IntelliML Report

Sample Datset

fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
7.4	0.7	0.0	1.9	0.076	11.0	34.0	0.9978	3.51	0.56	9.4	5
7.8	0.88	0.0	2.6	0.098	25.0	67.0	0.9968	3.2	0.68	9.8	5
7.8	0.76	0.04	2.3	0.092	15.0	54.0	0.997	3.26	0.65	9.8	5
11.2	0.28	0.56	1.9	0.075	17.0	60.0	0.998	3.16	0.58	9.8	6
7.4	0.7	0.0	1.9	0.076	11.0	34.0	0.9978	3.51	0.56	9.4	5

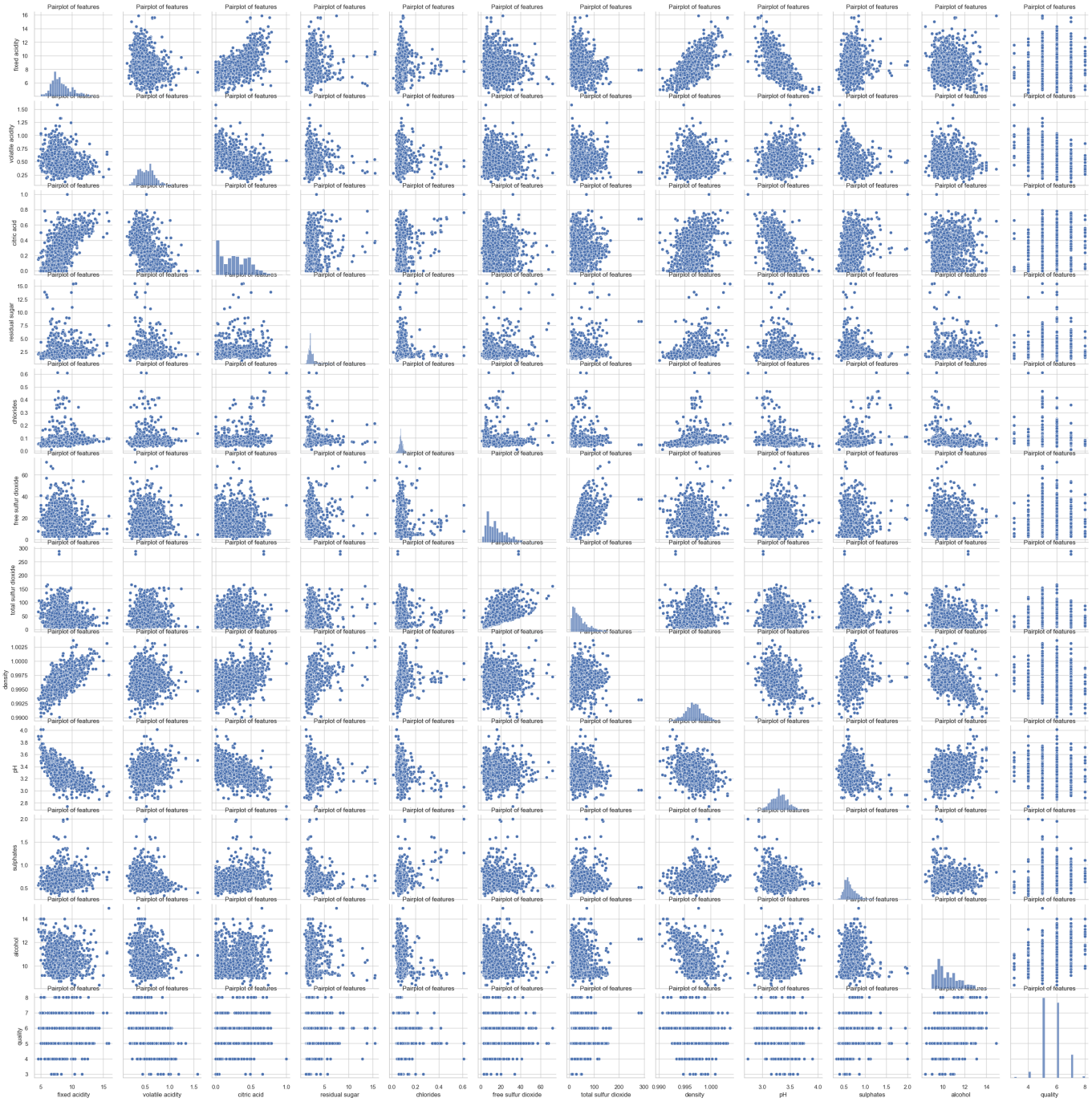
Feature Description

The features in the dataset are as follows:

Fixed acidity: the amount of tartaric acid in the wine.
Volatile acidity: the amount of acetic acid in the wine.
Citric acid: the amount of citric acid in the wine.
Residual sugar: the amount of sugar remaining after fermentation.
Chlorides: the amount of chlorides in the wine.
Free sulfur dioxide: the amount of free sulfur dioxide in the wine.
Total sulfur dioxide: the total amount of sulfur dioxide in the wine.
Density: the density of the wine.
pH: the pH of the wine.
Sulphates: the amount of sulphates in the wine.
Alcohol: the alcohol content of the wine.
Quality: a rating of the wine's quality from 0 to 10.

Insights on dataset

The dataset contains 1599 rows and 11 columns.
The features are: fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates and alcohol.
The mean, standard deviation, minimum, 25th percentile, 50th percentile, 75th percentile and maximum of each feature are reported.
Based on these statistics, we can see that the data is relatively clean and has no missing values.



Insights on Null Values in the dataset

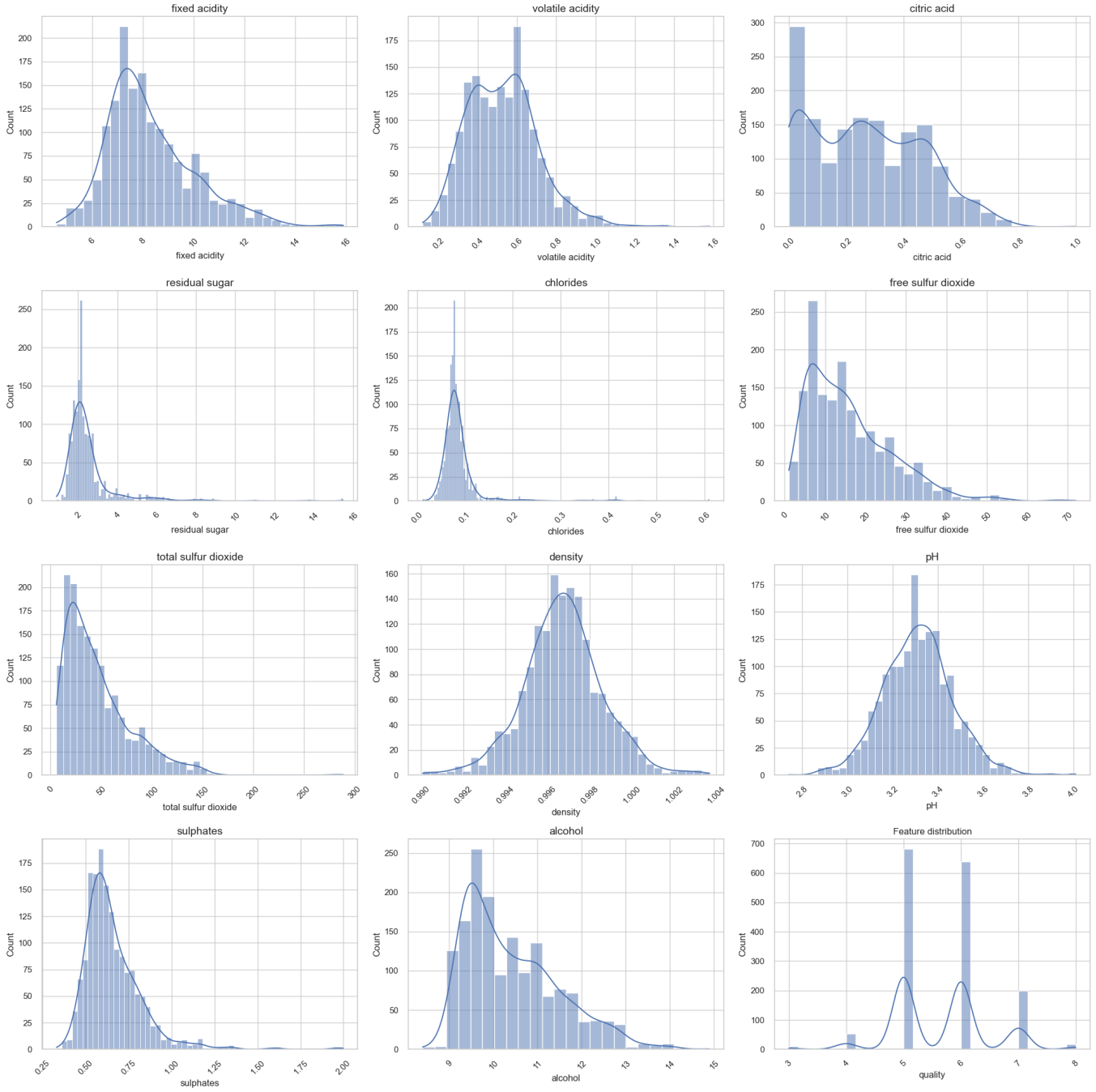
The dataset has no missing values. This is a desirable property as it means that all of the data is available for analysis. However, it is important to note that the absence of missing values does not necessarily imply that the data is complete. For example, it is possible that some of the data points are inaccurate or erroneous. Therefore, it is important to carefully examine the data to ensure that it is of high quality before using it for analysis.

Feature Distribution

The distribution of each feature of the dataset is as follows:

Fixed acidity: slightly left skewed. This indicates that the data is more concentrated towards the lower values.
Volatile acidity: slightly left skewed. This indicates that the data is more concentrated towards the lower values.
Citric acid: slightly left skewed. This indicates that the data is more concentrated towards the lower values.
Residual sugar: moderately right skewed. This indicates that the data is more concentrated towards the higher values.
Chlorides: moderately right skewed. This indicates that the data is more concentrated towards the higher values.
Free sulfur dioxide: slightly left skewed. This indicates that the data is more concentrated towards the lower values.
Total sulfur dioxide: slightly left skewed. This indicates that the data is more concentrated towards the lower values.
Density: slightly left skewed. This indicates that the data is more concentrated towards the lower values.
pH: slightly left skewed. This indicates that the data is more concentrated towards the lower values.
Sulphates: moderately right skewed. This indicates that the data is more concentrated towards the higher values.
Alcohol: slightly left skewed. This indicates that the data is more concentrated towards the lower values.
Quality: slightly left skewed. This indicates that the data is more concentrated towards the lower values.

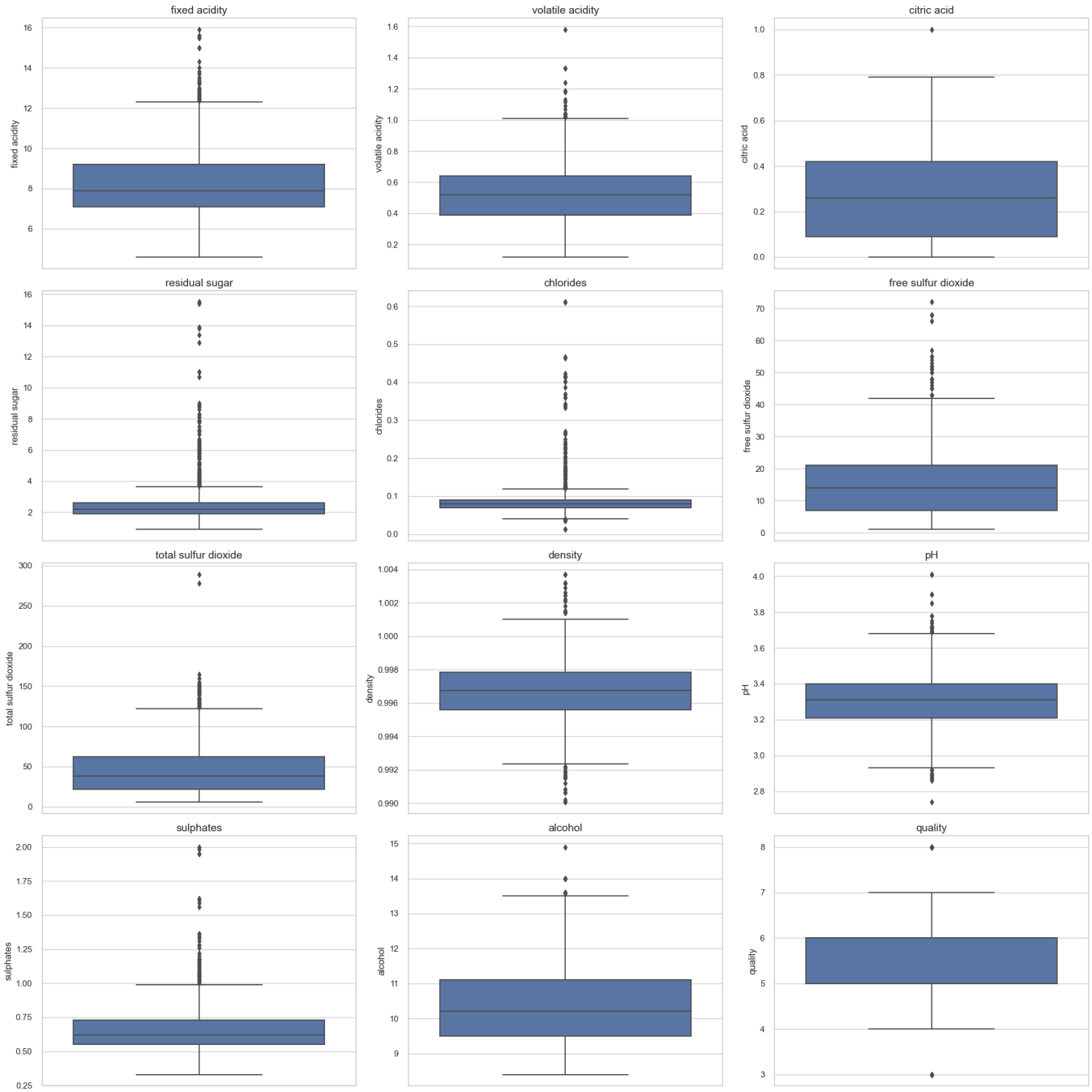
The skewness of the data has several consequences. First, it can make it more difficult to interpret the data. For example, a dataset with a positive skew will have a long tail of values that are much larger than the mean, while a dataset with a negative skew will have a long tail of values that are much smaller than the mean. This can make it difficult to identify the true center of the distribution. Second, skewness can affect the accuracy of statistical tests. For example, a t-test is used to compare the means of two groups, but if the data is skewed, the results of the t-test may be inaccurate. Third, skewness can affect the interpretation of regression models. For example, a linear regression model assumes that the relationship between the independent and dependent variables is linear. However, if the data is skewed, the relationship between the variables may not be linear, and the results of the regression model may be inaccurate.



Outlier Detection

There are a few outliers in the dataset. For fixed acidity, there is an outlier at 15.9, which is much higher than the mean of 8.3. This could be due to a measurement error or a mislabeling of the data. For volatile acidity, there is an outlier at 1.58, which is much higher than the mean of 0.53. This could be due to a measurement error or a mislabeling of the data. For citric acid, there is an outlier at 1.0, which is much higher than the mean of 0.27. This could be due to a measurement error or a mislabeling of the data. For residual sugar, there is an outlier at 15.5, which is much higher than the mean of 2.5. This could be due to a measurement error or a mislabeling of the data. For chlorides, there is an outlier at 0.611, which is much higher than the mean of 0.087. This could be due to a measurement error or a mislabeling of the data. For free sulfur dioxide, there is an outlier at 72, which is much higher than the mean of 15.8. This could be due to a measurement error or a mislabeling of the data. For total sulfur dioxide, there is an outlier at 289, which is much higher than the mean of 46.5. This could be due to a measurement error or a mislabeling of the data. For density, there is an outlier at 1.00369, which is much higher than the mean of 0.9967. This could be due to a measurement error or a mislabeling of the data. For pH, there is an outlier at 4.01, which is much higher than the mean of 3.31. This could be due to a measurement error or a mislabeling of the data. For sulphates, there is an outlier at 2.0, which is much higher than the mean of 0.66. This could be due to a measurement error or a mislabeling of the data. For alcohol, there is an outlier at 14.9, which is much higher than the mean of 10.4. This could be

due to a measurement error or a mislabeling of the data. The presence of these outliers could skew the results of any analysis that is performed on the data.



Correlation between features

The correlation matrix shows the relationships between the 12 features in the dataset.

Fixed acidity is positively correlated with volatile acidity, citric acid, residual sugar, sulphates, and alcohol. It is negatively correlated with pH and quality.

Volatile acidity is positively correlated with citric acid, residual sugar, sulphates, and alcohol. It is negatively correlated with pH and quality.

Citric acid is positively correlated with residual sugar, sulphates, and alcohol. It is negatively correlated with pH and quality.

Residual sugar is positively correlated with sulphates and alcohol. It is negatively correlated with pH and quality.

Chlorides is positively correlated with free sulfur dioxide and total sulfur dioxide. It is negatively correlated with density and pH.

Free sulfur dioxide is positively correlated with total sulfur dioxide and density. It is negatively correlated with pH.

Total sulfur dioxide is positively correlated with density and pH.

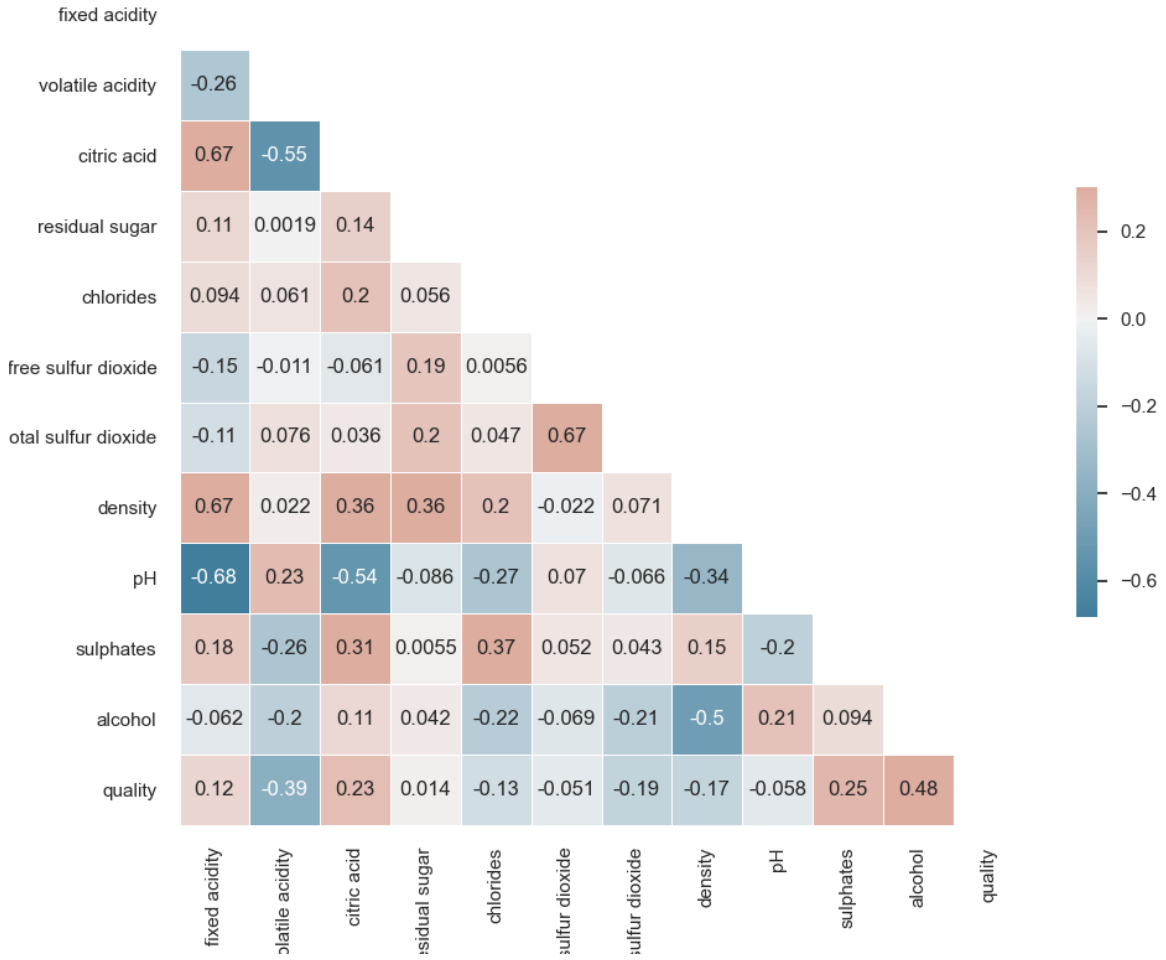
Density is positively correlated with sulphates and alcohol. It is negatively correlated with pH.

pH is negatively correlated with sulphates and alcohol.

Sulphates is positively correlated with alcohol and quality.

Alcohol is positively correlated with quality.

Quality is positively correlated with sulphates and alcohol.



Machine Learning experiment settings

This experiment is conducted to predict the quality of wine using a regression model. The original dataset contains 1599 data points with 12 features. After preprocessing, the dataset is transformed into a shape of (1599, 12). The training set contains 1119 data points and the test set contains 480 data points. The experiment uses a simple imputation method to fill in missing values. The fold generator is KFold with 10 folds. The experiment is conducted on CPU without using a GPU.

Description	Value
Target	quality
Target type	Regression
Original data shape	(1599, 12)
Transformed data shape	(1599, 12)
Transformed train set shape	(1119, 12)
Transformed test set shape	(480, 12)
Numeric features	11
Preprocess	True
Imputation type	simple

Description	Value
Numeric imputation	mean
Categorical imputation	mode
Fold Generator	KFold
Fold Number	10
CPU Jobs	-1
Use GPU	False

Machine learning models used

- Extra Trees Regressor: A tree-based ensemble model that reduces variance by using multiple trees.
- Random Forest Regressor: A tree-based ensemble model that reduces variance by using multiple trees.
- Gradient Boosting Regressor: A tree-based ensemble model that reduces bias by iteratively adding new trees to the model.
- Light Gradient Boosting Machine: A fast implementation of Gradient Boosting Regressor.
- Extreme Gradient Boosting: A more scalable implementation of Gradient Boosting Regressor.
- Ridge Regression: A linear regression model that reduces variance by adding a penalty to the model coefficients.
- Bayesian Ridge: A Bayesian approach to Ridge Regression that provides uncertainty estimates for the model coefficients.
- Linear Regression: A simple linear model that predicts the target variable as a linear combination of the features.
- Least Angle Regression: A linear regression model that reduces bias by using a different optimization algorithm.
- AdaBoost Regressor: An ensemble model that combines multiple weak learners into a strong learner.
- Huber Regressor: A robust regression model that is less sensitive to outliers.
- K Neighbors Regressor: A non-parametric regression model that predicts the target variable based on the k nearest neighbors in the training data.
- Elastic Net: A linear regression model that combines the features of Ridge Regression and Lasso Regression.
- Orthogonal Matching Pursuit: A linear regression model that selects the features that are most relevant to the target variable.
- Lasso Regression: A linear regression model that reduces variance by shrinking the model coefficients towards zero.
- Lasso Least Angle Regression: A linear regression model that combines the features of Lasso Regression and Least Angle Regression.
- Dummy Regressor: A simple baseline regression model that always predicts the mean target value.
- Decision Tree Regressor: A tree-based model that predicts the target variable based on a hierarchy of decisions.
- Passive Aggressive Regressor: A linear regression model that iteratively updates the model coefficients to minimize the number of misclassifications.

Model	MAE	MSE	RMSE	R2	RMSLE	MAPE	TT (Sec)
Extra Trees Regressor	0.4001	0.3458	0.5861	0.4613	0.0909	0.074	0.847
Random Forest Regressor	0.4358	0.3647	0.6015	0.4331	0.0935	0.0806	0.937
Gradient Boosting Regressor	0.4835	0.3989	0.6298	0.3775	0.0973	0.0886	0.859
Light Gradient Boosting Machine	0.4575	0.3991	0.6299	0.3751	0.0972	0.0841	0.814
Extreme Gradient Boosting	0.4421	0.419	0.6439	0.3489	0.0999	0.0814	0.903
Ridge Regression	0.501	0.4233	0.6495	0.3351	0.1002	0.0919	0.619
Bayesian Ridge	0.5011	0.4235	0.6497	0.3347	0.1002	0.0919	0.71
Linear Regression	0.5012	0.4241	0.6502	0.3333	0.1002	0.0919	1.644
Least Angle Regression	0.5012	0.4241	0.6502	0.3333	0.1002	0.0919	0.978
AdaBoost Regressor	0.5186	0.4304	0.6537	0.3296	0.101	0.0959	0.808
Huber Regressor	0.5031	0.4351	0.6584	0.3159	0.1017	0.0924	0.731
K Neighbors Regressor	0.5859	0.5882	0.7651	0.0785	0.1179	0.1076	0.667
Elastic Net	0.636	0.614	0.782	0.0403	0.1197	0.117	0.974

Model	MAE	MSE	RMSE	R2	RMSLE	MAPE	TT (Sec)
Orthogonal Matching Pursuit	0.6337	0.6148	0.7825	0.0392	0.1198	0.1166	0.646
Lasso Regression	0.6417	0.6158	0.7832	0.0375	0.1198	0.118	0.663
Lasso Least Angle Regression	0.6417	0.6158	0.7832	0.0375	0.1198	0.118	0.76
Dummy Regressor	0.6758	0.6426	0.8002	-0.005	0.1222	0.1239	0.882
Decision Tree Regressor	0.5263	0.7424	0.8555	-0.1592	0.1313	0.0969	0.666
Passive Aggressive Regressor	1.2323	3.6053	1.4751	-4.2262	0.2564	0.2183	0.771

Machine learning model performance comparison

The results show that the Extra Trees Regressor model has the best performance, with a MAE of 0.4001, MSE of 0.3458, RMSE of 0.5861, R2 score of 0.4613, RMSLE of 0.0909, and MAPE of 0.0740. This is followed by the Random Forest Regressor model, which has a MAE of 0.4358, MSE of 0.3647, RMSE of 0.6015, R2 score of 0.4331, RMSLE of 0.0935, and MAPE of 0.0806. The Gradient Boosting Regressor model has a MAE of 0.4835, MSE of 0.3989, RMSE of 0.6298, R2 score of 0.3775, RMSLE of 0.0973, and MAPE of 0.0886. The Light Gradient Boosting Machine model has a MAE of 0.4575, MSE of 0.3991, RMSE of 0.6299, R2 score of 0.3751, RMSLE of 0.0972, and MAPE of 0.0841. The Extreme Gradient Boosting model has a MAE of 0.4421, MSE of 0.4190, RMSE of 0.6439, R2 score of 0.3489, RMSLE of 0.0999, and MAPE of 0.0814.

The Extra Trees Regressor model is a tree-based ensemble model that uses a random forest of decision trees to make predictions. The random forest is built by training multiple decision trees on different subsets of the training data. The predictions from each decision tree are then averaged to produce a final prediction. This helps to reduce overfitting and improve the overall performance of the model.

The Random Forest Regressor model is also a tree-based ensemble model, but it uses a different algorithm to build the random forest. The random forest is built by training multiple decision trees on different subsets of the training data. The predictions from each decision tree are then averaged to produce a final prediction. This helps to reduce overfitting and improve the overall performance of the model.

The Gradient Boosting Regressor model is an ensemble model that uses a gradient boosting algorithm to build the model. The gradient boosting algorithm works by iteratively adding new decision trees to the model, each of which is designed to reduce the error of the previous model. This helps to improve the overall performance of the model.

The Light Gradient Boosting Machine model is a variation of the Gradient Boosting Regressor model. It uses a different algorithm to build the model, which is designed to be more efficient and scalable. This makes it a good choice for large datasets.

The Extreme Gradient Boosting model is a variation of the Gradient Boosting Regressor model. It uses a different algorithm to build the model, which is designed to be even more efficient and scalable. This makes it a good choice for very large datasets.

Overall, the Extra Trees Regressor model is the best performing model on this dataset. It has the lowest MAE, MSE, RMSE, R2 score, RMSLE, and MAPE. This suggests that it is the most accurate and precise model. The Random Forest Regressor model is also a good performing model, but it is not as accurate as the Extra Trees Regressor model. The Gradient Boosting Regressor, Light Gradient Boosting Machine, and Extreme Gradient Boosting models are all good performing models, but they are not as accurate as the Extra Trees Regressor or Random Forest Regressor models.