Analysis of Wine Quality and Prediction Using Logistic Regression

Alix, Paramveer, Susannah, Zoe

2024-12-08

Contents

Summary

This analysis investigates the relationship between physicochemical properties and wine quality using the Wine Quality dataset from the UCI Machine Learning Repository, containing data for both red and white wine. Through comprehensive exploratory data analysis, we examined 11 physicochemical features and their correlations with wine quality scores. Our analysis revealed that higher quality wines typically have higher alcohol content and lower volatile acidity, with white wines generally receiving higher quality scores than red wines. Most features showed right-skewed distributions with notable outliers, particularly in sulfur dioxide and residual sugar measurements. The quality scores themselves followed a normal distribution centered around scores 5-6.

We implemented a logistic regression model with standardized features and one-hot encoded categorical variables, using randomized search cross-validation to optimize the regularization parameter. The final model achieved an accuracy of 54.0% on the test set. While this performance suggests room for improvement, the analysis provides valuable insights for future research directions.

Introduction

The quality of wine is influenced by various chemical properties and sensory factors that determine its taste, aroma, and overall acceptability. Here, we aim to predict the quality of wine using a publicly available wine quality dataset. Machine learning-based predictive modeling is commonly used in the field of wine quality to identify patterns and relationships in key features such as alcohol, sulfates, and volatile acidity, which are critical factors impacting wine quality (Jain 2023). By applying machine learning model, we seek to enhance the accuracy of wine quality predictions and contribute to the advancement of data-driven approaches in wine evaluation methodologies.

Methods

Data

The dataset used in this project is the Wine Quality dataset from the UCI Machine Learning Repository (Cortez 2009) and can be found here. These datasets are related to red and white variants of the Portuguese "Vinho Verde" wine. They contains physicochemical properties (e.g., acidity, sugar content, and alcohol) of different wine samples, alongside a sensory score representing the quality of the wine, rated by experts on a scale from 0 to 10. Each row in the dataset represents a wine sample, with the columns detailing 11 physicochemical attributes and the quality score. The classes are ordered and not balanced (e.g. there are many more normal wines than excellent or poor ones).

Due to privacy and logistic issues, only physicochemical (inputs) and sensory (the output) variables are available (e.g. there is no data about grape types, wine brand, wine selling price, etc.).

1.EDA

1.1 Distribution of quality scores across numerical features

From the distribution plots in Figure 1, we have the following findings:

1. Higher quality wines tend to have higher alcohol content
2. Higher quality wines generally have lower volatile acidity
3. pH seems to have little discrimination power for quality (all quality levels overlap significantly)
4. The `density` feature does not showing any meaningful relationship with wine quality

1.2 Distribution of quality scores by categorical feature (wine color)

Figure 2 simply shows that white wine in average tends to have higher quality scores than red wine.

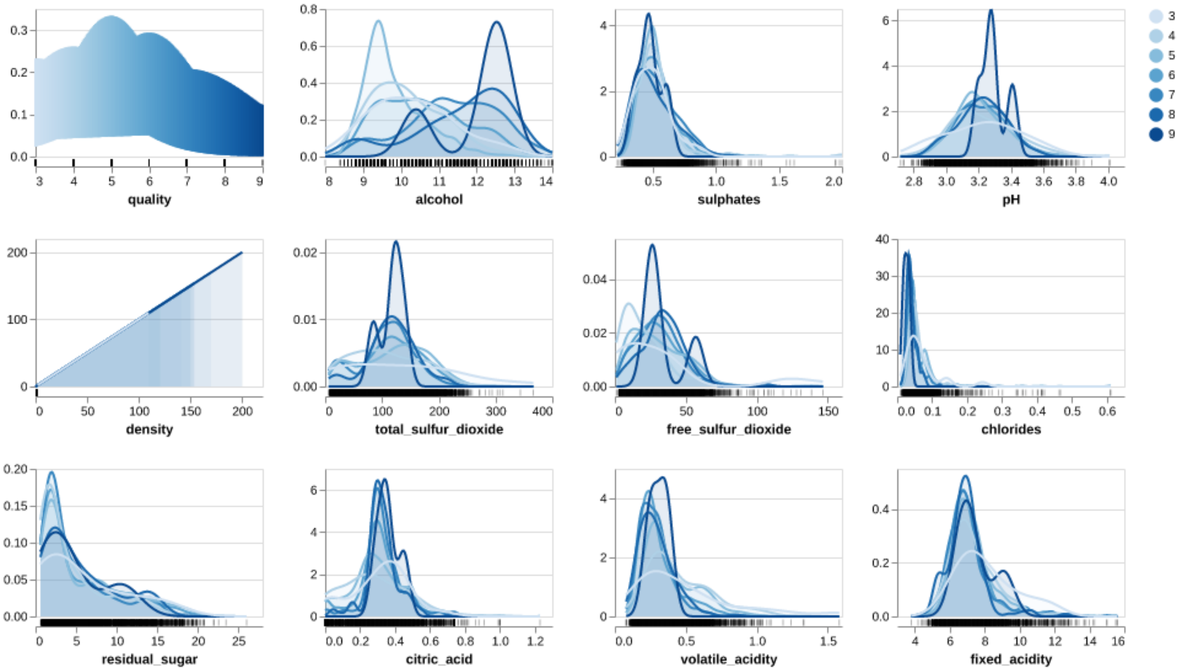

Figure 1: Distribution of wine quality scores by feature.

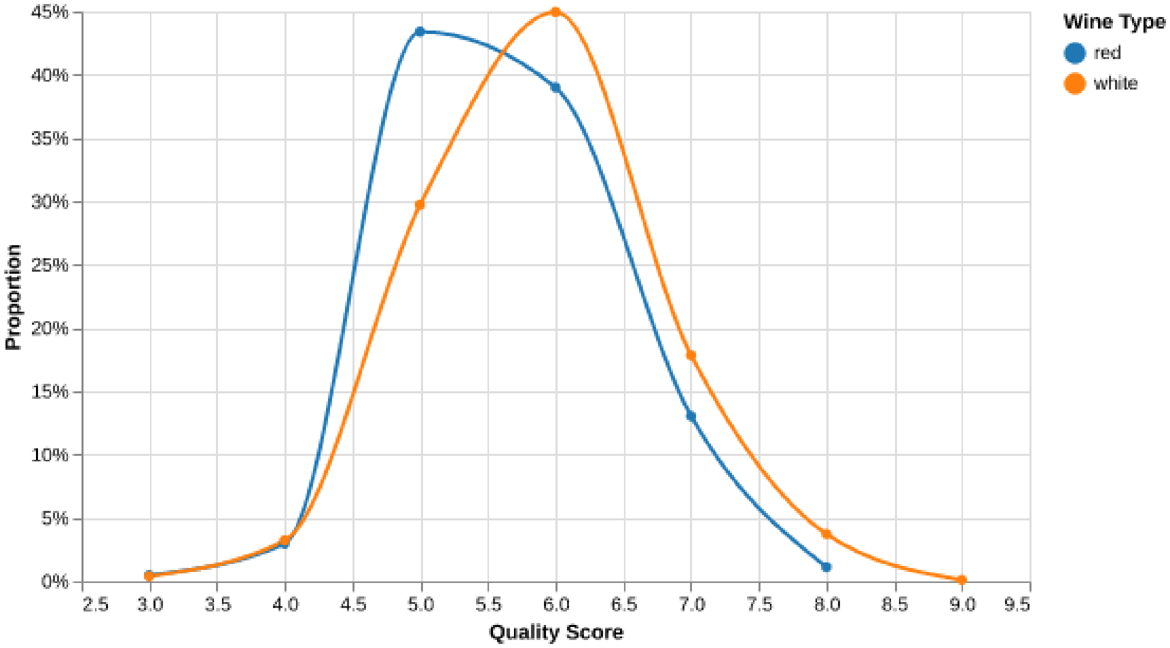

Figure 2: Comparison of red and white wine quality scores.

1.3 Correlation matrix

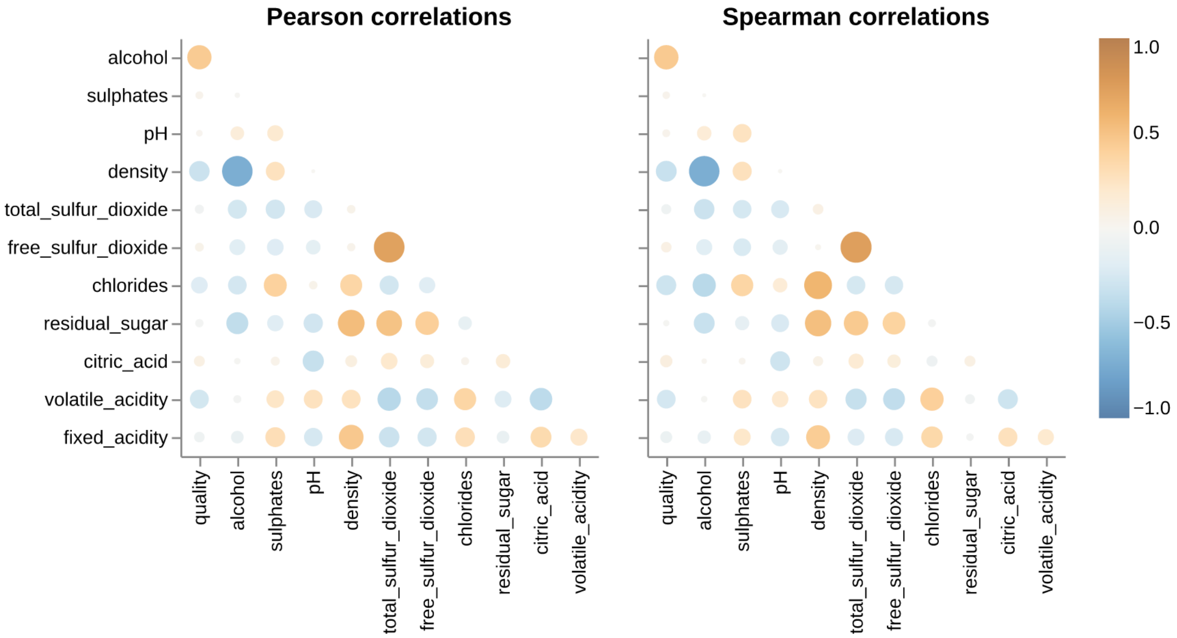

Figure 3: Correlation matrix of all features.

As Figure 3 shows, it seems that the correlation between total sulfur dioxide and free sulfur dioxide is high, we might want to use one of them to represent the other. But let's see the scatter plot for these two features first.

From the scatter plot in Figure 4, we can see that there is a positive linear correlation between between free and total sulfur dioxide, but the relationship is not perfectly linear. Since keeping both features would not make the model too complex, we will leave them both in the model for now.

1.4 Outlier detection

From Figure 5, we have the following findings:

1. Outliers:
   - Many features show significant outliers
   - Particularly noticeable in sulfur dioxide and residual sugar
2. Distributions:
   - Most features show right-skewed distributions
   - pH shows relatively normal distribution for both types

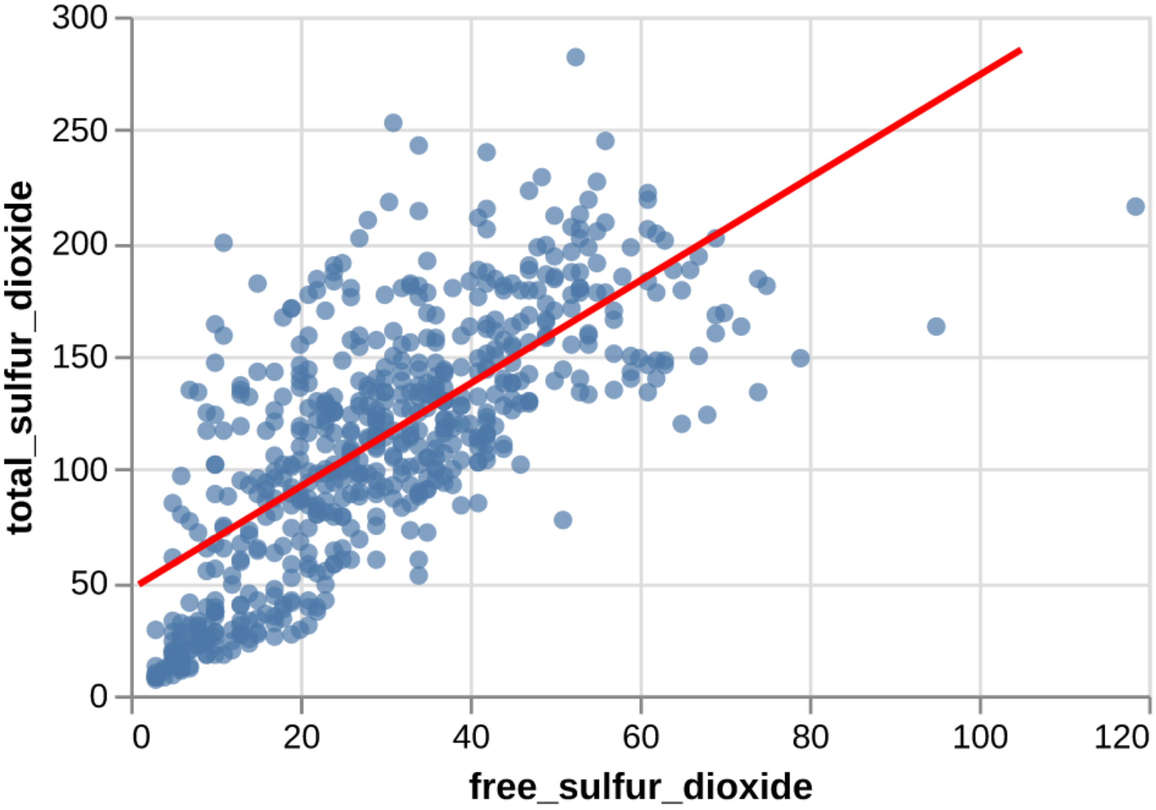

Figure 4: Comparison of levels between total Sulfur Dioxide vs free Sulfur Dioxide.

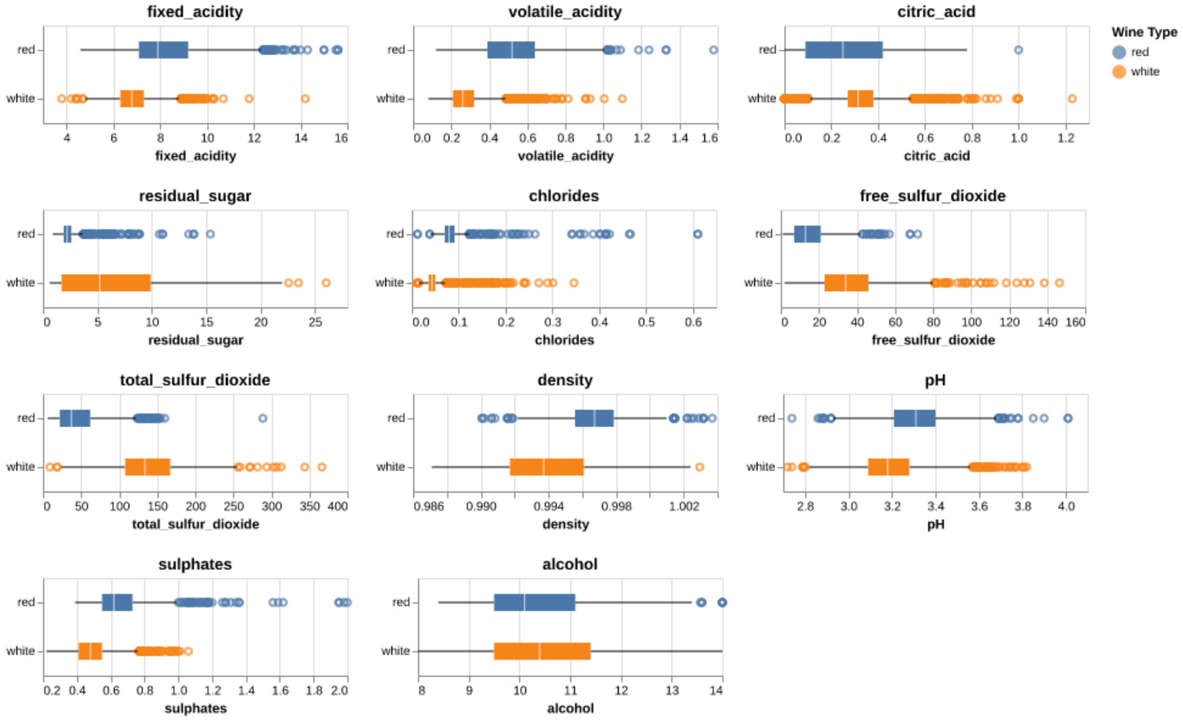

Figure 5: Comparison of levels for all features between red and wine wines.

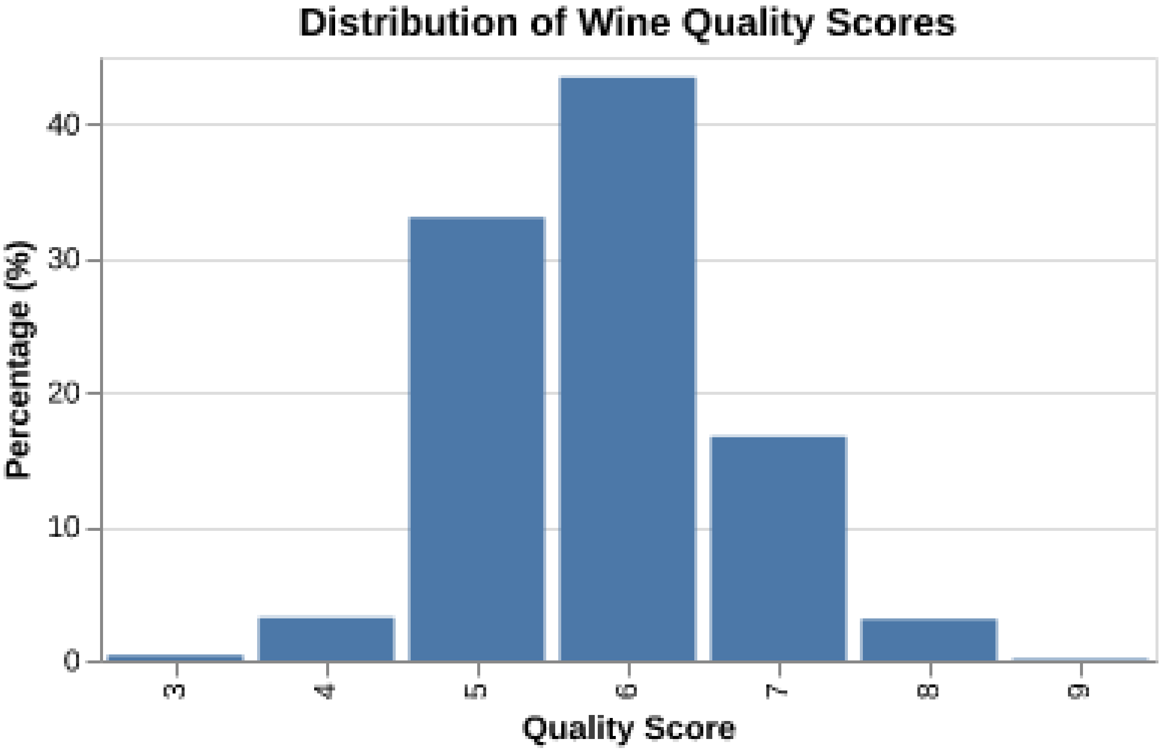
Distribution of Wine Quality Scores
Percentage (%)
0
10
20
30
40
3
4
5
6
7
8
9
Quality Score

Figure 6

1.5 The distribution of the target variable(quality)

We can see from Figure 6 our target variable has a normal distribution. The scores are centered around 5-6, with symmetric decreasing frequencies on both sides, forming a classic bell-shaped curve.

Analysis

The Logistic Regression algorithm was used to build a classification model to predict the quality as an ordinal and numeric integer (found in the `quality` column of the data set). All variables included in the original data set, including wine color (i.e. red or white) were used to fit the model. This is similar to the models suggested by Kniazieva (2023, October 12). Data was split with 80% being partitioned into the training set and 20% being partitioned into the test set. The hyperparameter C was chosen using 3-fold cross validation with the accuracy score as the classification metric. All variables were standardized just prior to model fitting. `color` column is converted to a single binary column with one hot encoding and its `drop='if_binary'` parameter.

Results and Discussion

We split and transform the data (i.e. wine color into binary variable and using standard scalers for all other features) and build our logistic regression model. Using RandomSearchCV, we find the best hyperparamter C for the model: {'logisticregression___C': 95.07243064099163}.

With our tuned model using the best C hyperparameter, we find the accuracy score of our predictions, comparing them to actual wine quality in the test set to be 0.54.

While the performance of this model is not likely very useful in predicting wine quality, as we observed an accuracy score of 0.54, we gained insights on directions that could be further explored. First, we chose logistic regression as it is an intuitive first-step to approach a dataset with largely numeric features representing measurements of contents inside wines. Therefore, further analysis inspecting presence of linear relationships can be conducted using logistic regression results. We can then propose another model, e.g. tree-based ones like Random Forest (Aich 2018), to see whether it does better in wine quality prediction should there be weak linear relationships observed. Second, data cleaning might benefit our decision in choosing an optimal model as outliers have been widely observed across many features, according to our EDA in the previous section. It might be worth it to understand what all features represent and apply human knowledge to modify and "treat" the data so that it is more suitable for training than how it is currently presented. This involves speaking with professionals that understand wine makeup and qualities and seek their insights on reasons of outlier presence and their indications. We believe conducting the above two next-steps will give us a better knowledge foundation in order for us to choose a model that performs better in the future.

References

Aich, Al-Absi, S. 2018. *A Classification Approach with Different Feature Sets to Predict the Quality of Different Types of Wine Using Machine Learning Techniques.* https://doi.org/10.23919/ICACT.2018.8323674.

Cortez, Cerdeira, P. 2009. *Wine Quality [Dataset]. UCI Machine Learning Repository.* https://doi.org/10.24432/C56S3T.

Jain, Kaushik, K. 2023. "Machine Learning-Based Predictive Modelling for the Enhancement of Wine Quality." *Scientific Reports* 13 (17042). https://doi.org/10.1038/s41598-023-44111-9.

Kniazieva, Y. 2023, October 12. *A Digital Sommelier: Machine Learning for Wine Quality Prediction.* https://labelyourdata.com/articles/machine-learning-for-wine-quality-prediction.