# WINE QUALITY PREDICTION REPORT

## TABLE OF CONTENTS

## Contents

# INTRODUCTION

Wine quality assessment has been a subject of interest for vintners, sommeliers, and wine enthusiasts for centuries. With the advent of modern technology and data science, we can now leverage machine learning techniques to predict the quality of wine based on its physicochemical properties. This project aims to develop predictive models to classify wines into good and poor-quality categories using various machine learning algorithms. By doing so, we hope to provide an automated, objective method for wine quality assessment that can assist winemakers in improving and sustaining the quality of their products.

In this study, I utilized a dataset of wine samples, which includes several features such as acidity, sugar content, pH, and alcohol levels. These features are crucial indicators of the wine's quality. I applied data preprocessing techniques to clean and normalize the data, followed by exploratory data analysis to understand the underlying patterns. Various machine learning models, including Logistic Regression, Random Forest, Gradient Boosting, and Support Vector Machine, were trained and evaluated to determine the most effective model for predicting wine quality. The performance of these models was compared using metrics such as accuracy, F1-score, and mean absolute error, providing valuable insights into their predictive capabilities.

## PROBLEM STATEMENT

Wine quality is traditionally assessed by human experts through sensory evaluation, which can be subjective and inconsistent. This manual assessment process is time-consuming and requires a high level of expertise, which might not always be readily available. Moreover, the traditional methods of quality evaluation are not scalable, especially when dealing with large volumes of wine production. This can lead to variations in quality and potentially affect the marketability and consumer satisfaction of the wine products.

To address these challenges, there is a need for an automated and reliable system that can predict wine quality based on measurable physicochemical properties. By developing a machine learning-based predictive model, we can provide a consistent, objective, and scalable solution for wine quality assessment. This would not only streamline the quality control process for winemakers but also ensure that consumers receive high-quality products consistently. The goal of this project is to leverage data science and machine learning techniques to create a robust model capable of accurately classifying wines into good and poor-quality categories based on their chemical composition.

### OBJECTIVES

i. Extract meaningful insights from the wine dataset.
ii. Identify key physicochemical factors that contribute to wine quality.
iii. Develop a robust classification model to accurately predict wine quality.
iv. Optimize and refine the classification model to enhance predictive performance.

## DATA DESCRIPTION

The dataset consists of 13 features and a sample size of 1143, each representing different physicochemical properties of wines along with their quality ratings. The features in the dataset are as follows:

1. **fixed acidity**: Fixed acids such as tartaric acid in wine, measured in g/dm³.
2. **volatile acidity**: Volatile acids such as acetic acid in wine, measured in g/dm³.
3. **citric acid**: Citric acid in wine, measured in g/dm³.
4. **residual sugar**: Residual sugar left after fermentation, measured in g/dm³.
5. **chlorides**: Chlorides in wine, measured in g/dm³.
6. **free sulfur dioxide**: Free SO2 in wine, measured in mg/dm³.
7. **total sulfur dioxide**: Total SO2 in wine, measured in mg/dm³.
8. **density**: Density of the wine, measured in g/cm³.
9. **pH**: pH level of the wine.
10. **sulphates**: Sulphates in wine, measured in g/dm³.
11. **alcohol**: Alcohol content in wine, measured in % vol.
12. **quality**: Quality rating of the wine, scored between 3 and 8.
13. **Id**: Unique identifier for each wine sample.

In the analysis, I converted the 'quality' column into a binary classification problem, where wines rated 7 or above are considered "good-quality" (labelled as 1), and those rated below 7 are considered "poor-quality" (labelled as 0). The 'Id' column was dropped from the dataset as it does not contribute to the prediction of wine quality.

# DATA ACQUISITION

The dataset was sourced from a publicly available wine quality dataset in Kaggle: https://www.kaggle.com/datasets/yasserh/wine-quality-dataset

# DATA PRE-PROCESSING

Before proceeding with modeling, the dataset underwent several preprocessing steps to ensure its suitability for analysis:

- **Column Removal**: The 'Id' column was removed from the dataset as it did not contribute to the predictive modeling process.
- **Binary Encoding**: The quality scores were converted into a binary classification where wines with a score of 7 or higher were labeled as 'good quality' and others as 'poor quality'. This transformation simplified the target variable for classification purposes.
- **Handling Missing Values**: The dataset was inspected for any missing values. Fortunately, no missing values were present in the dataset, eliminating the need for imputation or removal of entries.

# EXPLORATORY DATA ANALYSIS

Exploratory Data Analysis (EDA) was conducted to gain insights into the dataset and understand the relationships between different variables.

## SUMMARY STATISTICS

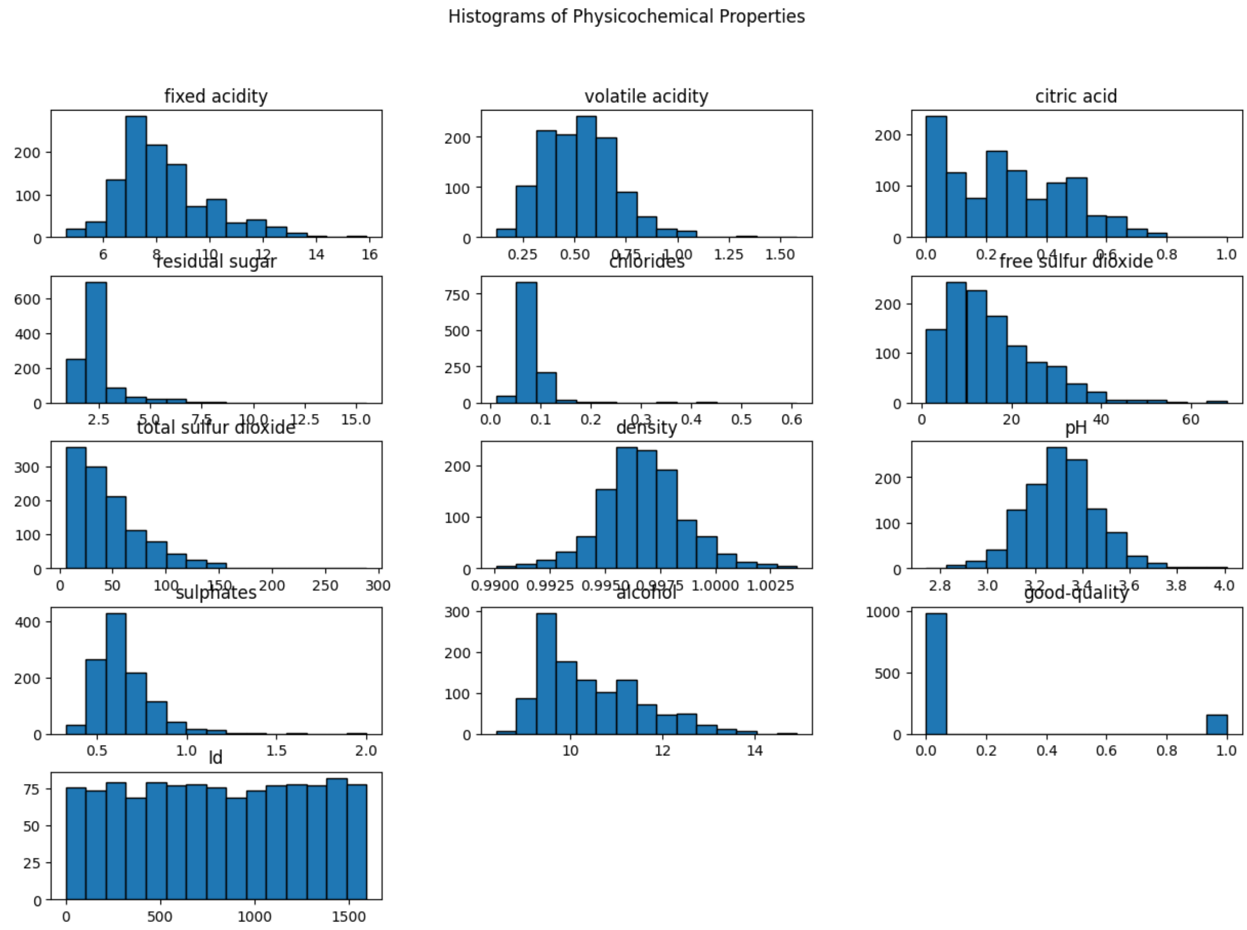
Summary statistics, including mean, median, standard deviation, minimum, and maximum values, were computed for each physicochemical property. These metrics provided a comprehensive overview of the central tendency and spread of the dataset.

- The alcohol content ranges from 8.4 to 14.9 with a mean of approximately 10.44.

- The distribution of alcohol content suggests that most wines have alcohol levels around the mean, with some wines having notably lower or higher alcohol content.
- Quality is represented on a scale from 0 to 8, with a mean quality score of approximately 5.66. This indicates that the majority of wines fall within the middle range of quality
- The total count is 1143

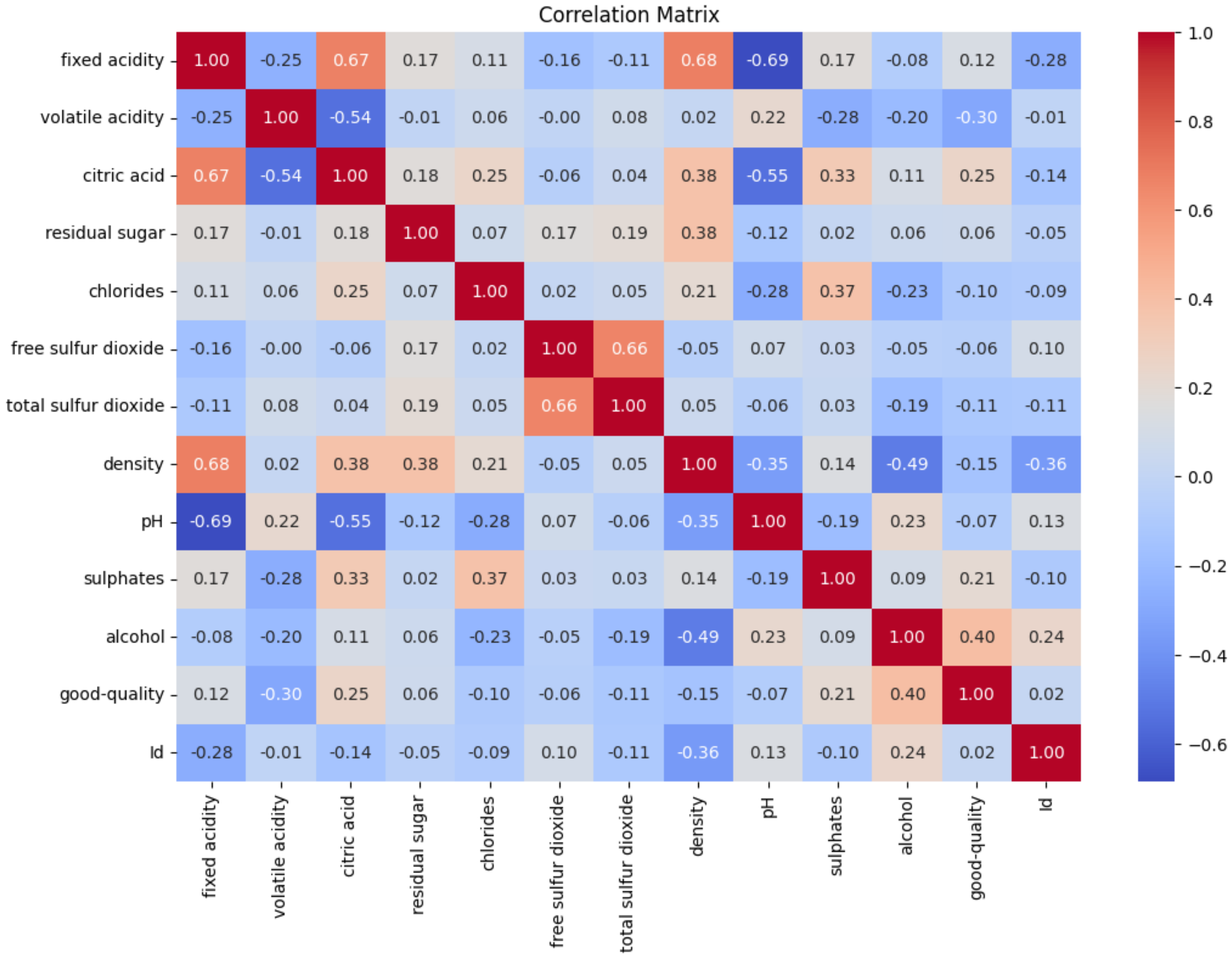
## DISTRIBUTION OF FEATURES

- Histograms were used to visualize the distribution of physicochemical properties such as fixed acidity, volatile acidity, citric acid, etc.
- These plots revealed the skewness and kurtosis of the feature distributions.
- Most of the features were right-skewed with a few like pH and density being approximately normally distributed.



## CORRELATION ANALYSIS

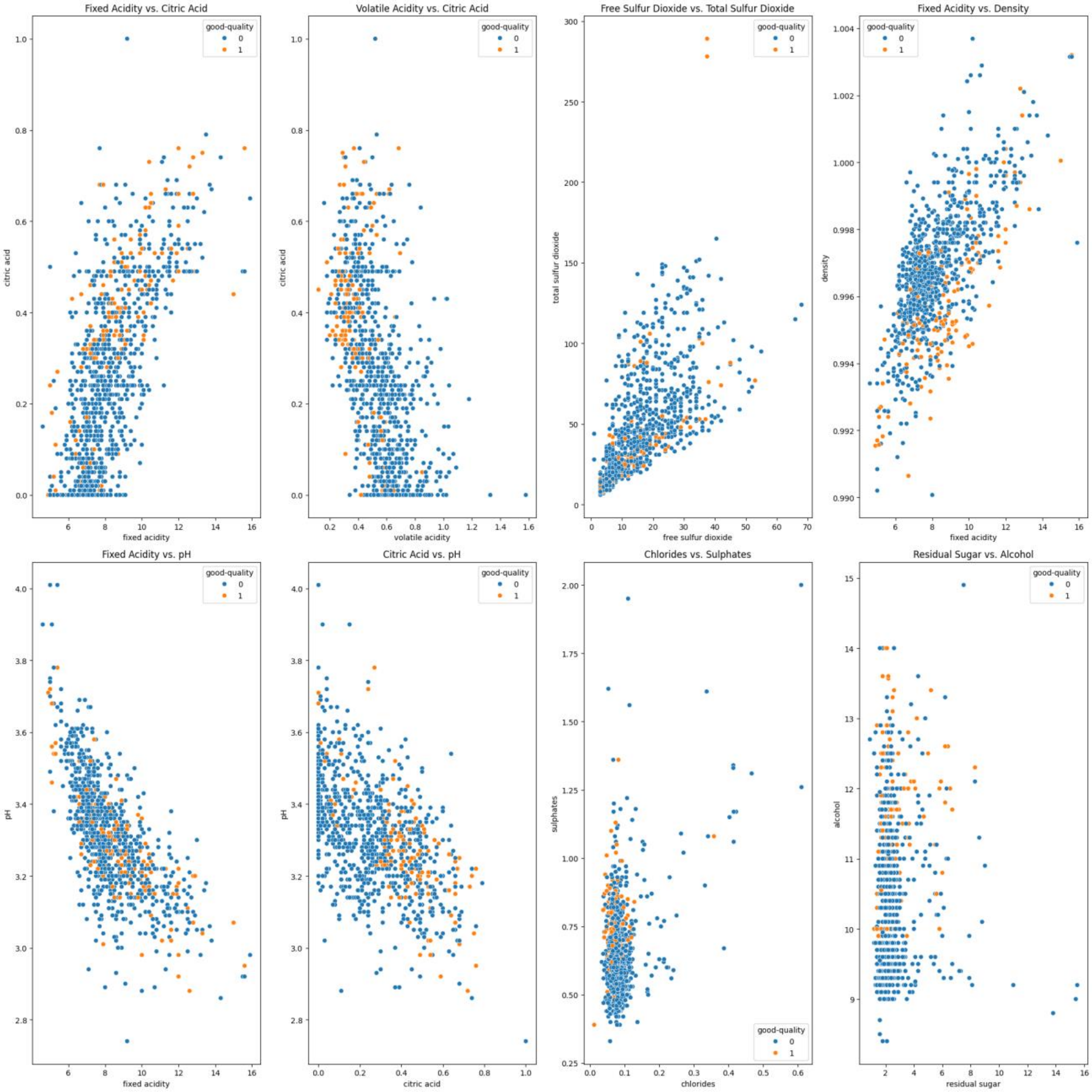
A correlation matrix was computed to examine the relationships between different features and the target variable (quality). Heatmap was used to visualize the correlation coefficients, highlighting strong positive or negative correlations between variables.



- The heatmap visualizes the strength and direction of linear relationships between pairs of variables.
- Positive correlations are indicated by warmer colors (e.g., red) for example 0.68 between 'density' and 'fixed acidity', while negative correlations are indicated by cooler colors (e.g., blue) for example -0.69 between 'pH' and 'fixed acidity'.

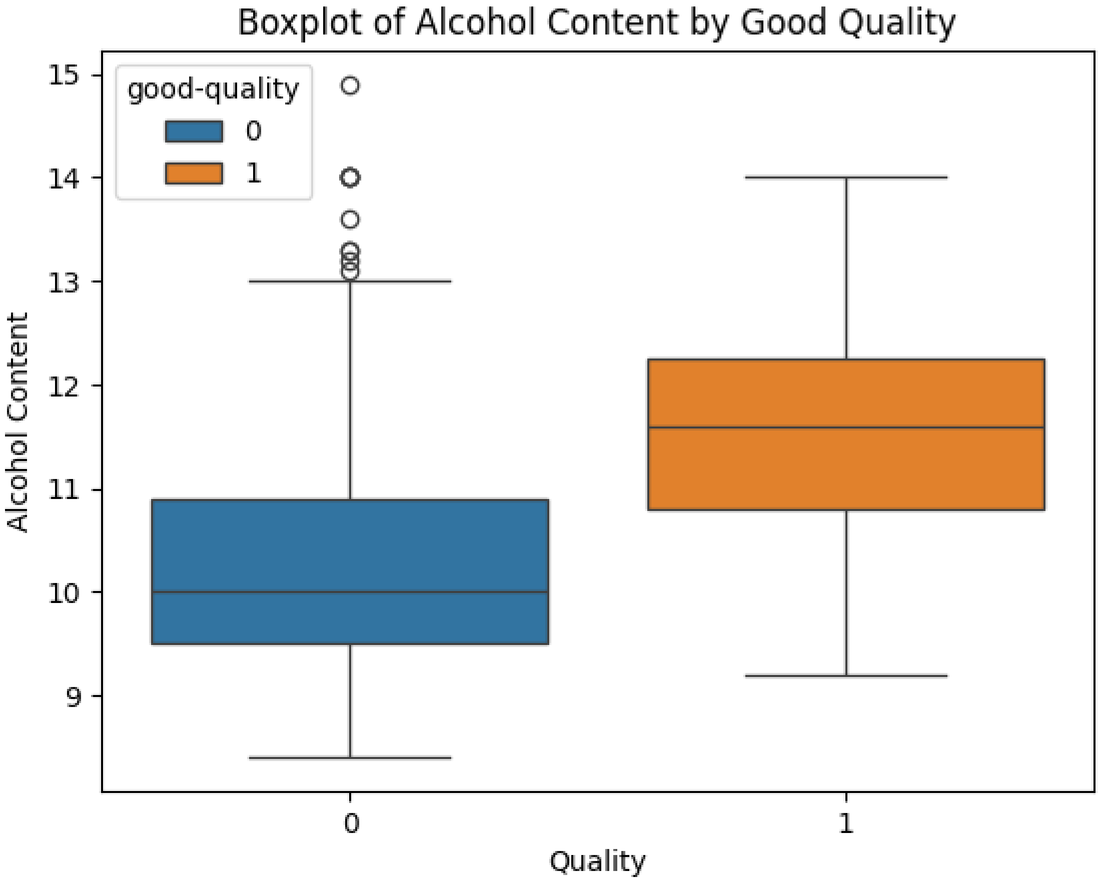
## VISUALIZATION OF RELATIONSHIPS

Scatter plots and boxplots were utilized to visualize relationships between pairs of variables. These plots helped identify potential patterns or trends in the data, such as linear or non-linear relationships between physicochemical properties.



- There is a positive linear relationship between citric acid and fixed acidity. As citric acid increases so does the fixed acidity.

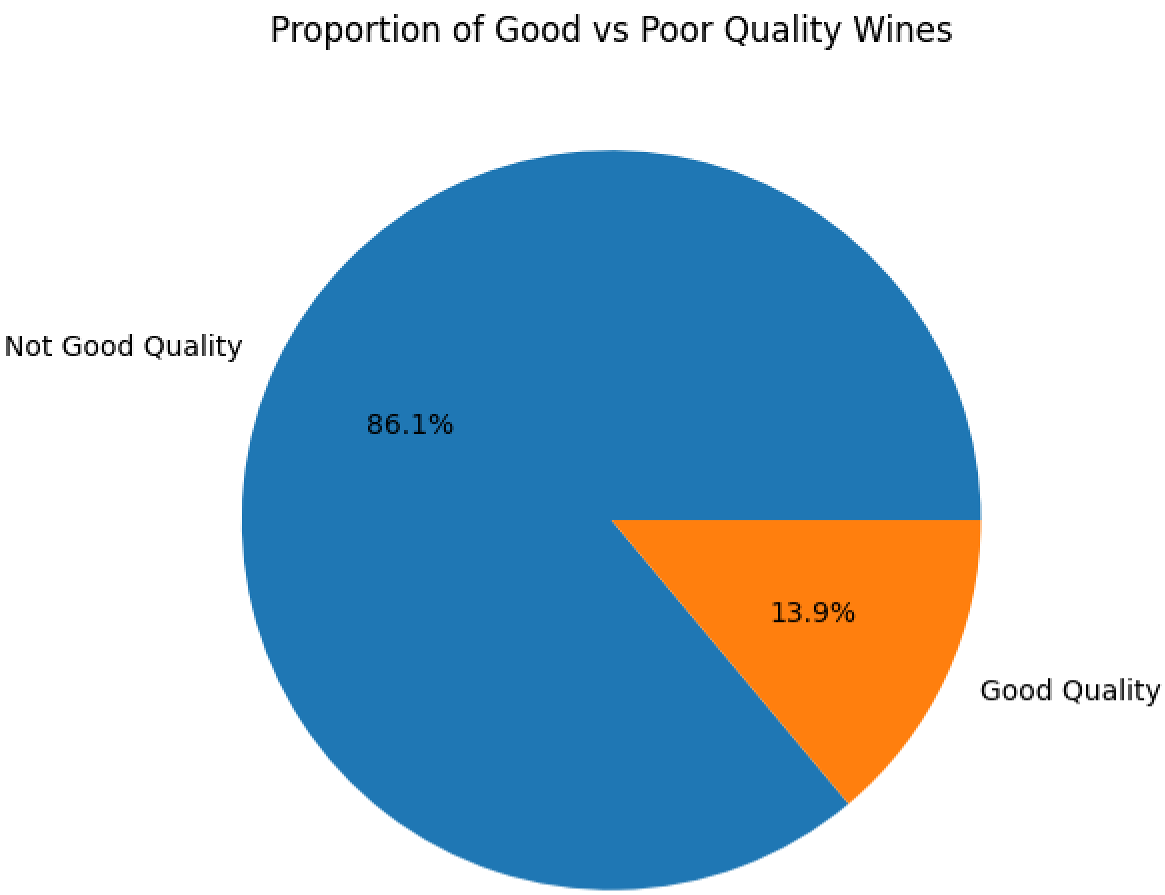
- Combinations that have citric acid show that the wine is of good quality with higher citric acid concentrations. This is indicated by the hue of the data points.
- There is negative linear relationship between pH and fixed acidity. As fixed acidity increases, the pH lowers.



- Good quality wine has higher alcohol content than poor quality wine. This means that high alcohol content results in higher quality of wine.

## PIE CHART OF PROPORTIONS

A pie chart was generated to visualize the proportion of good-quality wines versus poor-quality wines in the dataset. This chart provided a clear visual representation of the distribution of wine quality labels, indicating the balance or imbalance between the two classes.



- Most of the wine was not of good quality as indicated by the 86.1 percentage

# FEATURE ENGINEERING

## FEATURE SELECTION AND EXTRACTION

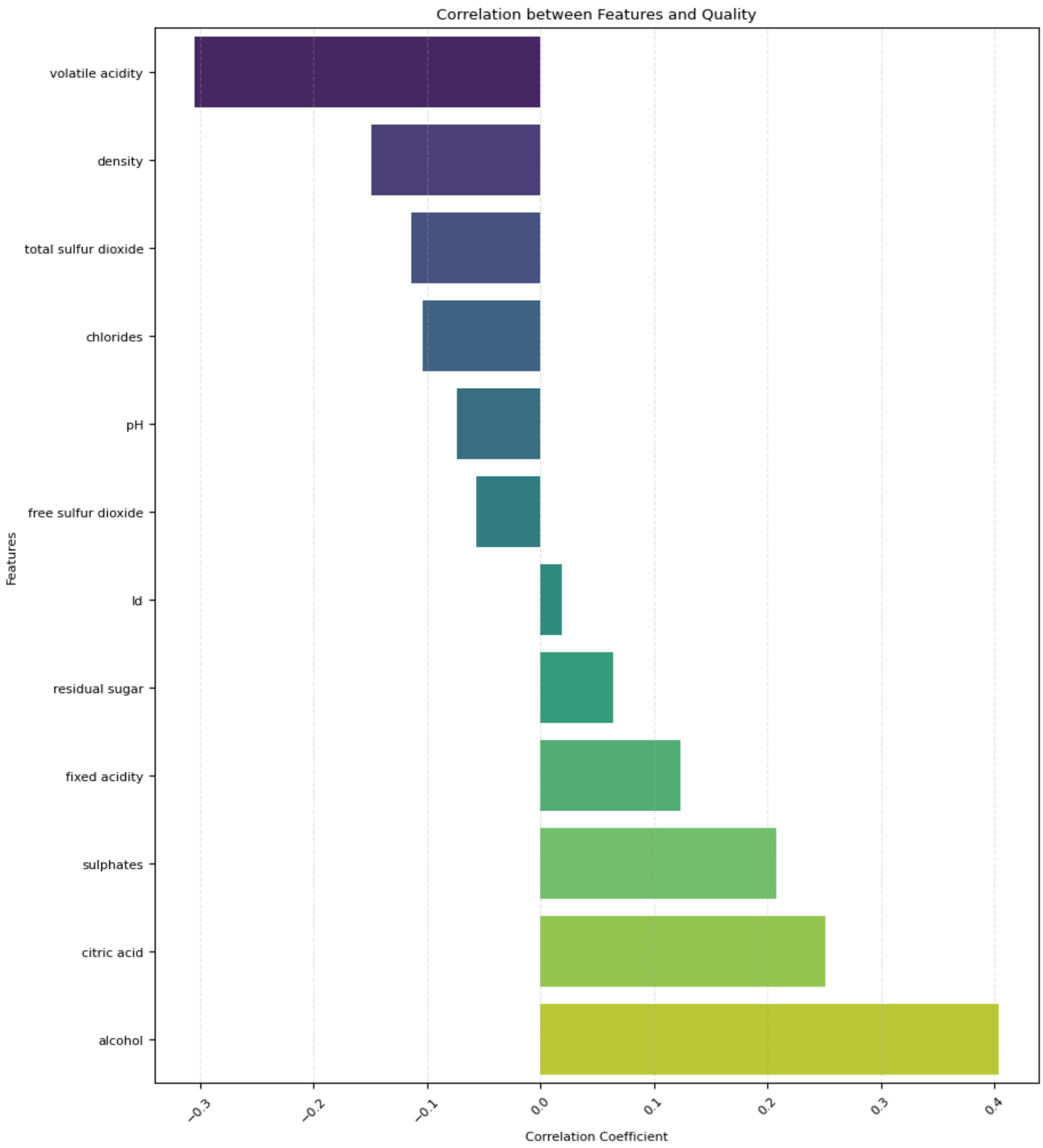
This involves identifying and choosing the most relevant features from the dataset for model training.

## FEATURE SCALING

In feature scaling, I ensured that all features have the same scale. This was achieved using the ‘**StandardScaler’** from scikit-learn, which standardizes features by removing the mean and scaling to unit variance.

## FEATURE CORRELATION WITH TARGET VARIABLE

I calculated the correlation coefficients between each feature and the target variable using Pearson correlation coefficient. This was done to identify features that are most predictive of the target variable.



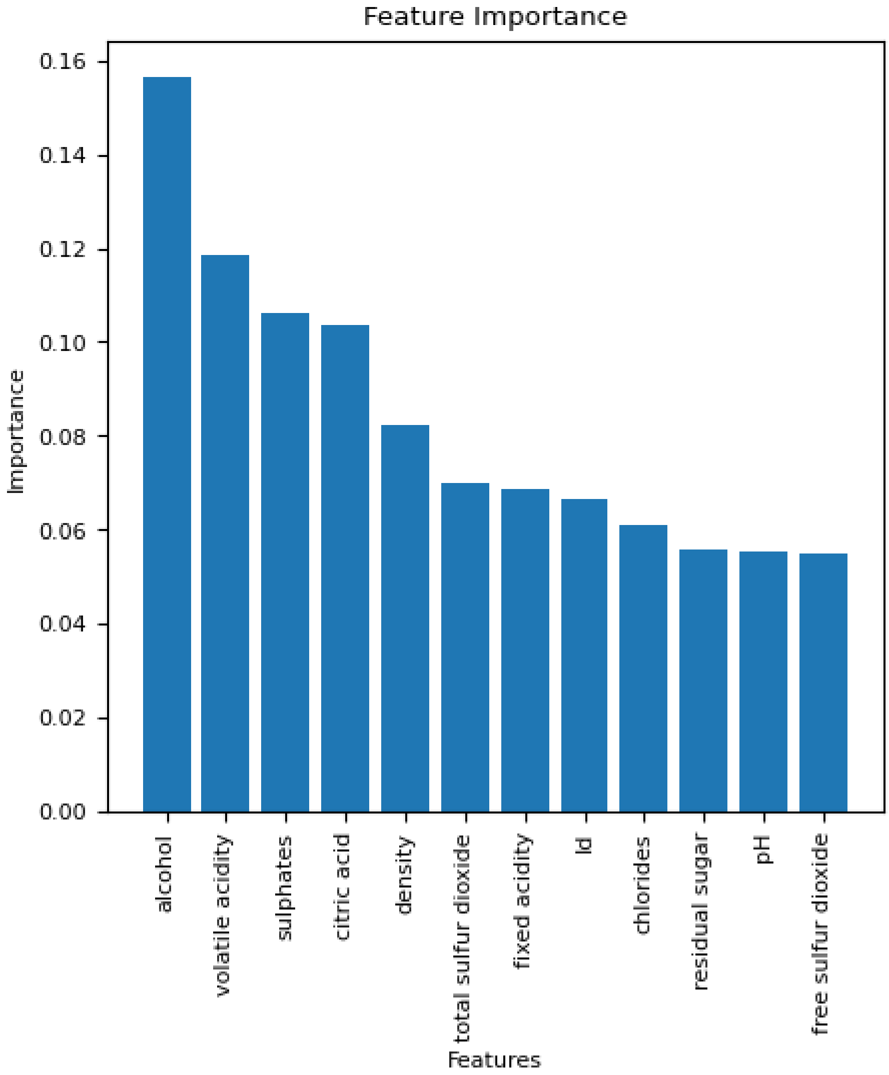
- The feature most positively correlated with 'Quality' is 'alcohol'

- The feature most negatively correlated with 'Quality' is 'volatile acidity'

## FEATURE IMPORTANCE

Feature importance was determined using models that provide feature importance scores, such as Random Forest. Random Forest model was trained and extracted the feature importance.

Feature importance helps in understanding the relative significance of different features in predicting the target variable. It provides insights into which features are most influential in the model's decision-making process.



## MODEL SELECTION

For model selection, the study considered logistic regression, random forest, gradient boosting, and support vector machine (SVM) classifiers. These models were chosen for their suitability for classification tasks, ranging from linear to ensemble and kernel-based methods. Each model offers unique strengths, such as interpretability in logistic regression, robustness in random forest, and high predictive power in gradient boosting and SVM. By evaluating multiple models, the study aimed to identify the most effective algorithm for the dataset, balancing performance and interpretability.

## MODEL TRAINING, TESTING AND EVALUATION

1. **Model Training**: Each selected model was trained using the training dataset. This involved feeding the features (X_train) and corresponding labels (y_train) into each model and allowing it to learn the underlying patterns in the data.

2. **Model Testing**: After training, the trained models were used to make predictions on the testing dataset (X_test). This step helped assess how well each model generalized to unseen data.

3. **Model Evaluation**: To evaluate the performance of each model, several metrics were calculated:

    - **Accuracy Score**: This metric measures the proportion of correctly predicted labels out of all predictions.

    - **F1-Score**: It is the harmonic mean of precision and recall, providing a balance between the two metrics.

    - **Mean Absolute Error (MAE)**: This metric calculates the average absolute differences between the predicted and actual values.

    - **Root Mean Squared Error (RMSE)**: It is the square root of the average of squared differences between predicted and actual values, providing an indication of the model's prediction error.

4. **Classification Report**: A classification report for each model was generated, which includes precision, recall, F1-score, and support for each class. This report provides a detailed summary of the model's performance across different classes.
5. **Confusion Matrix**: A confusion matrix visualizes the model's performance by comparing predicted labels against true labels. It provides insights into the model's ability to correctly classify instances into different classes.

# RESULTS

Based on the evaluation results, the random forest model emerged as the top performer, achieving the highest accuracy and F1-score among the evaluated models. Its robust performance in identifying wines of good quality makes it the preferred choice for predicting wine quality based on physicochemical attributes.

A key takeaway from this analysis accentuates the critical role of alcohol content, volatile acidity, sulphates and citric acid as important factors influencing wine quality.