Wine Quality Dataset Evaluation and Visualization Report

Data Science II

COSC 4337

Submitted to

DR. Ricardo Vilalta

Submitted by

Dylan Flores (1872416)

Introduction

Understanding the complex relationship between the physicochemical properties of wine and its perceived quality is not only a scientific endeavor but also a key factor in maintaining and enhancing market reputation and consumer satisfaction. This is why I have chosen to perform an in-depth analysis of a comprehensive dataset that encompasses a wide range of wines, each characterized by detailed measurements of their chemical makeup and corresponding quality assessments by wine experts.

This dataset, representing a diverse collection of the Portuguese "Vinho Verde" wine, provides a unique opportunity to apply scientific analysis to uncover patterns and predictors of wine quality. By leveraging advanced data analysis techniques, I aimed to discern which physicochemical attributes most significantly influence wine quality ratings, and how these insights can be translated into practical applications in wine production processes.

The objective of this report is to distill complex data analyses into actionable insights that can be easily understood and applied to any wine production team. Using predictive modeling, I aimed to create a robust tool that can forecast wine quality based on measurable attributes, which enables preemptive adjustments to enhance the quality of product.

This report will outline the methods used in my analysis, discuss the implications of the findings, and present visualizations that highlight key data trends and modeling results. Ultimately, the insights derived from this work are expected to aid winemakers in making informed decisions that positively impact the quality and consistency of wine offerings.

Data Analysis Overview

The dataset included various chemical attributes of wine such as acidity levels, sugar content, and alcohol concentration, along with expert ratings for wine quality. I started a detailed examination of these properties to understand how they influence the perceived quality of wine.

- **Understanding Wine Characteristics:** I explored the basic characteristics of the wine data, identifying typical ranges and distributions for features like acidity and sugar content. This helped see the typical profiles of wines currently in production.
- **Relationships Between Features:** I analyzed how different chemical properties interact with each other. For example, I looked at how acidity levels correlate with the presence of sulfates, which can inform adjustments in wine processing to achieve desired taste profiles.
- **Quality Distribution Analysis:** I examined the distribution of quality ratings across the wines, identifying trends and outliers. This analysis was crucial to spot any consistent issues in batches that received lower ratings.

Predictive Modeling and Insights

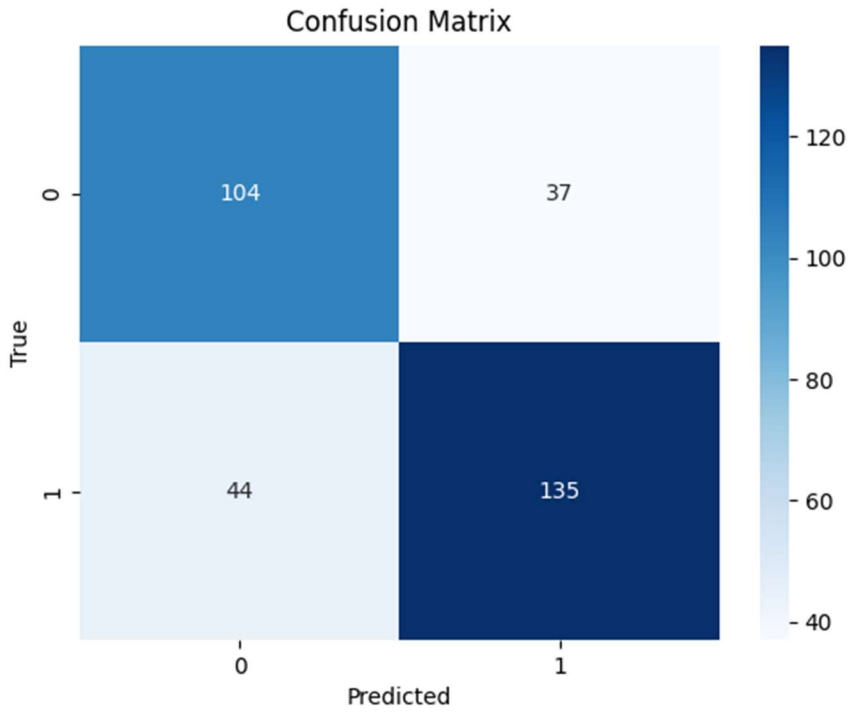
Using statistical models, I predicted wine quality based on its chemical properties, aiming to understand the impact of each attribute on the overall quality rating.

- **Model Performance:** I evaluated various models to find the most accurate one for predicting wine quality. The models helped understand which attributes most strongly predict higher quality, guiding quality improvement efforts.
- **Key Findings:**
    - Wines with lower volatile acidity and higher levels of sulfates tend to be rated higher in quality.
    - Alcohol level is a significant predictor of quality, with higher alcohol content often correlating with higher ratings.

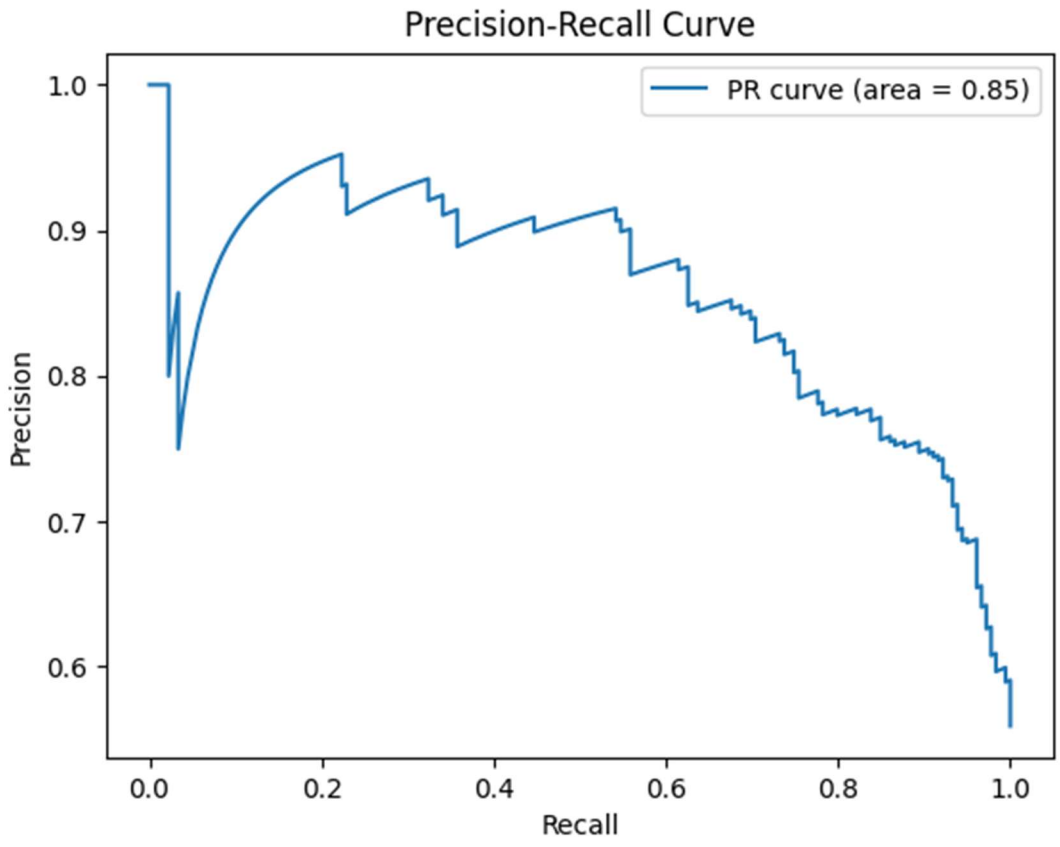
Visualizations and Their Implications

Several visualizations were created to aid in understanding and decision-making:

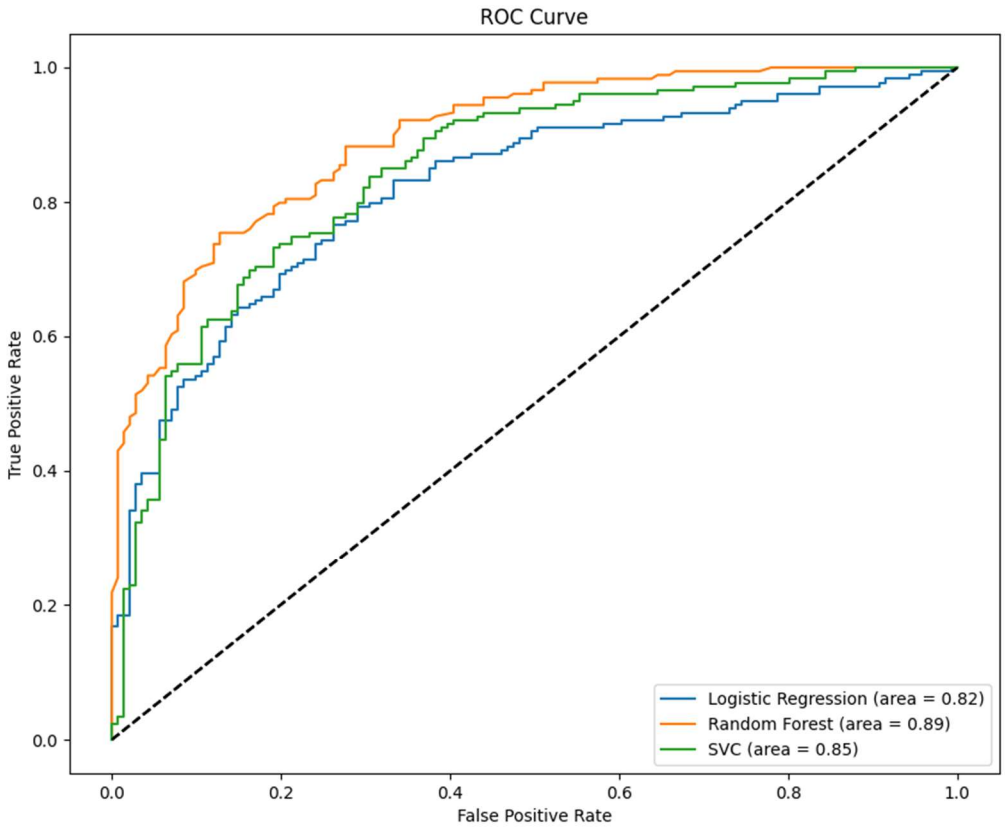
1. **Confusion Matrix:** This chart helps see the accuracy of our predictions versus actual ratings. It highlights cases where our predictions match the actual ratings and where discrepancies occur. For instance, in our analysis, the model predicted that 104 wines were of below-average quality, and this was accurate for those instances. It also accurately identified 135 wines as above-average quality. However, there were discrepancies: 37 instances were falsely predicted as below-average, and 44 instances were mistakenly labeled as above-average. This visualization is critical as it highlights the model's precision and areas where improvements are necessary, particularly in reducing false negatives, which could lead to undervaluing a potentially high-quality batch of wine.



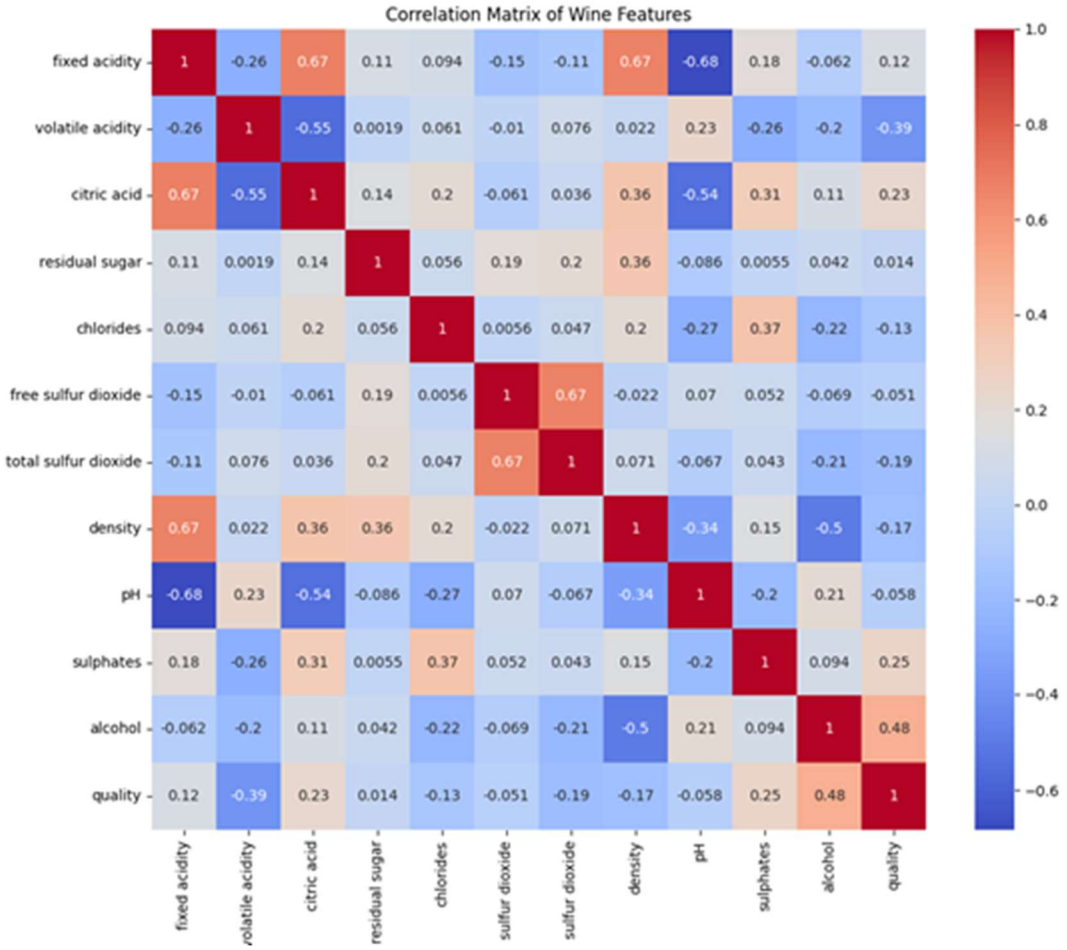
2. **Precision-Recall Curve:** This graph shows the trade-off between precision and recall for our best model. A higher area under the curve indicates that our model is effective at identifying high-quality wines. This precision-recall curve demonstrated the primary model exhibited an area under the curve (AUC) of 0.85. This high score indicates that our model has a strong capability to discern between lower and higher-quality wines reliably, maintaining a high precision across varying levels of recall. High precision ensures that when the model predicts a wine as high quality, it is very likely to be so, which is crucial in maintaining brand reputation and ensuring customer satisfaction from the wines. This curve helps in understanding the trade-off between catching as many positive cases as possible (high recall) and ensuring those positive predictions are correct (high precision).



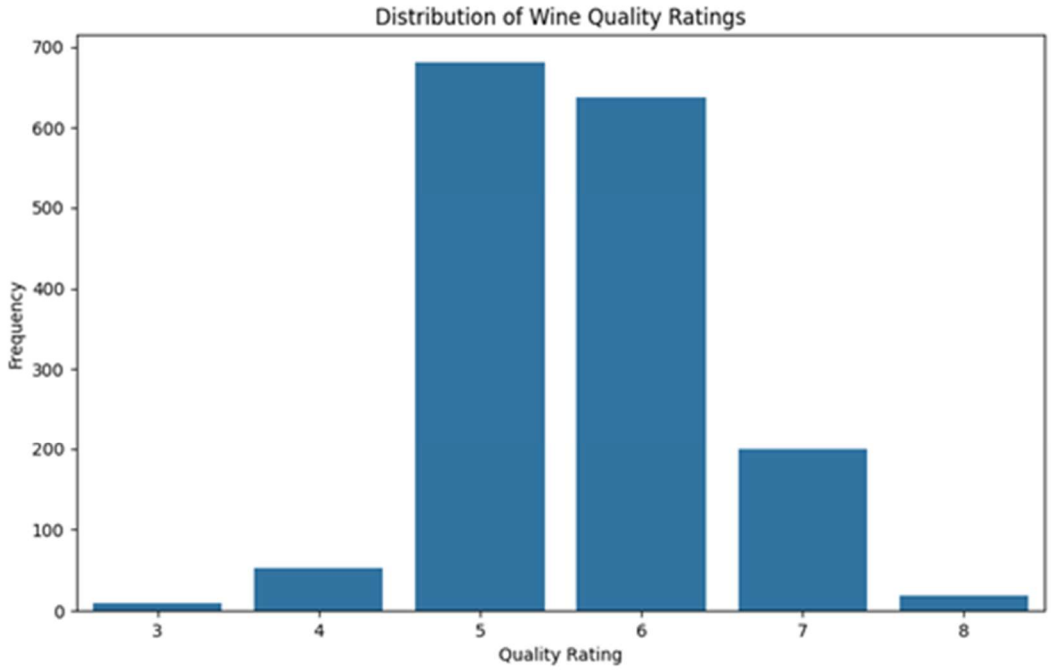
3. **ROC Curve:** Demonstrates the effectiveness of the models at distinguishing between higher and lower quality wines. The closer the curve follows the left and top borders, the more effective the model. The ROC curve was instrumental in comparing the efficacy of different models employed in the study. For example, the Random Forest model exhibited an AUC of 0.89, showcasing its superior ability to discriminate between the quality classes over Logistic Regression and SVC, which recorded AUCs of 0.82 and 0.85, respectively. This high AUC signifies that the Random Forest model has a high true positive rate and a low false positive rate, essential for minimizing misclassifications of wine quality.



4. **Correlation Matrix:** This matrix provided a snapshot of how different chemical properties of wine interrelate and their correlation to the quality rating. For instance, attributes like alcohol and sulphates showed positive correlations with quality, suggesting that higher levels might be associated with higher quality perceptions. This information is crucial for making informed adjustments in wine formulation.



5. **Distribution of Wine Quality Ratings:** By visualizing the distribution of quality ratings, I identified the most frequently occurring ratings and noted potential skews in data. This chart highlighted the concentration of wines rated around the median, guiding focus towards improving wines in this common quality range to shift the overall quality distribution upwards.



Conclusion From Visualizations:

These visualizations serve as compelling evidence of the models' capabilities and the critical relationships within the wine data. They guide the decision-making process, ensuring that efforts are focused on the most impactful factors. Moving forward, these visual tools will continue to play a vital role in the iterative process of quality enhancement, helping fine-tune production methods and better align with consumer expectations and industry standards.

Practical Applications and Recommendations

Based on my findings, I recommend the following actions to enhance wine quality:

- **Adjusting Chemical Properties:** Focus on controlling volatile acidity and optimizing sulfate levels during the winemaking process to enhance overall wine quality.
- **Targeted Quality Improvements:** Use the model predictions to identify batches that are predicted to have lower quality and prioritize these for chemical adjustments and taste testing.
- **Continuous Monitoring and Adjustments:** Implement regular testing of wine batches using the model, adjusting processes in real-time to ensure consistent quality.

Conclusion

The comprehensive analysis conducted on the Wine Quality dataset has yielded significant insights into the factors that contribute to the quality of wine. By integrating advanced data analysis techniques and predictive modeling, I have uncovered meaningful relationships between the physicochemical properties of wine and its quality ratings. These findings not only enhance the understanding of wine characteristics but also serve as a catalyst for refining many winemaking processes.

The predictive models, particularly the Random Forest model, which can be visualized in the ROC curve, demonstrated a strong ability to identify high-quality wines based on their chemical profiles. This capability allows predicting the quality of wine early in the production process, providing an opportunity for timely interventions that can significantly enhance the final product. The models' effectiveness, as illustrated through various metrics and visualizations such as the

ROC curve and precision-recall graph, confirms their reliability and practical utility in any program that deals with quality assurance.

Furthermore, the analysis has pinpointed specific chemical attributes, such as volatile acidity and sulfate level, that are critical predictors of wine quality. Adjusting these properties within targeted ranges can lead to consistent improvements in wine quality, directly impacting factors like consumer satisfaction and market competitiveness. The visual representations of data, particularly the confusion matrices and distribution charts, have provided clear and intuitive insights that are easily interpretable.

In conclusion, this project has not only reinforced the importance of data-driven decision-making in the wine industry but has also set a pathway for continuous improvement and innovation in production techniques. If Wine manufacturers committed to utilizing these techniques to ensure that each bottle of wine produced exceeds the expectations of wine enthusiasts, there would be a reduction in the frequency of lower ratings we see in the Wine Quality Dataset used for this project.

Steps for running the code attached

1. Import both “Visualizations_for_report_3.py” and “winequality-red.csv” into an open PyCharm project.
2. Ensure there are no errors with the libraries at the top of the file being imported and if there are errors make sure that you install each library before trying to run the code.
3. Once there are no errors and both files are in the directory of the PyCharm project simply click run and all visualizations should appear.