data Mining Approach", *Formosa Journal of Multidisciplinary Research (FJMR)*. Available at: https://www.researchgate.net/publication/361078760_Modeling_Red_Wine_Quality_Based_on_Physicochemical_Tests_A_Data_Mining_Approach (Accessed: 28 November 2024)

Shalom, N., Asras, T., Gal, E., Demasia, T., Tarab, E., Ezekiel, N., Nikapros, O., Semimufar, O., Gladky, E., Karpenko, M., Sason, D., Maslov, D. and Mor, O. (2022). 'Wine quality and type prediction from physicochemical properties using neural networks for machine learning: A free software for winemakers and customers', *MetaArXiv*. Available at: https://osf.io/preprints/metaarxiv/ph4cu (Accessed: 28 November 2024)

Zaiontz, C. (2023). *Real Statistics Using Excel*. Available at: www.real-statistics.com (Accessed: 20 November 2024)

Visual sources

"Oak wine barrels in cellar" from El Coto (https://elcoto.com/en/types-wine-barrels/), author unknown.

Software

The multinomial logistic regression in this paper was generated with the help of the Real Statistics Resource Pack software (Release 8.9.1). Copyright (2013 – 2023) Charles Zaiontz. www.real-statistics.com.