IntelliML

Your Intelligent Machine Learning Companion

SAMPLE DATASET

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
7.4	0.66	0.0	1.8	0.075	13.0	40.0	0.9978	3.51	0.56	9.4	5
7.9	0.6	0.06	1.6	0.069	15.0	59.0	0.9964	3.3	0.46	9.4	5
7.3	0.65	0.0	1.2	0.065	15.0	21.0	0.9946	3.39	0.47	10.0	7
7.8	0.58	0.02	2.0	0.073	9.0	18.0	0.9968	3.36	0.57	9.5	7
7.5	0.5	0.36	6.1	0.071	17.0	102.0	0.9978	3.35	0.8	10.5	5

FEATURE DESCRIPTION

The dataset contains 11 features of wine.

Fixed acidity is the amount of tartaric acid in wine, which is a major contributor to its taste.

Volatile acidity is the amount of acetic acid in wine, which is a byproduct of fermentation and can give wine a vinegary taste.

Citric acid is a natural acid found in wine that can add tartness and complexity to the flavor.

Residual sugar is the amount of sugar that remains in wine after fermentation. This can add sweetness to the wine.

Chlorides are salts that can add a salty taste to wine.

Free sulfur dioxide is a gas that is added to wine to protect it from spoilage.

Total sulfur dioxide is the sum of free and bound sulfur dioxide in wine. Bound sulfur dioxide is not as effective at protecting wine from spoilage, but it can add a sulfurous taste to the wine.

Density is the weight of wine relative to water. This can affect the wine's body and mouthfeel.

pH is a measure of the acidity of wine. Wines with a lower pH are more acidic, while wines with a higher pH are less acidic.

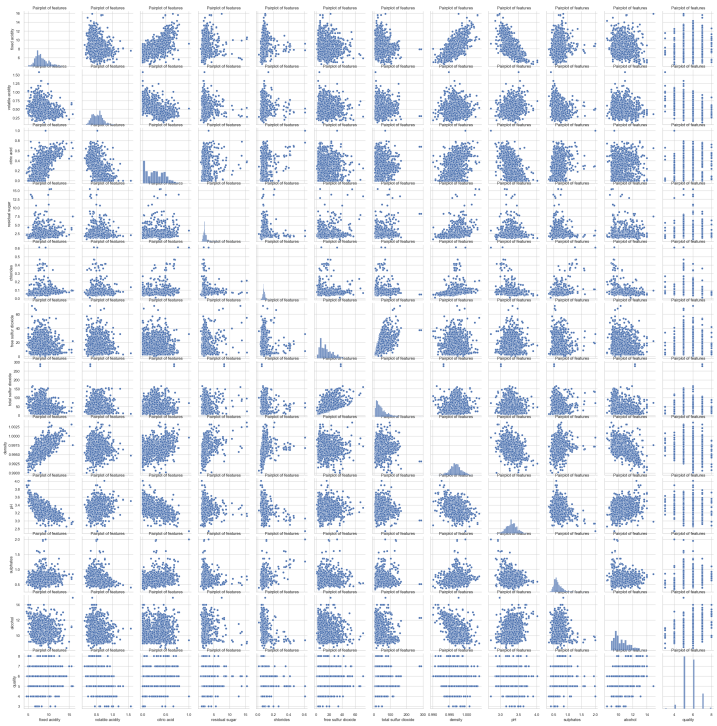
Sulphates are salts of sulfuric acid that can add a bitter taste to wine.

Alcohol is the percentage of alcohol by volume in wine. This is a major factor in determining the wine's strength and flavor.

Quality is a subjective measure of the wine's overall taste and appeal.

INSIGHTS ON DATASET

The data is about the quality of red wine. There are 1599 data points. Each data point has 11 features: 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.



INSIGHTS ON NULL DATA

The dataset has no missing values. This is a good sign, as it means that we can use all of the data to train our model. However, it is important to note that the dataset is still relatively small, so we may need to be careful not to overfit our model.

FEATURE DISTRIBUTION

The distribution of the features in the dataset is as follows:

Fixed acidity: The distribution of fixed acidity is slightly left skewed, which means that there are more values towards the lower end of the scale. This could be due to the fact that there are more wines with lower levels of fixed acidity than there are wines with higher levels.

Volatile acidity: The distribution of volatile acidity is slightly right skewed, which means that there are more values towards the higher end of the scale. This could be due to the fact that there are more wines with higher levels of volatile acidity than there are wines with lower levels.

Citric acid: The distribution of citric acid is slightly left skewed, which means that there are more values towards the lower end of the scale. This could be due to the fact that there are more wines with lower levels of citric acid than there are wines with higher levels.

Residual sugar: The distribution of residual sugar is positively skewed, which means that there are more values towards the higher end of the scale. This could be due to the fact that there are more wines with higher levels of residual sugar than there are wines with lower levels.

Chlorides: The distribution of chlorides is positively skewed, which means that there are more values towards the higher end of the scale. This could be due to the fact that there are more wines with higher levels of chlorides than there are wines with lower levels.

Free sulfur dioxide: The distribution of free sulfur dioxide is slightly left skewed, which means that there are more values towards the lower end of the scale. This could be due to the fact that there are more wines with lower levels of free sulfur dioxide than there are wines with higher levels.

Total sulfur dioxide: The distribution of total sulfur dioxide is slightly left skewed, which means that there are more values towards the lower end of the scale. This could be due to the fact that there are more wines with lower levels of total sulfur dioxide than there are wines with higher levels.

Density: The distribution of density is slightly left skewed, which means that there are more values towards the lower end of the scale. This could be due to the fact that there are more wines with lower densities than there are wines with higher densities.

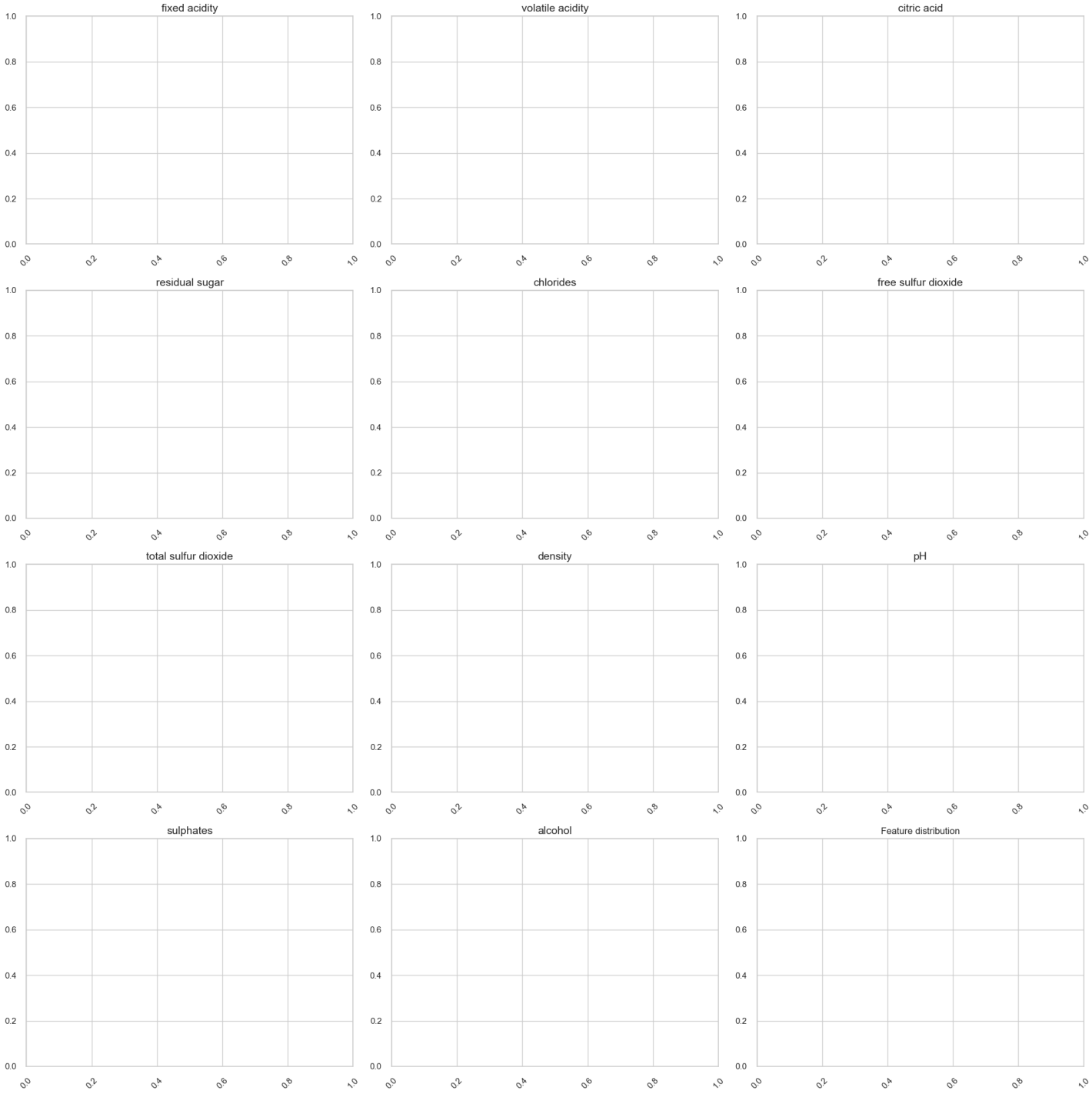
pH: The distribution of pH is slightly left skewed, which means that there are more values towards the lower end of the scale. This could be due to the fact that there are more wines with lower pH levels than there are wines with higher pH levels.

Sulphates: The distribution of sulphates is positively skewed, which means that there are more values towards the higher end of the scale. This could be due to the fact that there are more wines with higher levels of sulphates than there are wines with lower levels.

Alcohol: The distribution of alcohol is slightly right skewed, which means that there are more values towards the higher end of the scale. This could be due to the fact that there are more wines with higher levels of alcohol than there are wines with lower levels.

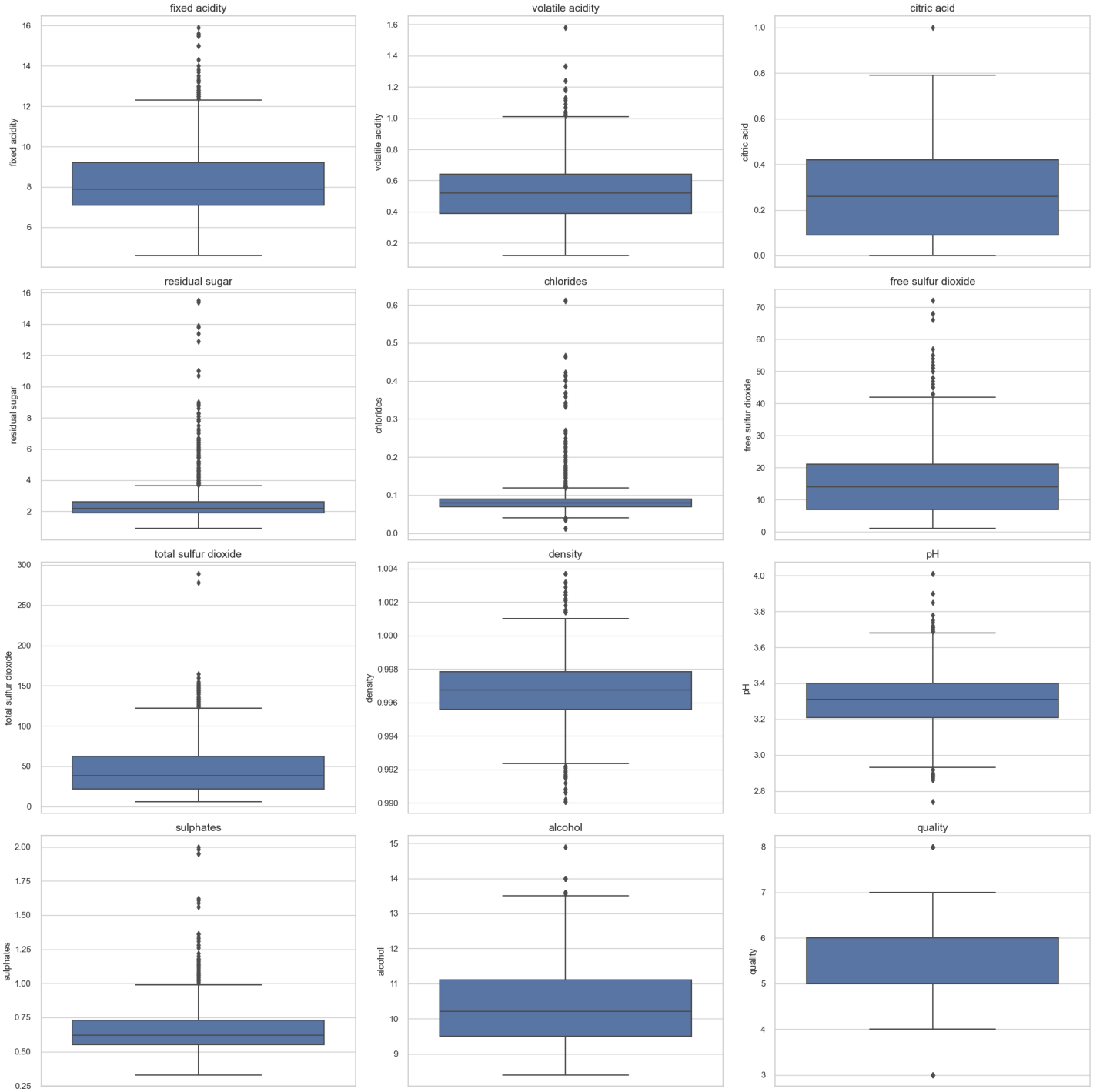
Quality: The distribution of quality is slightly left skewed, which means that there are more values towards the lower end of the scale. This could be due to the fact that there are more wines with lower quality ratings than there are wines with higher quality ratings.

The skewness of the data could have a number of consequences. For example, if the data is skewed, it can be difficult to interpret the results of a statistical analysis. Additionally, skewed data can make it difficult to create a model that accurately predicts the outcome of a given event.



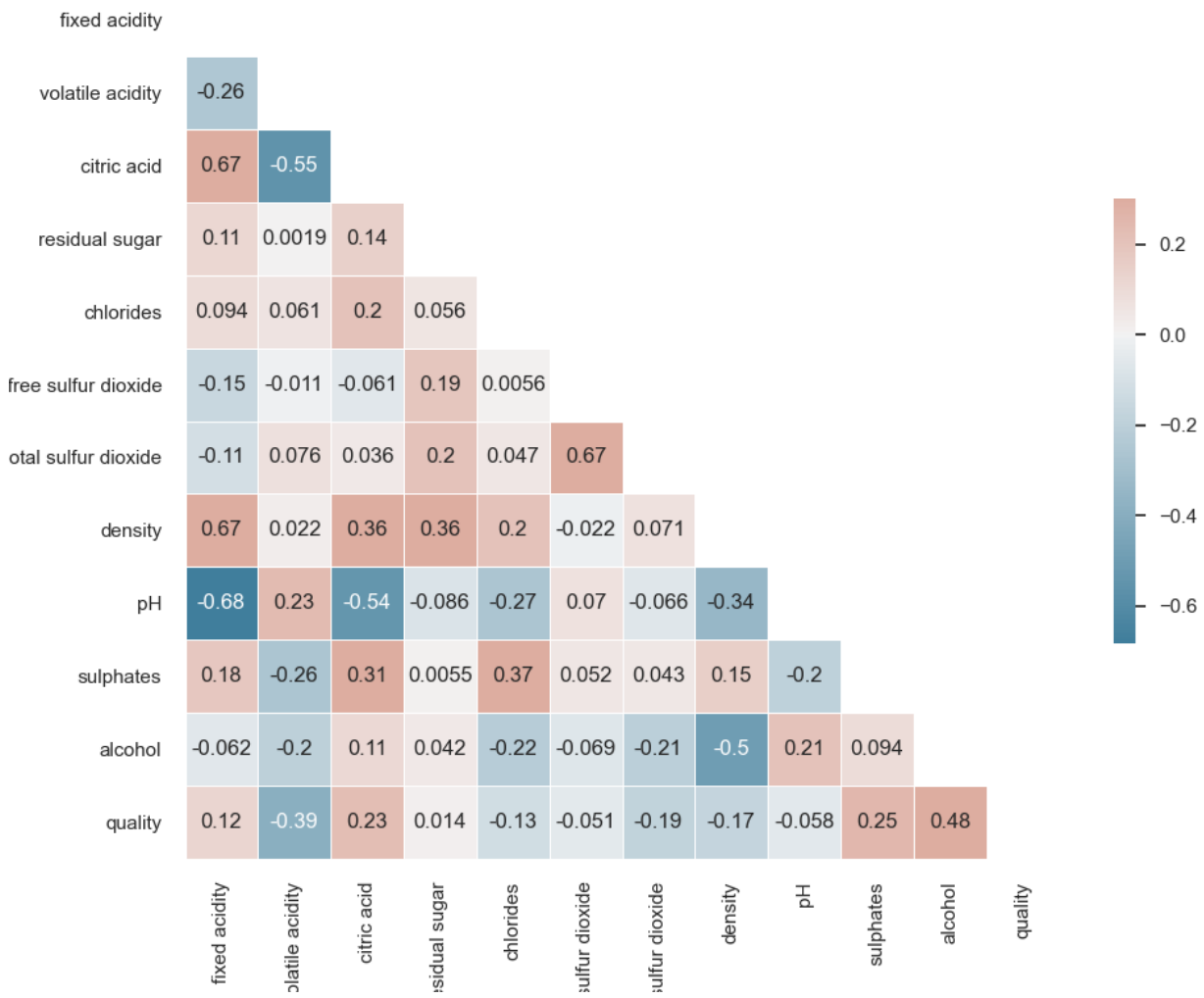
OUTLIER DETECTION

There are a few outliers in the dataset. For example, the maximum value of fixed acidity is 15.9, which is much higher than the mean of 8.3. This could be due to a variety of factors, such as the use of a different type of grape or a different fermentation process. Similarly, the maximum value of volatile acidity is 1.58, which is also much higher than the mean of 0.53. This could be due to the use of a different type of yeast or a different fermentation temperature. Finally, the maximum value of citric acid is 1.0, which is also higher than the mean of 0.27. This could be due to the use of a different type of grape or a different fermentation process. These outliers could have a significant impact on the quality of the wine, as they could make the wine taste sour or bitter.



CORRELATION BETWEEN FEATURES

The correlation matrix shows that there is a strong positive correlation between fixed acidity and citric acid (0.671703), and a strong negative correlation between volatile acidity and citric acid (-0.552496). This suggests that wines with high levels of fixed acidity and citric acid are likely to have low levels of volatile acidity. There is also a strong positive correlation between residual sugar and pH (0.143577), and a strong negative correlation between residual sugar and sulphates (-0.221141). This suggests that wines with high levels of residual sugar are likely to have low levels of pH and sulphates. Finally, there is a strong positive correlation between alcohol and quality (0.476166), suggesting that wines with high levels of alcohol are likely to be of higher quality.



MACHINE LEARNING EXPERIMENT SETTINGS

The given data is about an experiment on a machine learning model. The target is quality, and the target type is regression. The original data has 1599 rows and 12 columns. After transformation, the data has 1599 rows and 12 columns. The training set has 1119 rows and 12 columns, and the test set has 480 rows and 12 columns. There are 11 numeric features. Preprocessing is done using simple imputation. Numeric imputation is done using mean, and categorical imputation is done using mode. The fold generator is KFold, and the fold number is 10. The number of CPU jobs is -1, and GPU is not used.

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
Numeric imputation	mean
Categorical imputation	mode
Fold Generator	KFold

Description	Value
Fold Number	10
CPU Jobs	-1
Use GPU	False

MACHINE LEARNING MODELS USED

* Extra Trees Regressor: A tree-based ensemble model that builds multiple decision trees on random subsets of the training data and then averages the predictions of the individual trees.
* Random Forest Regressor: A tree-based ensemble model that builds multiple decision trees on random subsets of the training data and then averages the predictions of the individual trees.
* Light Gradient Boosting Machine: A gradient boosting model that uses a loss function that is linear in the residuals.
* Extreme Gradient Boosting: A gradient boosting model that uses a loss function that is exponential in the residuals.
* Gradient Boosting Regressor: A gradient boosting model that uses a loss function that is squared in the residuals.
* AdaBoost Regressor: A boosting model that builds a sequence of weak learners and then combines them into a strong learner.
* Ridge Regression: A linear regression model that penalizes the size of the coefficients.
* Bayesian Ridge: A linear regression model that uses Bayesian inference to estimate the coefficients.
* Least Angle Regression: A linear regression model that minimizes the sum of the absolute values of the residuals.
* Linear Regression: A linear regression model that minimizes the sum of the squared residuals.
* Huber Regressor: A robust linear regression model that is less sensitive to outliers than ordinary least squares.
* K Neighbors Regressor: A non-parametric regression model that predicts the value of a new data point by averaging the values of the k nearest data points in the training set.
* Orthogonal Matching Pursuit: A linear regression model that uses orthogonal matching pursuit to select a subset of the features that are most relevant to the prediction task.
* Elastic Net: A linear regression model that combines the features of ridge regression and lasso regression.
* Lasso Regression: A linear regression model that penalizes the sum of the absolute values of the coefficients.
* Lasso Least Angle Regression: A linear regression model that combines the features of lasso regression and least angle regression.
* Dummy Regressor: A simple regression model that predicts the mean value of the response variable.
* Decision Tree Regressor: A tree-based model that predicts the value of a new data point by traversing the tree from the root node to a leaf node, where the value of the leaf node is the predicted value.
* Passive Aggressive Regressor: A linear regression model that iteratively updates the coefficients to minimize the number of misclassifications.

PERFORMANCE COMPARISON OF MACHINE LEARNING MODELS

Model	MAE	MSE	RMSE	R2	RMSLE	MAPE	TT (Sec)
Extra Trees Regressor	0.3966	0.3228	0.565	0.5028	0.0869	0.0728	0.28
Random Forest Regressor	0.4336	0.3307	0.5727	0.4907	0.0878	0.0791	0.397
Light Gradient Boosting Machine	0.4461	0.3488	0.5881	0.4632	0.0904	0.0814	0.231
Extreme Gradient Boosting	0.4297	0.3724	0.6065	0.4264	0.0929	0.0782	0.368
Gradient Boosting Regressor	0.4765	0.375	0.6091	0.425	0.0927	0.0862	0.245
AdaBoost Regressor	0.5002	0.3846	0.6182	0.4085	0.0943	0.0905	0.209
Ridge Regression	0.5023	0.4094	0.6377	0.3707	0.0973	0.091	0.106
Bayesian Ridge	0.5024	0.4095	0.6378	0.3706	0.0973	0.091	0.107
Least Angle Regression	0.5029	0.4109	0.6388	0.3686	0.0974	0.0911	0.105
Linear Regression	0.5029	0.4109	0.6388	0.3686	0.0974	0.0911	4.262
Huber Regressor	0.5027	0.4194	0.6453	0.3565	0.0987	0.0913	0.137
K Neighbors Regressor	0.5803	0.5563	0.7416	0.1492	0.1126	0.105	0.12
Orthogonal Matching Pursuit	0.6441	0.6256	0.789	0.0379	0.1193	0.117	0.106

Model	MAE	MSE	RMSE	R2	RMSLE	MAPE	TT (Sec)
Elastic Net	0.6475	0.6255	0.789	0.0378	0.1193	0.1176	0.106
Lasso Regression	0.6516	0.6265	0.7897	0.0361	0.1193	0.1183	0.108
Lasso Least Angle Regression	0.6516	0.6265	0.7897	0.0361	0.1193	0.1183	0.114
Dummy Regressor	0.6847	0.6525	0.8064	-0.0059	0.1216	0.1241	0.179
Decision Tree Regressor	0.4977	0.6478	0.8029	-0.0129	0.1226	0.0908	0.116
Passive Aggressive Regressor	1.1619	2.3743	1.4181	-2.6582	0.2344	0.2124	0.105

The Extra Trees Regressor model is the best performing model with a MAE of 0.3966, MSE of 0.3228, RMSE of 0.5650, R2 score of 0.5028, RMSLE of 0.0869 and MAPE of 0.0728. This model is a type of ensemble learning model that builds multiple decision trees on random subsets of the training data and then averages the predictions of the individual trees. This helps to reduce the variance of the model and improve its performance.

The other models that performed well are the Random Forest Regressor, Light Gradient Boosting Machine, and Gradient Boosting Regressor. These models are all ensemble learning models that build multiple decision trees on the training data. They all have similar performance metrics, with the Random Forest Regressor having the lowest MAE and MSE and the Gradient Boosting Regressor having the highest R2 score.

The worst performing model is the Passive Aggressive Regressor. This model is a linear model that iteratively updates its weights to minimize the number of misclassifications. It has the highest MAE, MSE, RMSE, R2 score, RMSLE, and MAPE of all the models.