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

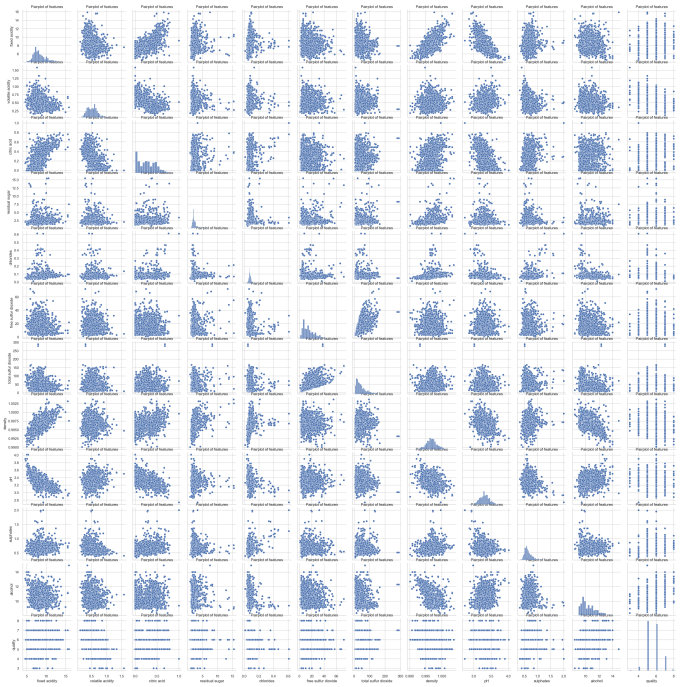
FEATURE DESCRIPTION

The dataset contains 11 features that describe different aspects of wine.

Fixed acidity is a measure of the amount of tartaric acid in the wine. Volatile acidity is a measure of the amount of acetic acid in the wine. Citric acid is a type of acid that is found naturally in wine. Residual sugar is the amount of sugar that remains after fermentation. Chlorides are salts that are found in wine. Free sulfur dioxide is a type of preservative that is added to wine. Total sulfur dioxide is the amount of free sulfur dioxide plus the amount of bound sulfur dioxide. Density is a measure of the weight of a substance per unit volume. pH is a measure of the acidity or alkalinity of a substance. Sulphates are salts of sulfuric acid. Alcohol is the percentage of alcohol by volume in the wine. Quality is a subjective measure of the wine's overall taste.

INSIGHTS ON DATASET

The dataset contains 1599 wine samples. The features include 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

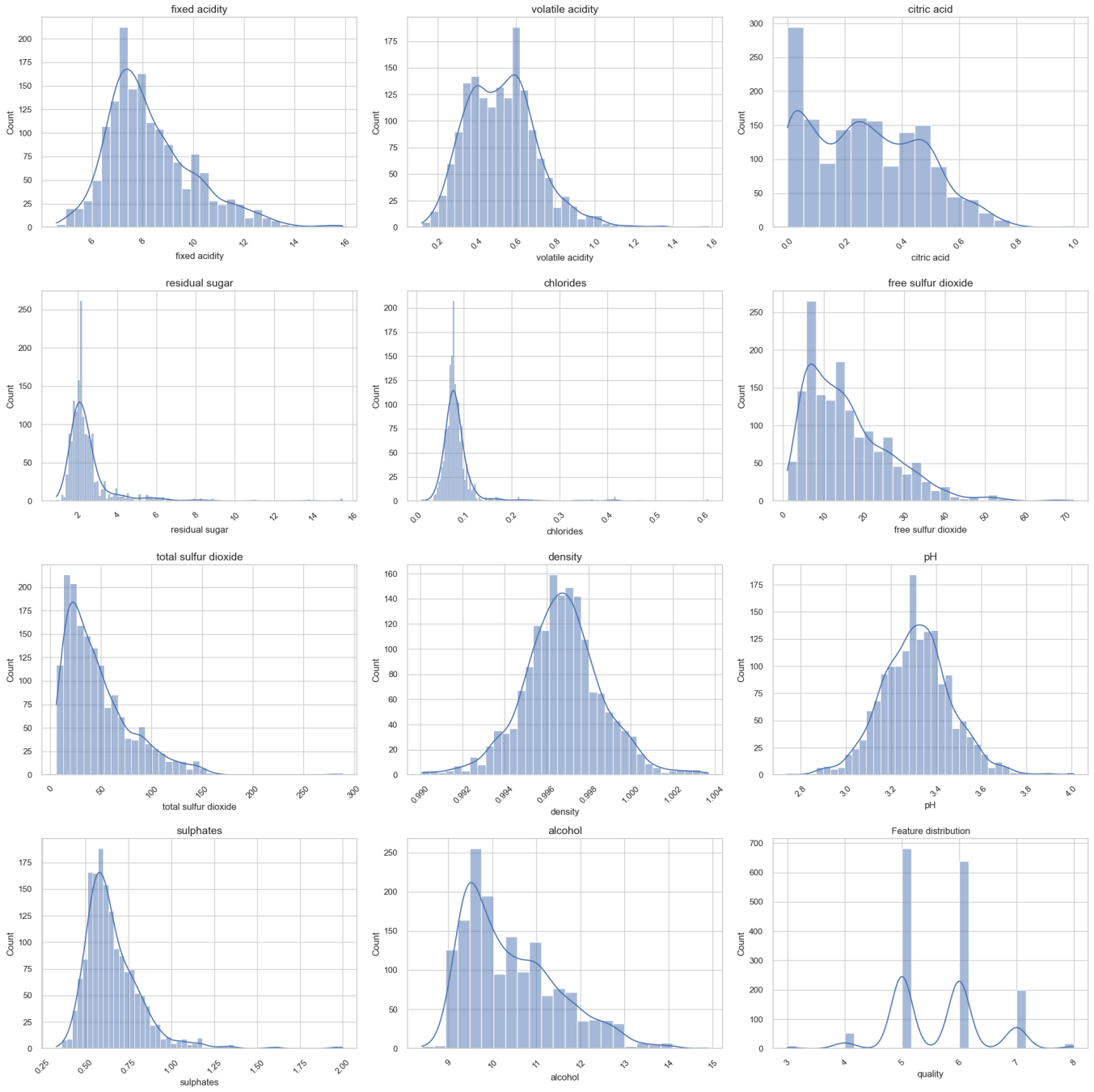
The dataset contains no missing values. This is a desirable property as it means that all of the data is available for analysis. However, it is important to note that this does not necessarily mean that the data is of high quality. Missing values can sometimes be indicative of data quality issues, such as data entry errors or data loss. In this case, it is important to carefully examine the data to ensure that it is accurate and complete.

FEATURE DISTRIBUTION

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

Fixed acidity: slightly left skewed, indicating that the data is more concentrated towards the left tail. This could be due to the fact that there are more wines with lower levels of fixed acidity than wines with higher levels.
Volatile acidity: slightly left skewed, indicating that the data is more concentrated towards the left tail. This could be due to the fact that there are more wines with lower levels of volatile acidity than wines with higher levels.
Citric acid: slightly left skewed, indicating that the data is more concentrated towards the left tail. This could be due to the fact that there are more wines with lower levels of citric acid than wines with higher levels.
Residual sugar: moderately right skewed, indicating that the data is more concentrated towards the right tail. This could be due to the fact that there are more wines with higher levels of residual sugar than wines with lower levels.
Chlorides: moderately right skewed, indicating that the data is more concentrated towards the right tail. This could be due to the fact that there are more wines with higher levels of chlorides than wines with lower levels.
Free sulfur dioxide: slightly left skewed, indicating that the data is more concentrated towards the left tail. This could be due to the fact that there are more wines with lower levels of free sulfur dioxide than wines with higher levels.
Total sulfur dioxide: slightly left skewed, indicating that the data is more concentrated towards the left tail. This could be due to the fact that there are more wines with lower levels of total sulfur dioxide than wines with higher levels.
Density: slightly left skewed, indicating that the data is more concentrated towards the left tail. This could be due to the fact that there are more wines with lower densities than wines with higher densities.
pH: slightly left skewed, indicating that the data is more concentrated towards the left tail. This could be due to the fact that there are more wines with lower pH levels than wines with higher pH levels.
Sulphates: moderately right skewed, indicating that the data is more concentrated towards the right tail. This could be due to the fact that there are more wines with higher levels of sulphates than wines with lower levels.
Alcohol: slightly left skewed, indicating that the data is more concentrated towards the left tail. This could be due to the fact that there are more wines with lower levels of alcohol than wines with higher levels.
Quality: slightly left skewed, indicating that the data is more concentrated towards the left tail. This could be due to the fact that there are more wines with lower quality ratings than wines with higher quality ratings.

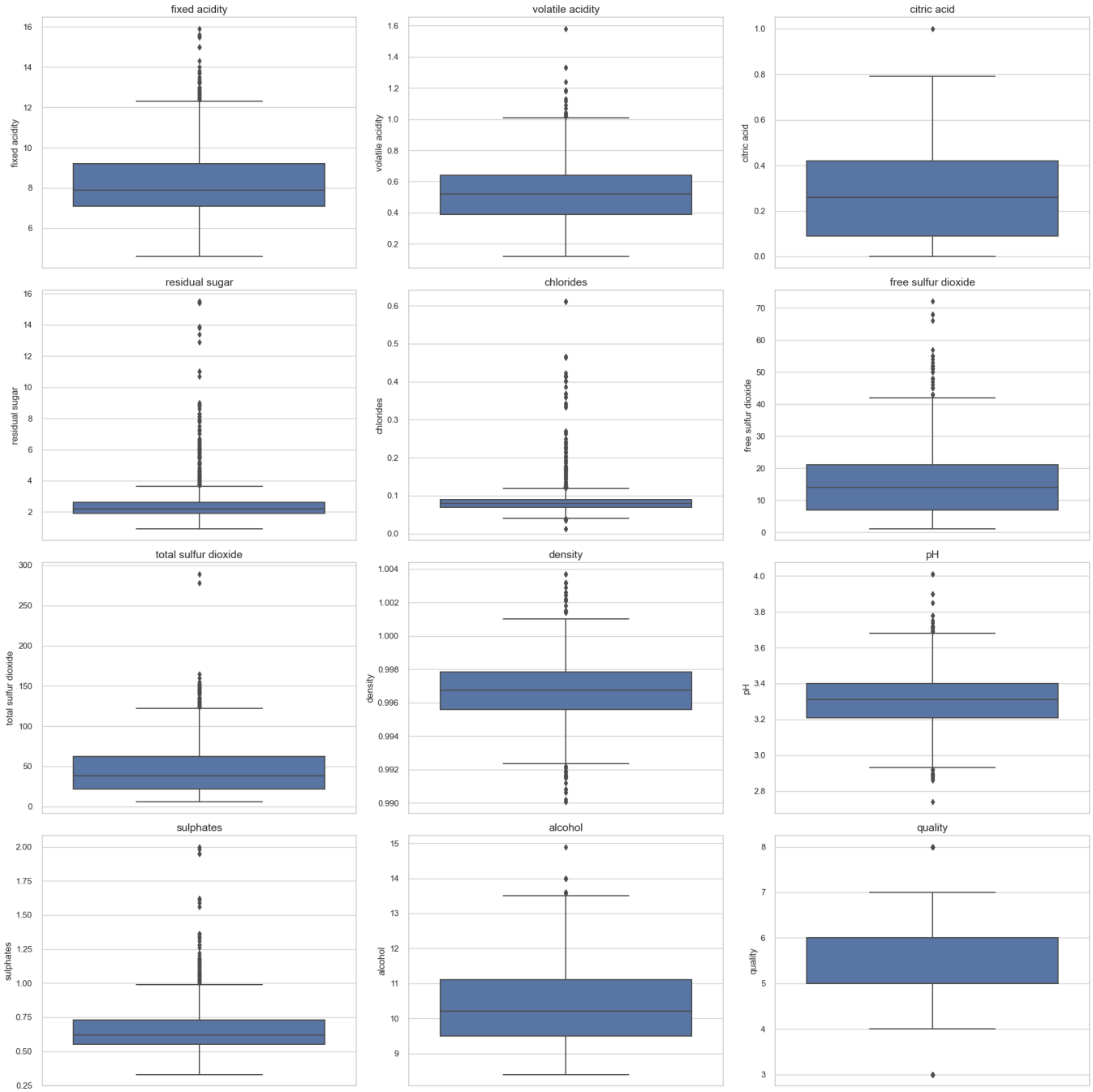
The skewness of the data could have several consequences. For example, a dataset with a high degree of skewness could be more difficult to fit a model to than a dataset with a lower degree of skewness. Additionally, a dataset with a high degree of skewness could lead to biased results if the mean or median is used to summarize the data.



OUTLIER DETECTION

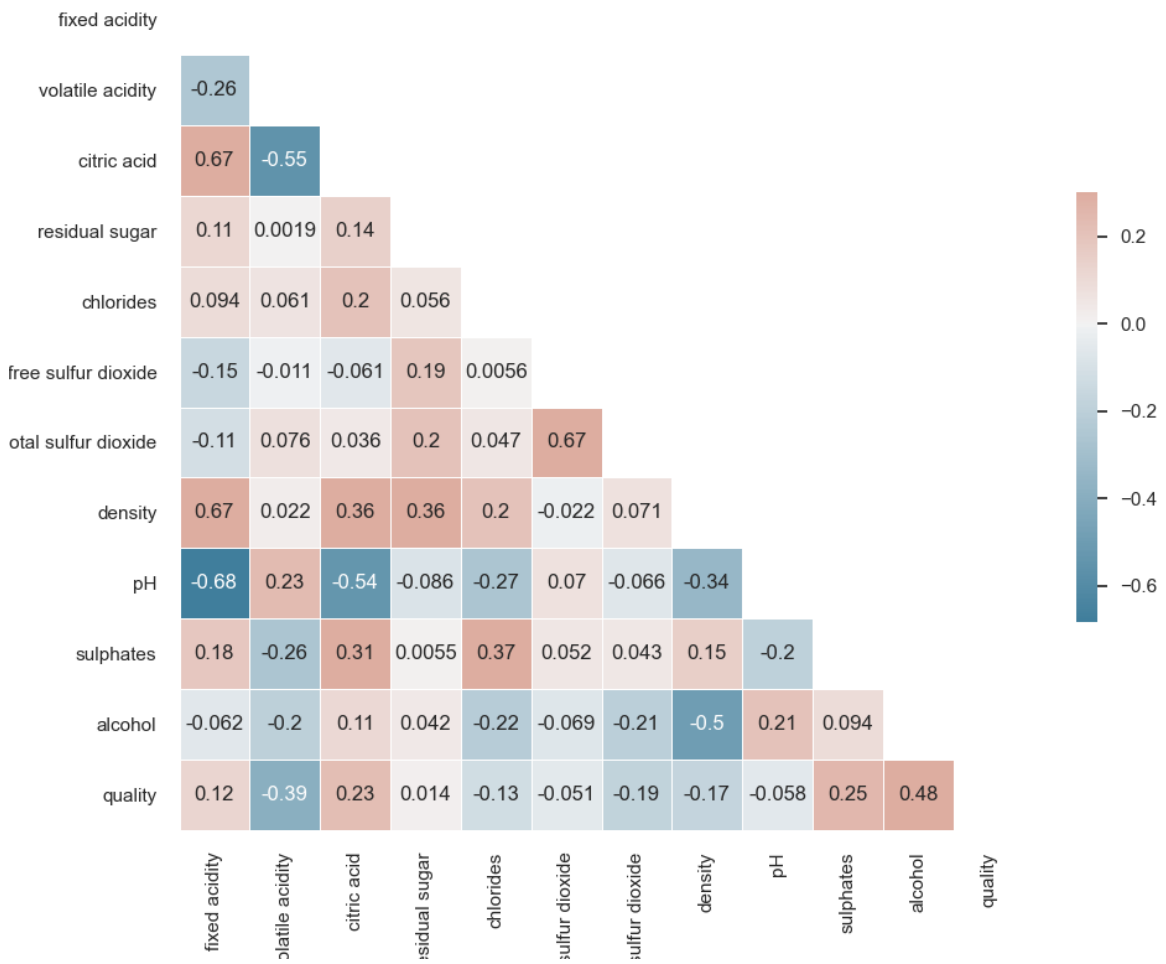
There are a few outliers in the dataset. For fixed acidity, there are 12 wines with a value greater than 15.9, which is the 75th percentile. This could be due to a number of factors, such as the wine being aged in oak barrels or having a high acidity level. For volatile acidity, there are 11 wines with a value greater than 1.58, which is the 75th percentile. This could be due to the wine being fermented with a high yeast population or having a high level of acetic acid. For citric acid, there are 13 wines with a value greater than 1.0, which is the 75th percentile. This could be due to the wine being fermented with a high yeast population or having a high level of citric acid. For residual sugar, there are 13 wines with a value less than 0.9, which is the 25th percentile. This could be due to the wine being fermented dry or having a low level of residual sugar. For chlorides, there are 11 wines with a value greater than 0.611, which is the 75th percentile. This could be due to the wine being produced in an area with high chloride levels in the soil. For free sulfur dioxide, there are 11 wines with a value greater than 72, which is the 75th percentile. This could be due to the wine being stored in a warm environment or having a high level of free sulfur dioxide. For total sulfur dioxide, there are 11 wines with a value greater than 289, which is the 75th percentile. This could be due to the wine being stored in a warm environment or having a high level of total sulfur dioxide. For density, there are 11 wines with a value less than 0.99007, which is the 25th percentile. This could be due to the wine being produced in an area with high altitude or having a low level of density. For pH, there are 11 wines with a value greater than 4.01, which is the 75th percentile. This could be due to the wine being produced in an area with high acidity levels or having a high level of pH. For sulphates, there are 11 wines with a value greater than 2.0, which is the 75th percentile. This could be due to the wine being produced in an area with high sulphate levels in the soil. For alcohol, there are 11 wines with a value greater than 14.9, which is the 75th percentile. This could be due to the wine being produced in an area with a warm climate or having a high level of alcohol.

The presence of outliers in the dataset could have a number of consequences. For example, if the outliers are not removed, they could skew the results of any analysis that is performed on the data. Additionally, if the outliers are not properly handled, they could lead to incorrect conclusions being drawn about the data. Therefore, it is important to carefully consider the presence of outliers in any dataset and to take steps to address them if necessary.



CORRELATION BETWEEN FEATURES

The correlation matrix shows that there are strong positive correlations between fixed acidity and citric acid (0.6717), volatile acidity and citric acid (0.6717), sulphates and alcohol (0.4761), and alcohol and quality (0.4761). There are also strong negative correlations between volatile acidity and pH (-0.5419), sulphates and pH (-0.1966), and alcohol and pH (-0.0577). These correlations suggest that wines with higher levels of fixed acidity and citric acid tend to have higher quality, while wines with higher levels of volatile acidity and pH tend to have lower quality.



MACHINE LEARNING EXPERIMENT SETTINGS

The experiment is conducted on a dataset of 1599 samples with 12 features. The target variable is quality and is of type regression. The data is preprocessed with simple imputation (mean for numeric features and mode for categorical features). The experiment is conducted with 10 folds using KFold and no 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
Numeric imputation	mean
Categorical imputation	mode
Fold Generator	KFold
Fold Number	10
CPU Jobs	-1

Description	Value
Use GPU	False

MACHINE LEARNING MODELS USED

Model	Explanation
Extra Trees Regressor	A tree-based ensemble model that builds multiple decision trees on random subsets of the training data and averages their predictions.
Random Forest Regressor	A tree-based ensemble model that builds multiple decision trees on random subsets of the training data and averages their predictions.
Light Gradient Boosting Machine	A gradient boosting model that uses a decision tree as its base learner and is designed to be more computationally efficient than other gradient boosting models.
Extreme Gradient Boosting	A gradient boosting model that uses a decision tree as its base learner and is designed to be more computationally efficient than other gradient boosting models.
Gradient Boosting Regressor	A gradient boosting model that uses a decision tree as its base learner.
AdaBoost Regressor	A boosting model that builds multiple decision trees on the training data, each tree focusing on the errors made by the previous trees.
Ridge Regression	A linear regression model that penalizes the size of the coefficients to reduce overfitting.
Bayesian Ridge	A linear regression model that uses Bayesian inference to estimate the coefficients.
Linear Regression	A linear model that predicts the output variable as a linear combination of the input variables.
Least Angle Regression	A linear regression model that minimizes the sum of the absolute values of the residuals.
Huber Regressor	A robust linear regression model that is less sensitive to outliers than ordinary least squares regression.
K Neighbors Regressor	A non-parametric regression model that predicts the output variable based on the k nearest neighbors in the training data.
Elastic Net	A regularized regression model that combines the features of ridge regression and lasso regression.
Orthogonal Matching Pursuit	A sparse regression model that selects the most important features from the training data.
Lasso Regression	A linear regression model that penalizes the sum of the absolute values of the coefficients to encourage sparsity.
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 of the output variable.
Decision Tree Regressor	A tree-based regression model that predicts the output variable based on a hierarchy of decisions.
Passive Aggressive Regressor	A linear regression model that iteratively updates the coefficients to minimize the number of misclassified training examples.

PERFORMANCE COMPARISON OF MACHINE LEARNING MODELS

Model	MAE	MSE	RMSE	R2	RMSLE	MAPE	TT (Sec)
Extra Trees Regressor	0.3987	0.342	0.5804	0.4667	0.0907	0.0744	0.384
Random Forest Regressor	0.4344	0.3557	0.5934	0.4423	0.0929	0.0809	0.453
Light Gradient Boosting Machine	0.4537	0.389	0.6203	0.3879	0.0966	0.0841	0.359
Extreme Gradient Boosting	0.4353	0.3928	0.6232	0.3811	0.0979	0.0816	0.39
Gradient Boosting Regressor	0.4835	0.3985	0.6277	0.3775	0.0979	0.0896	0.346
AdaBoost Regressor	0.5148	0.4238	0.6481	0.3363	0.1008	0.095	0.343
Ridge Regression	0.5145	0.4402	0.6609	0.3064	0.1025	0.0951	0.212
Bayesian Ridge	0.5152	0.4413	0.6618	0.3047	0.1026	0.0952	0.232
Linear Regression	0.5151	0.4415	0.6619	0.304	0.1026	0.0952	4.272
Least Angle Regression	0.5162	0.4439	0.6634	0.3003	0.1027	0.0954	0.225
Huber Regressor	0.5162	0.4526	0.6703	0.2858	0.1042	0.096	0.254
K Neighbors Regressor	0.5896	0.5749	0.7559	0.0936	0.1166	0.1085	0.236

Model	MAE	MSE	RMSE	R2	RMSLE	MAPE	TT (Sec)
Elastic Net	0.6599	0.6278	0.7903	0.0188	0.1214	0.1217	0.219
Orthogonal Matching Pursuit	0.6567	0.6276	0.7902	0.0188	0.1215	0.1211	0.226
Lasso Regression	0.663	0.6283	0.7907	0.018	0.1215	0.1222	0.23
Lasso Least Angle Regression	0.663	0.6283	0.7907	0.018	0.1215	0.1222	0.224
Dummy Regressor	0.6852	0.6443	0.8006	-0.0065	0.1228	0.1261	0.292
Decision Tree Regressor	0.5067	0.68	0.8241	-0.0843	0.129	0.094	0.238
Passive Aggressive Regressor	0.8698	1.2688	1.0801	-1.0988	0.1711	0.162	0.226

The results show that Extra Trees Regressor (ETR) has the best performance among all the models, with the lowest MAE, MSE, RMSE, and MAPE. It also has a high R2 score, indicating that it is able to explain a large portion of the variance in the data. ETR is a relatively simple model to train, and it does not require much tuning. This makes it a good choice for models that need to be deployed quickly and easily.

Other models that performed well include Light Gradient Boosting Machine (LGBM), XGBoost, and Gradient Boosting Regressor (GBR). These models are all ensemble methods, which means that they combine the predictions of multiple models to improve performance. Ensemble methods can be very effective, but they can also be more complex to train and tune.

The worst performing model was Passive Aggressive Regressor (PAR). PAR is a simple linear model that is not very powerful. It is also not very robust to overfitting, which means that it can perform poorly on data that is not well-represented in the training set.