Predicting Wine Quality Using SEMMA Methodology

Neeharika Singh

Abstract

This research paper aims to predict the quality of wine based on various physicochemical properties. We employ the SEMMA (Sample, Explore, Modify, Model, Assess) methodology to guide the data science process. Our findings reveal that a Random Forest model performs reasonably well on both validation and unseen datasets.

1 Introduction

Predicting the quality of wine based on its physicochemical features is an intriguing yet challenging problem. This paper employs the SEMMA methodology, providing a structured approach for this data science task.

2 Sample: Getting to Know the Data

2.1 Dataset Structure

The dataset comprises 1123 rows and 13 columns, including features like fixed acidity, volatile acidity, citric acid, residual sugar, and a target variable of wine quality.

2.2 Key Findings

The dataset has no missing values, and the features have different scales and ranges.

3 Explore: Digging Deeper

3.1 Missing Values

No missing values are present in the dataset, making it ready for further analysis.

3.2 Feature Distributions

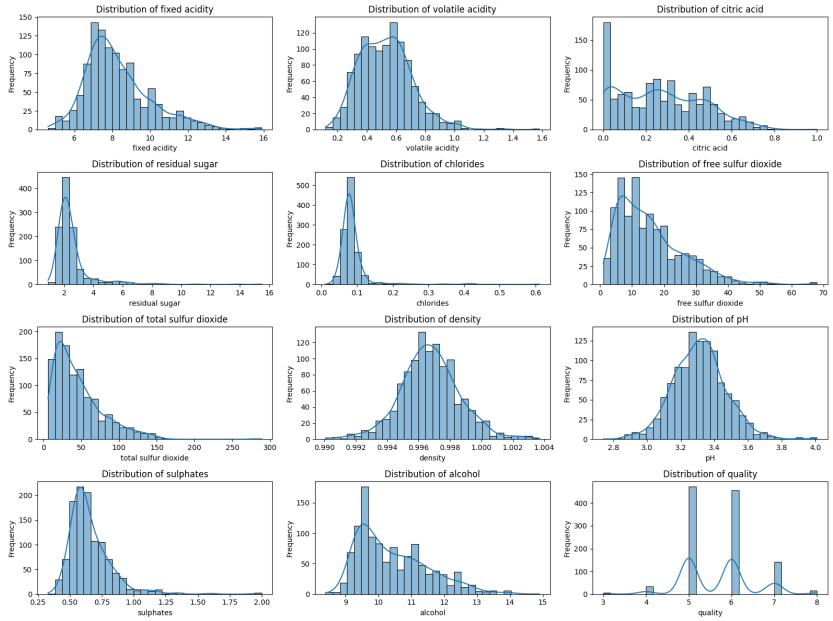
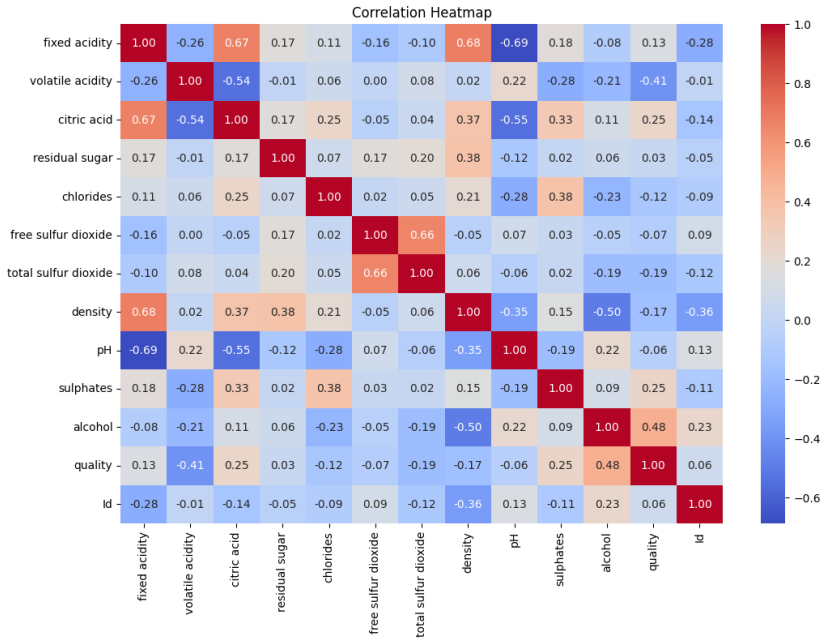


Figure 1: Histograms of feature distributions

3.3 Correlations

A heatmap and correlation coefficients reveal that alcohol content positively correlates with wine quality, whereas volatile acidity negatively correlates



with it.

4 Modify: Data Preparation

4.1 Feature Scaling

Standard scaling is used to normalize the features, making them ready for machine learning algorithms.

4.2 Data Splitting

The dataset is split into 80% for training and 20% for validation.

5 Model: Building the Prediction Model

5.1 Model Choice

A Random Forest Regressor is chosen for its ability to capture complex feature interactions and its robustness to overfitting.

5.2 Model Training and Validation

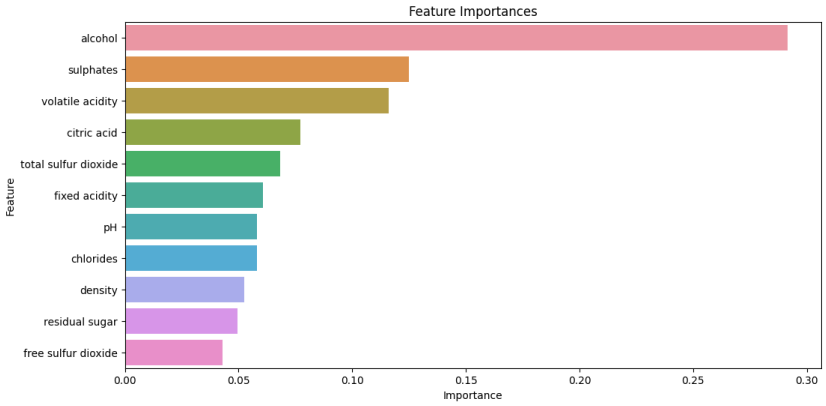
The model is trained on the training set, and initial validation metrics are as follows:

$$\text{RMSE} = 0.582$$
$$R^2 = 0.441$$

6 Assess: Model Evaluation

6.1 Feature Importance

The most influential features are alcohol content, followed by sulphates and volatile acidity.



6.2 Model Evaluation on Unseen Data

The model performs well on an unseen dataset with the following metrics:

$$\text{RMSE (Unseen)} = 0.428$$
$$R^2(Unseen) = 0.601$$

7 Conclusion

The Random Forest model provides a reasonably good prediction of wine quality. Future work may include hyperparameter tuning, exploring advanced models, and feature engineering.