Machine Learning Algorithm Selection Report for Wine Quality Prediction

Farshad Ghorbanishovaneh

Abstract

This report presents an evaluation of various machine learning algorithms designed to predict the quality of wine based on its physicochemical attributes. It includes detailed comparisons of performance metrics such as accuracy, mean squared error (MSE), mean absolute error (MAE), R-squared (R^2), and root mean squared error (RMSE). The study identifies the most effective algorithms by analyzing these metrics. Additionally, it examines the impact of feature importance, the necessity of scaling data for certain models, and the benefits of cross-validation.

1. Introduction

Predicting wine quality using machine learning is feasible even on a personal computer, thanks to the simplicity of the dataset. This study evaluates multiple algorithms on a dataset of wine samples, each characterized by features such as acidity, sugar content, and alcohol level, to determine the best predictors of quality. The focus is on both the performance of individual models and the efficacy of ensemble methods.

2. Dataset Description

The Wine Quality Dataset consists of 1,599 samples of wine, each characterized by 12 physicochemical properties and quality ratings. The properties include various acidity measures, sugar, sulfur dioxide levels, and alcohol content. All columns in the dataset are fully populated with non-null values. The quality ratings vary from 3 to 8, with the majority of the wines rated between 5 and 6. Here is a detailed breakdown of the dataset structure and a correlation heatmap to visualize the relationships between different properties:

Attribute	Description	Data Type
Fixed Acidity	Most acids involved with wine	float64
Volatile Acidity	Amount of acetic acid in wine	float64
Citric Acid	Found in small quantities, citric acid	float64
Residual Sugar	Amount of sugar remaining after fermentation	float64
Chlorides	Amount of salt in the wine	float64
Free Sulfur Dioxide	Free form of SO2; prevents microbial growth	float64
Total Sulfur Dioxide	Amount of free and bound forms of S02	float64
Density	Density of the wine	float64
pH	Describes the acidity or basicity of wine	float64
Sulphates	Wine additive which can contribute to sulfur dioxide gas (S02) levels	float64
Alcohol	Alcohol content of the wine	float64
Quality	Score between 3 and 8 (inclusive)	int64

Table 1. Dataset Description

Besides the 'quality' column, which serves as our target variable, all other attributes in our dataset are of the float data type. The distribution of these data points is depicted in the boxplot Figure 1. In Figure 2, you can observe the correlations between each variable within the dataset, reflecting how each physicochemical property relates to the others and to the overall quality of the wine. The strongest positive correlation with quality is exhibited by alcohol content, suggesting that higher alcohol levels often correspond to higher quality ratings. Conversely, volatile acidity has the strongest negative correlation

with quality, indicating that lower volatile acidity is associated with higher quality wines. Other properties, such as sulphates and citric acid, display weaker positive correlations with quality, suggesting a modest relationship to higher quality. In contrast, density and total sulfur dioxide show weaker negative correlations, implying that lower values in these properties might be characteristic of higher quality wines.

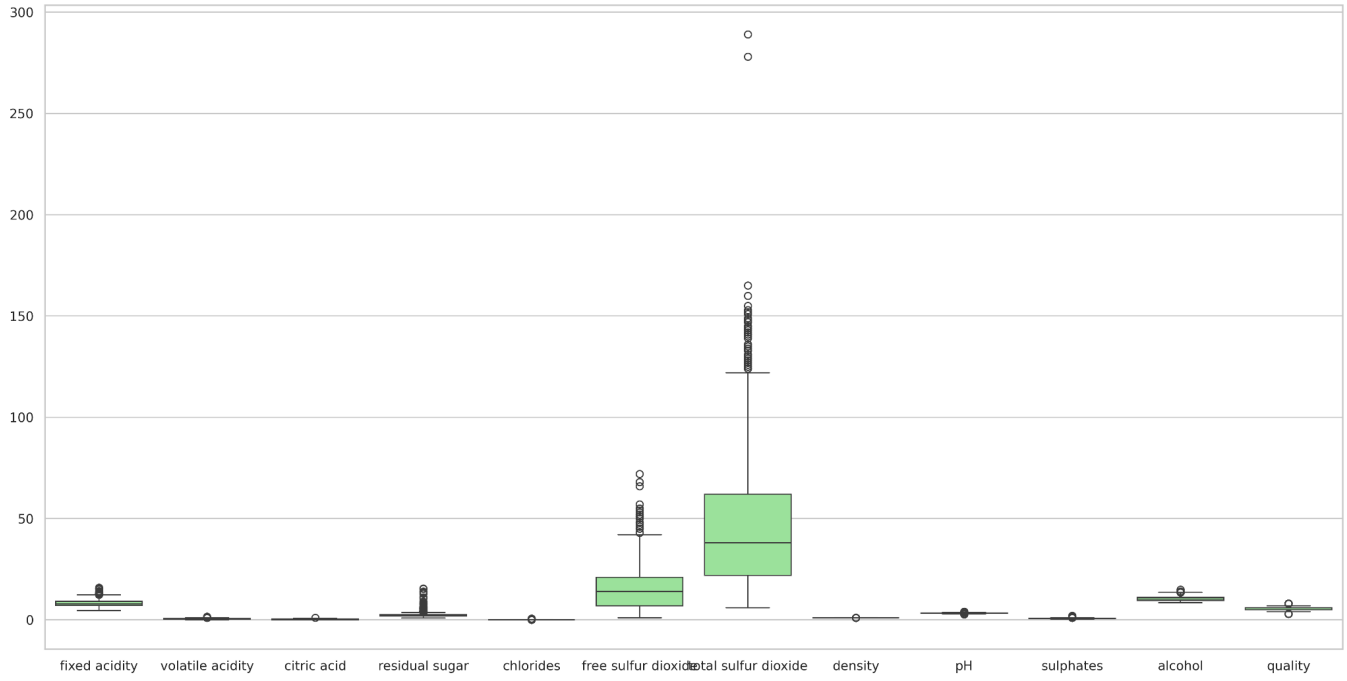

Figure1. The image shows a boxplot for each physicochemical property and quality rating in the Wine Quality Dataset.

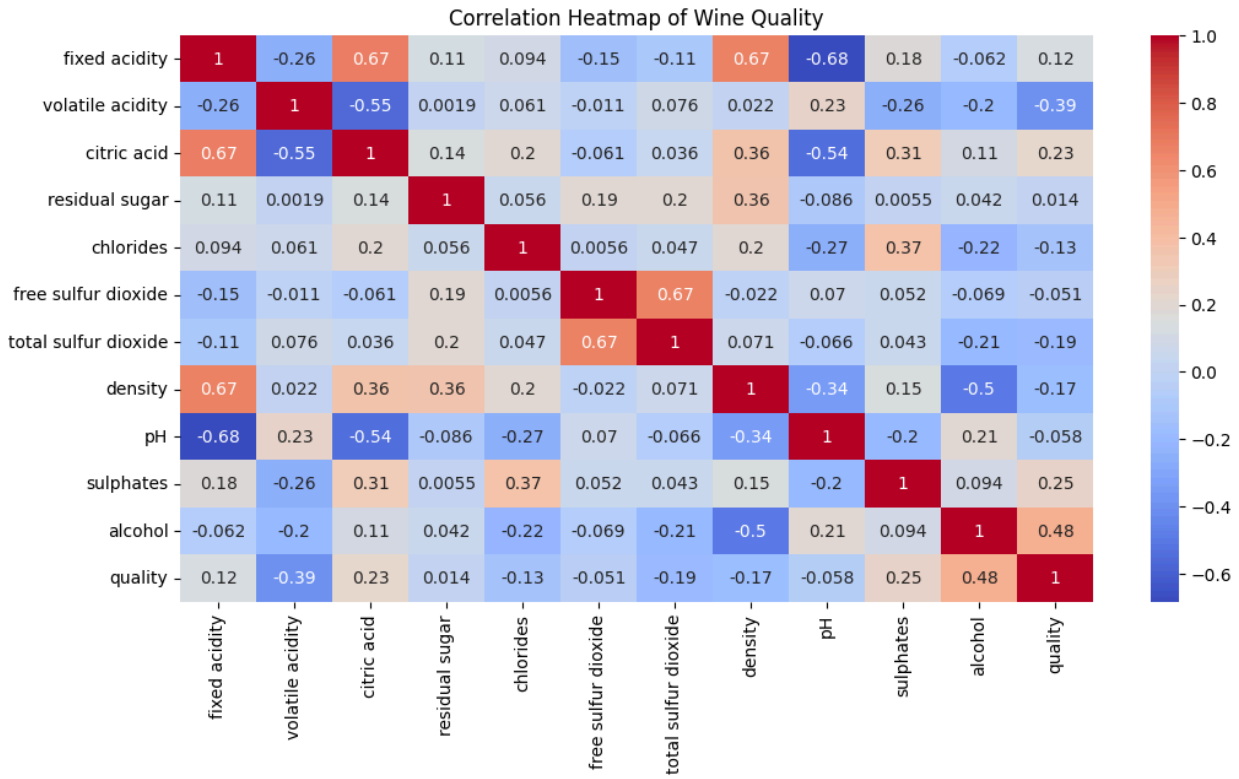

Figure 2. Correlation heatmap of wine quality dataset

3. Experimental Setup

To address the problem of predicting wine quality based on physicochemical properties, the following experimental setup is employed:

Programming Languages and Libraries

- **Python**: The primary programming language used for handling data.
- **Pandas**: Employed for efficient data manipulation and preprocessing.
- **NumPy**: Utilized for high-performance numerical computing.
- **Scikit-learn**: A versatile library that provides straightforward implementation for a wide array of machine learning algorithms.
- **Matplotlib and Seaborn**: Both are used for generating various plots and visualizations to analyze the data and interpret the model results.

- **Jupyter Notebooks**: Chosen for documenting the analysis process, enabling interactive development, and sharing the findings.

Data Preparation and Preprocessing

The initial steps taken to prepare the data for machine learning model ingestion involved the following procedures:

- Cleaning: Confirmed that the dataset has no missing or null values, which simplifies the preprocessing phase.
- Scaling: Applied standardization to the features to mitigate the risk of certain attributes disproportionately influencing the model due to their scale.
- Splitting: Segregated the data into training and testing sets to ensure a reliable evaluation of the models' predictive capabilities.

Model Selection and Implementation

A range of machine learning models were chosen and implemented to explore their efficacy in predicting the quality of wine:

Machine Learning Algorithms

- **Linear Models**:
    - **Linear Regression**: Serves as the benchmark model to set a baseline for performance. Its simplicity and interpretability make it a vital point of comparison.
    - **Logistic Regression**: Though traditionally used for classification, logistic regression is employed here to assess its performance in a probabilistic setting for quality ranking.
- **Tree-Based Models**:

    - **Decision Tree Regressor**: A model that breaks down the data by making decisions based on individual feature value thresholds, providing insight into the structural relationship between features and the target.
    - **Random Forest Regressor**: An ensemble of decision trees, this model increases the predictive robustness by averaging the results of individual trees, thus reducing the risk of overfitting.
- **Instance-Based Models**:
    - **K-Nearest Neighbors (KNN)**: This algorithm predicts the quality of wine by averaging the target values of the nearest data points, offering an intuitive, distance-based approach to prediction.
- **Probabilistic Models**:
    - **Gaussian Naive Bayes**: A classifier based on Bayes' theorem, which assumes feature independence. It's included to evaluate how probabilistic reasoning fares in quality prediction.
- **Dimensionality Reduction and Discriminant Analysis**:
    - **Linear Discriminant Analysis (LDA)**: While typically used for dimensionality reduction, LDA is also a powerful classifier, aiming to find the linear combinations of features that best separate classes.
- **Support Vector Machines**:
    - **Support Vector Machine (SVM)**: A robust classifier, SVM finds the hyperplane that best separates the classes in high-dimensional space, and is tested for its effectiveness in non-linear classification through the use of kernel tricks.

Performance Measurement and Validation

In order to gauge the effectiveness and accuracy of each model, a suite of performance metrics was selected, and a robust validation methodology was implemented:

Evaluation Metrics

- **Mean Squared Error (MSE):** Gauges the average squared difference between the estimated values and the actual value.
- **Root Mean Squared Error (RMSE):** The square root of MSE that represents the sample standard deviation of the differences between predicted values and observed values.
- **R-squared (R^2):** Reflects the percentage of the dependent variable's variance that the model explains.
- **Accuracy:** Used for classification models to measure the proportion of correctly predicted instances.

Validation Methodology

- Cross-Validation: Employed K-fold cross-validation to provide insight into the model's effectiveness across different data segments, ensuring that the evaluation is not biased by the particular split of the train-test data.

4. Result

The analysis comprises four categories: Scaled data, Unscaled data, Cross-validation with scaled data, and Cross-validation with unscaled data. Various machine learning models were trained on both scaled and unscaled data, with evaluation conducted using metrics such as Root Mean Squared Error (RMSE), Mean Squared Error (MSE), Mean Absolute Error (MAE), R-squared (R^2), and, in some instances, accuracy. Additionally, a 9-fold cross-validation approach was employed for assessing model performance.

Figure 3. Model performance

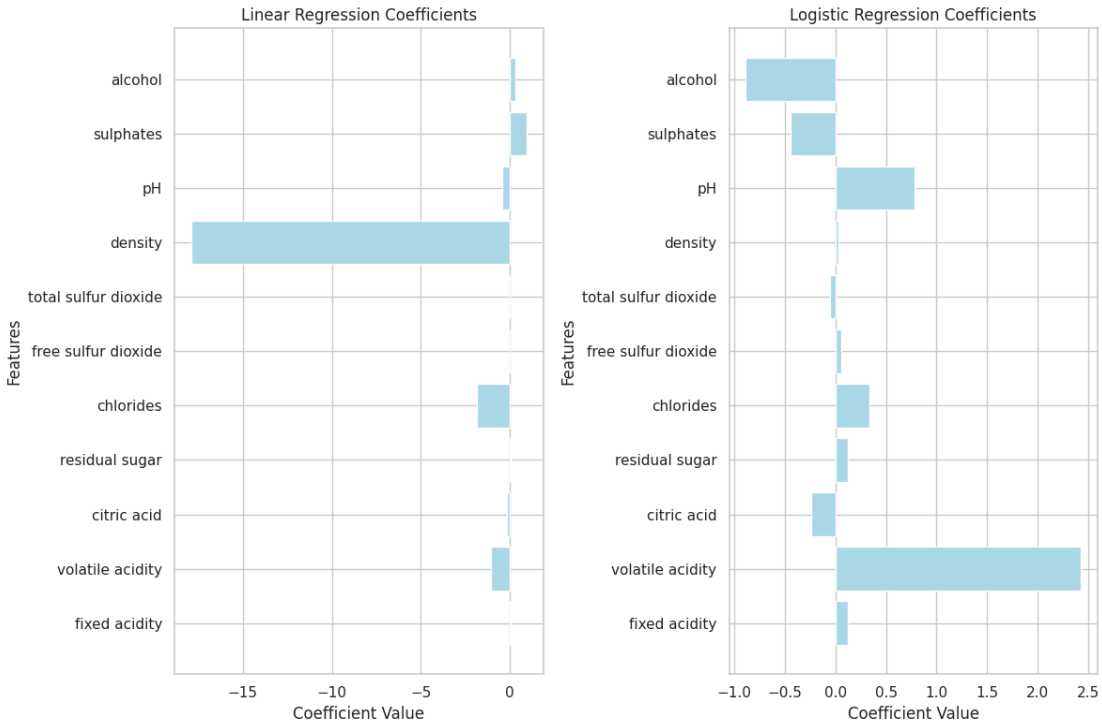

Figure 4. Linear Regression and Logistic Regression Coefficients

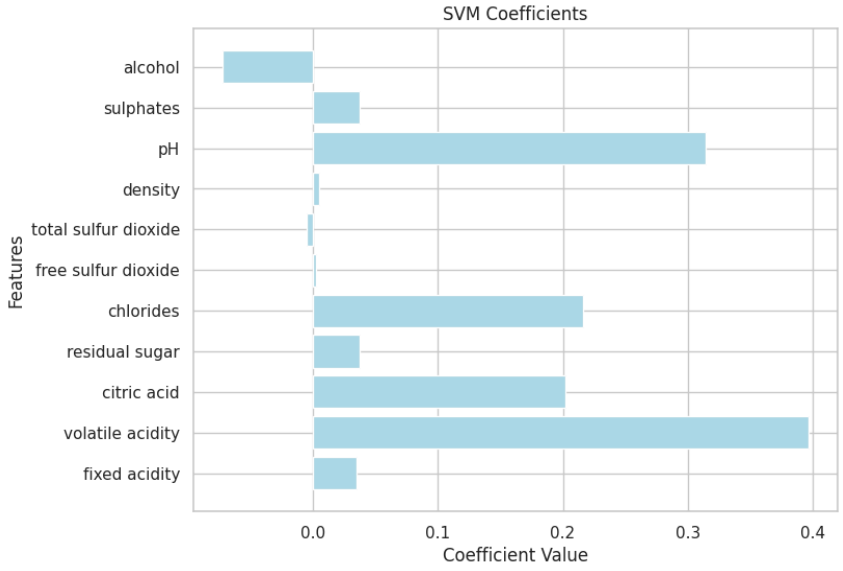

Figure 4. Support Vector Machine Coefficients

Feature importance, as depicted in Figures 5 and 6, highlights the significance of features in linear and logistic regression models. Positive coefficients denote a positive correlation, while negative coefficients indicate a negative correlation. Notably, larger coefficients suggest a stronger influence, exemplified by 'density' (-17.881) and 'sulphates' (0.916) in linear regression. In logistic regression, coefficients represent log-odds, impacting class likelihood, as observed in 'volatile acidity' (2.430) and 'alcohol' (-0.887). Additionally, in SVM, coefficients reflect feature weights affecting classification decisions, prominently seen in 'volatile acidity' (2.430) and 'alcohol' (-0.887) influencing classification outcomes.

Overall, Random Forest emerges as the top-performing model among those evaluated, with Logistic Regression also demonstrating respectable performance, particularly in terms of classification accuracy. Notably, the impact of scaling on model performance is evident, particularly for SVM and KNN, as illustrated in Figure 3.