**Wine Quality Prediction Model**

**Data Cleaning and Preparation:**

All missing values were for the dataset was replaced with the mean as the types was float64.

**Data Exploration and Visualisation**

The data was visualized and seen to be

Shape: (1599, 12) – 1599 instances or cases and 12 atrributes

Types: The types for all the data was float64

**Description of the Data**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Fixed Acidity | Volatile Acidity | …….. | Alcohol | Quality |
| count | 1599.00000 | 1599.00000 | …….. | 1599.000000 | 1599.000000 |
| mean | 8.319637 | 0.527821 |  | 10.422983 | 5.636023 |
| Std div. | 1.741096 | 0.179060 |  | 1.065668 | 0.807569 |
| min | 4.6000000 | 0.120000 |  | 8.400000 | 3.000000 |
| 25% | 7.100000 | 0.390000 |  | 9.500000 | 5.000000 |
| 50% | 7.900000 | 0.520000 |  | 10.200000 | 6.000000 |
| 75% | 9.200000 | 0.640000 |  | 11.100000 | 6.000000 |
| max | 15.900000 | 1.580000 |  | 14.900000 | 8.000000 |

Missing values:

fixed acidity --- 0

volatile acidity --- 0

citric acid --- 0

residual sugar --- 0

chlorides --- 0

free sulfur dioxide --- 0

total sulfur dioxide --- 0

density --- 0

pH --- 0

sulphates --- 0

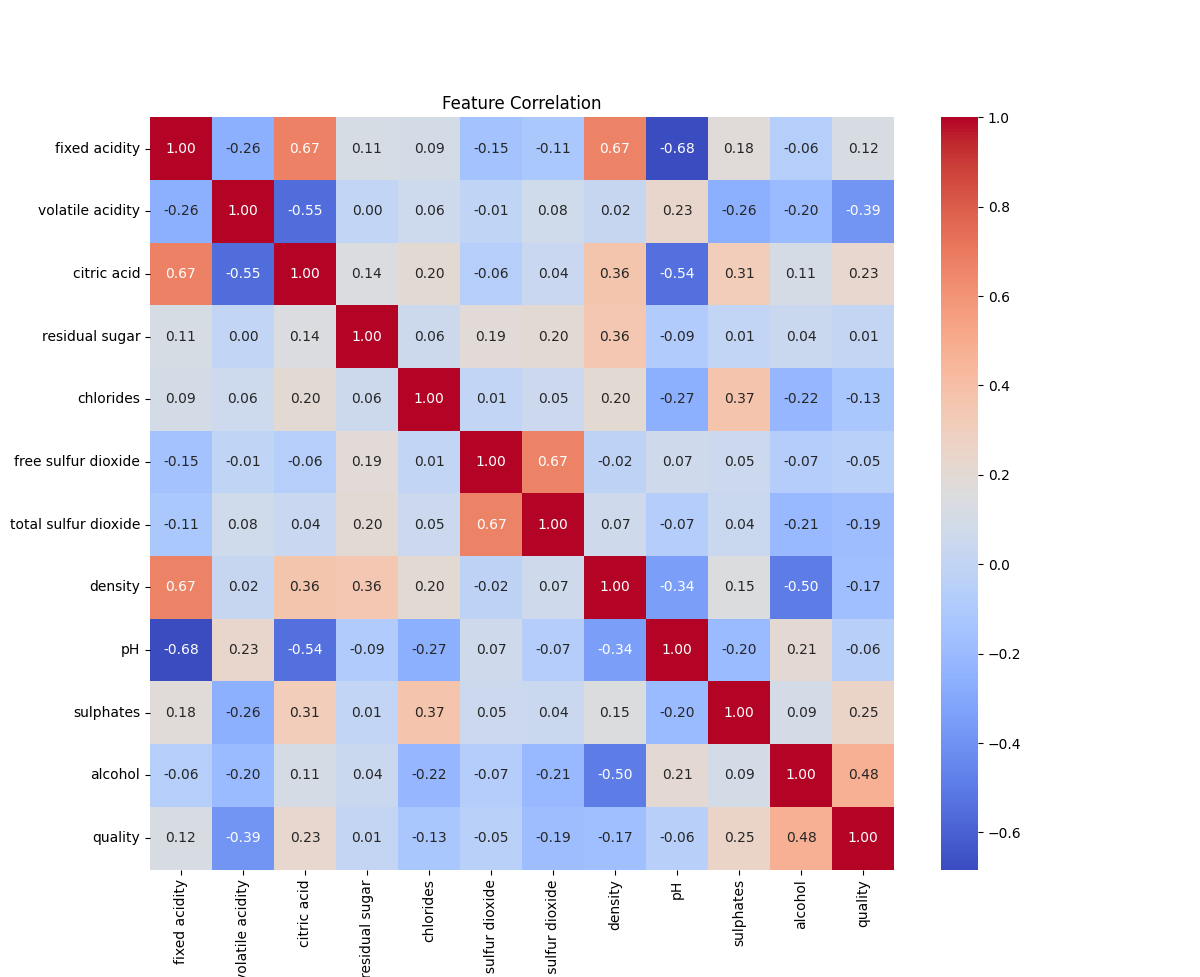
alcohol --- 0

quality --- 0

Quality Distribution

|  |  |
| --- | --- |
| Quality Distribution | Quality |
| 3 | 10 |
| 8 | 18 |
| 4 | 53 |
| 7 | 199 |
| 6 | 638 |
| 5 | 681 |

Correlation Heat Map



# Wine Quality Model Evaluation Report

## Overview

Two machine learning models, Linear Regression and Random Forest Regressor, were trained to predict wine quality (0-10 scale) using the UCI Wine Quality dataset. Performance was evaluated on a test set (~320 samples) using Mean Squared Error (MSE) and R² Score.

## Model Performance

### Linear Regression

* **MSE**: 0.39
  + Average prediction error is ~0.62 quality points (√0.39), indicating moderate accuracy.
* **R²**: 0.403
  + Explains ~40.3% of variance in quality scores, capturing linear relationships (e.g., alcohol’s impact) but missing non-linear patterns.

### Random Forest Regressor

* **MSE**: 0.30
  + Lower error (~0.55 points, √0.30), showing higher accuracy than Linear Regression.
* **R²**: 0.54
  + Explains ~53.9% of variance, better capturing non-linear relationships and feature interactions.

## Comparative Analysis

* **Accuracy**: Random Forest outperforms Linear Regression, with a ~23% lower MSE (0.30 vs. 0.39), reducing prediction error by ~0.07 points.
* **Explained Variance**: Random Forest’s R² is ~34% higher (0.54 vs. 0.403), indicating better modeling of complex patterns.
* **Why?**: Random Forest handles non-linearities and interactions (e.g., alcohol and sulphates), while Linear Regression is limited to linear assumptions.

## Conclusion

Both models perform well for the dataset, with Random Forest being superior (MSE: 0.30, R²: 0.54) due to its ability to model complex relationships. Performance is strong compared to typical results (MSE ~0.4-0.6, R² ~0.3-0.5), but the dataset’s imbalance and noise limit R² to ~0.5. Further improvements could involve hyperparameter tuning or feature engineering.