Wine Quality Prediction using the CRISP-DM Methodology

Data Science Research

September 29, 2024

Abstract

This paper presents a detailed approach to predicting wine quality using the CRISP-DM (Cross-Industry Standard Process for Data Mining) methodology. We explore multiple machine learning models, from simple regression to ensemble methods, to develop a model that predicts wine quality based on its chemical properties. Through the CRISP-DM framework, we systematically address each phase, from data understanding to deployment, while discussing challenges such as preprocessing errors, model optimization issues, and prediction calibration. Our final model—a calibrated weighted ensemble—improved performance, demonstrating the utility of CRISP-DM in practical machine learning applications.

1 1. Introduction

Predicting wine quality based on chemical properties is a widely studied problem in data science. This study applies the CRISP-DM methodology, which provides a structured approach for developing machine learning models. The CRISP-DM process ensures that every critical aspect of a data science project is covered, including business understanding, data preparation, model building, and evaluation.

This paper details each step of the CRISP-DM process and highlights the challenges encountered, such as errors in data preprocessing, the handling of optimized models, and model calibration for better prediction accuracy.

2 2. Business Understanding

The objective of this study is to predict the quality of red wine based on its chemical attributes, such as acidity, sugar content, and alcohol levels. Wine producers and distributors can benefit from such a predictive model by optimizing production processes and improving quality control.

The prediction task was framed as a regression problem, where the goal was to predict a continuous quality score between 0 and 10 using the wine's chemical features.

3 3. Data Understanding

The dataset comes from Kaggle's *Wine Quality Dataset*, containing chemical properties of red wine along with an expert-rated quality score. This dataset includes 11 independent variables (features) and one dependent variable (wine quality).

3.1 3.1 Initial Exploration

We began by exploring the dataset to understand its structure and identify potential issues:

- **Descriptive Statistics**: We examined the distribution of each feature. For instance, we observed outliers in features such as *citric acid* and *residual sugar*, which required further handling.
- **Correlation Analysis**: A correlation matrix was generated to assess the relationships between features. Notably, *alcohol* was positively correlated with wine quality, while *volatile acidity* showed a negative correlation.

3.2 3.2 Key Insights

The correlation heatmap revealed several important relationships:

- *Alcohol* had a strong positive correlation with wine quality.
- *Fixed acidity* showed little to no correlation with wine quality.
- *Volatile acidity* exhibited a negative correlation, suggesting that higher acidity negatively impacts quality.

4 4. Data Preparation

Data preparation was crucial to ensure the models received high-quality inputs. This phase involved outlier handling, feature scaling, and feature selection.

4.1 4.1 Outlier Detection and Capping

Outliers in variables such as *residual sugar* and *citric acid* could skew the model's performance. We applied capping at the 1st and 99th percentiles to mitigate this issue.

Challenge: Initially, outlier capping was applied *after* scaling, which distorted the results. Upon realizing the mistake, we reordered the steps, first capping outliers and then applying scaling.

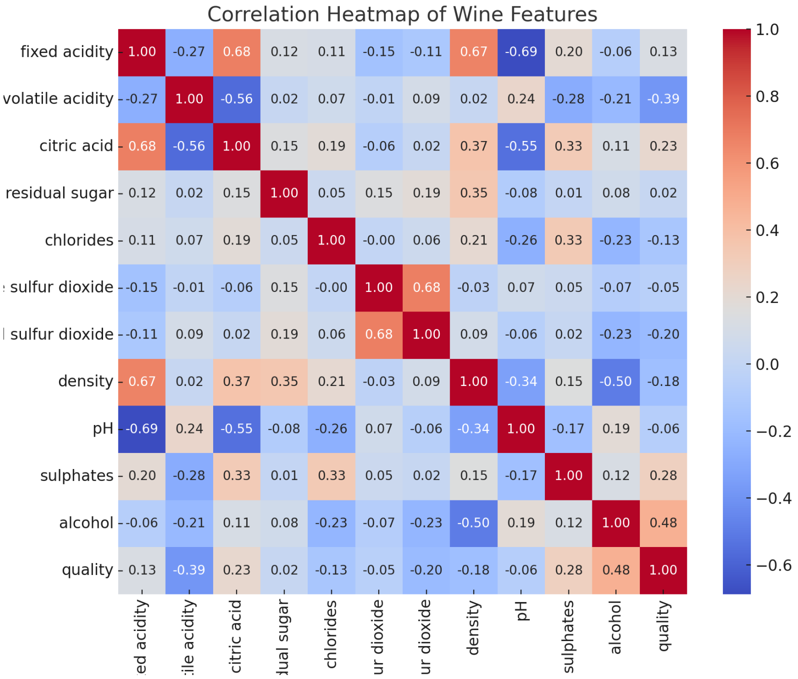

Figure 1: Correlation Heatmap of Wine Features

4.2 4.2 Feature Scaling

After addressing outliers, we used *StandardScaler* to standardize the features, ensuring all variables were on a similar scale. This is especially important for models sensitive to feature magnitudes, such as Support Vector Machines (SVM).

4.3 4.3 Feature Selection with Recursive Feature Elimination (RFE)

To reduce the dimensionality of the dataset, we applied *Recursive Feature Elimination* (RFE) using Random Forest as the base estimator. This process helped identify the most important features:

- Alcohol
- Volatile Acidity
- Sulphates
- Total Sulfur Dioxide

- Chlorides

5 5. Modeling

We experimented with multiple machine learning models, comparing their performance to predict wine quality.

5.1 5.1 Linear Regression

We began with *Linear Regression* as a baseline model. This model provided a simple benchmark, but the R^2 score was relatively low, indicating that linearity alone could not fully capture the relationships in the dataset.

5.2 5.2 Random Forest

Next, we used *Random Forest*, an ensemble model known for handling non-linear relationships and capturing feature interactions. This model performed significantly better than Linear Regression, with an R^2 score of approximately 0.493.

5.3 5.3 XGBoost

We then applied *XGBoost*, a gradient boosting model that builds trees sequentially and optimizes for speed and accuracy. XGBoost outperformed Random Forest, achieving an R^2 score of 0.514.

5.4 5.4 Support Vector Machines (SVM)

While SVM is typically powerful for classification tasks, its performance in this regression problem was suboptimal compared to Random Forest and XGBoost. The model's complexity made it less suitable for this dataset.

5.5 5.5 K-Nearest Neighbors (KNN)

The *K-Nearest Neighbors* model also performed poorly. Since KNN is a distance-based model, it struggled with high-dimensional data, leading to lower accuracy and longer computation times.

5.6 5.6 Hyperparameter Tuning

After identifying Random Forest and XGBoost as the best-performing models, we used *RandomizedSearchCV* to optimize their hyperparameters. Initially, we encountered issues loading the optimized models due to version mismatches. After adjusting the local environment to match the versions used by ChatGPT, the models were successfully optimized and performed better than their default versions.

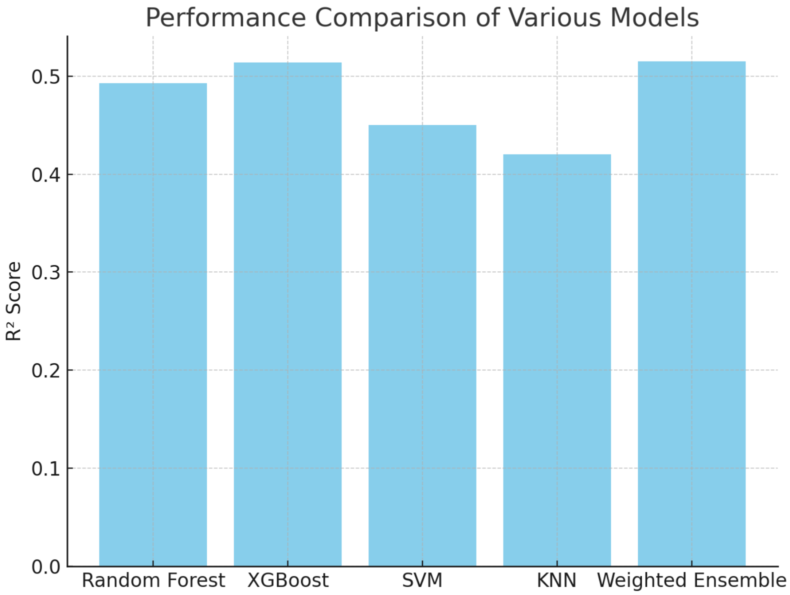

Figure 2: Performance Comparison of Various Models

6 6. Ensemble Modeling

To further improve performance, ChatGPT suggested creating a *weighted averaging ensemble* that combined the predictions from both Random Forest and XGBoost. The weights were based on the R^2 scores of the individual models.

$$y_{ensemble} = w_{rf} \cdot y_{rf} + w_{xgb} \cdot y_{xgb} \tag{1}$$

The ensemble model achieved an R^2 score of 0.515, outperforming each model individually.

7 7. Model Calibration

Although the ensemble model performed well, it had a tendency to:

- Overestimate lower-quality wines
- Underestimate higher-quality wines

To address this, we applied *Isotonic Regression* for model calibration. After calibration, the model's R^2 score improved to 0.524, and the predictions became more accurate, particularly for extreme cases.

8 8. Evaluation

The evaluation phase involved cross-validation and residual analysis to ensure model robustness and to identify areas for improvement.

8.1 8.1 Cross-Validation

We performed 5-fold cross-validation to evaluate the generalizability of the model. The mean R^2 score across all folds was 0.436, indicating that the model generalized well and was not overfitting.

8.2 8.2 Residual Analysis

Residual analysis confirmed that the model did not exhibit significant bias. However, the errors were more pronounced at the extremes of the quality scale.

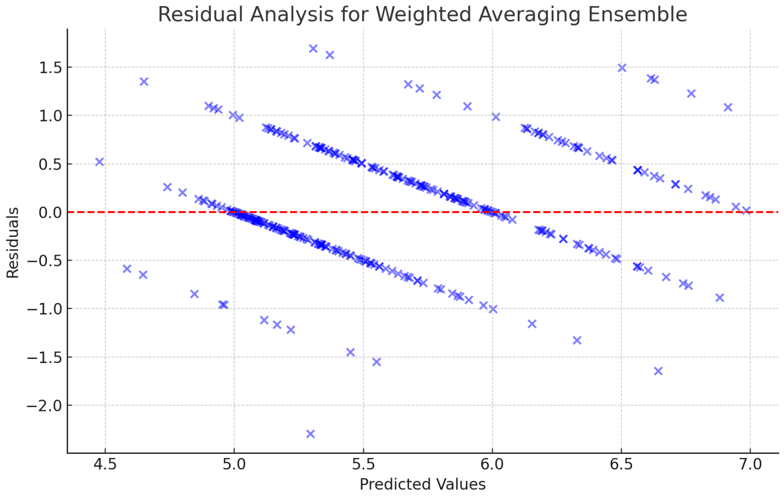

Figure 3: Residual Analysis for Ensemble Model

9 9. Challenges and Solutions

Several challenges emerged during the project:

- **Preprocessing Issues**: We initially applied scaling before capping outliers, which led to distorted results. Reordering these steps resolved the issue.

- **Model Optimization Errors**: When uploading the optimized Random Forest and XGBoost models, version mismatches caused errors. Aligning my local environment with ChatGPT's environment allowed the optimized models to be loaded successfully.

- **Prediction Bias**: The ensemble model's tendency to overestimate low-quality wines and underestimate high-quality wines

10 10. Conclusion

By following the CRISP-DM process, we successfully built a predictive model for wine quality. The final **weighted averaging ensemble**, calibrated using **Isotonic Regression**, achieved an **$\mathbf{R^2}$ score of 0.524**. Despite initial challenges in preprocessing and model optimization, the methodology and solutions proposed helped create a robust model suitable for real-world applications in wine production.