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 dataset contains 11 features that describe different aspects of wine.

Fixed acidity is the amount of tartaric acid present in the wine. Volatile acidity is a measure of the amount of acetic acid present in the wine. Citric acid is a type of acid that occurs naturally in grapes. 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 to prevent the growth of bacteria. 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 compared to the weight of an equal volume of water. 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 overall quality of the wine.

Insights on dataset

The dataset contains 1599 rows and 12 columns.
The features are: fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, alcohol, and quality.
The mean, standard deviation, minimum, 25th percentile, 50th percentile, 75th percentile, and maximum values of each feature are reported.
Based on the statistics, we can see that the fixed acidity has a mean of 8.319637, a standard deviation of 1.741096, a minimum of 4.6, a 25th percentile of 7.1, a 50th percentile of 7.9, a 75th percentile of 9.2, and a maximum of 15.9.
The volatile acidity has a mean of 0.527821, a standard deviation of 0.179060, a minimum of 0.12, a 25th percentile of 0.39, a 50th percentile of 0.52, a 75th percentile of 0.64, and a maximum of 1.58.
The citric acid has a mean of 0.270976, a standard deviation of 0.194801, a minimum of 0.0, a 25th percentile of 0.09, a 50th percentile of 0.26, a 75th percentile of 0.42, and a maximum of 1.0.

Insights on Null Values in the dataset

The dataset contains no null values for any of the features. This is an ideal situation, as it means that all of the data is available for analysis. However, it is important to note that this dataset is relatively small, and that null values may be more common in larger datasets. If a dataset contains null values, it is important to consider how to handle them before performing analysis. One option is to simply remove all rows that contain null values, but this may result in a loss of data. Another option is to impute the missing values, either with the mean, median, or mode of the non-null values, or with a random value. The choice of which imputation method to use will depend on the specific dataset and the analysis that is being performed.

Feature Distribution

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

* Fixed acidity: slightly right skewed, indicating that the data is more concentrated towards the lower values. This could be due to the fact that there is a minimum legal limit for fixed acidity in wine, so most wines will fall within this range.
* Volatile acidity: slightly left skewed, indicating that the data is more concentrated towards the higher values. This could be due to the fact that volatile acidity is a measure of the amount of acetic acid in wine, and acetic acid is a byproduct of fermentation.
* Citric acid: slightly left skewed, indicating that the data is more concentrated towards the higher values. This could be due to the fact that citric acid is a natural preservative in wine, and so wines with higher levels of citric acid are more likely to last longer.
* Residual sugar: moderately right skewed, indicating that the data is more concentrated towards the lower values. This could be due to the fact that residual sugar is a measure of the amount of sugar that remains in the

wine after fermentation.

* Chlorides: moderately right skewed, indicating that the data is more concentrated towards the lower values. This could be due to the fact that chlorides are a natural component of wine, and so wines with higher levels of chlorides are more likely to be salty.
* Free sulfur dioxide: slightly left skewed, indicating that the data is more concentrated towards the higher values. This could be due to the fact that free sulfur dioxide is a preservative that is added to wine to prevent it from spoiling.
* Total sulfur dioxide: slightly left skewed, indicating that the data is more concentrated towards the higher values. This could be due to the fact that total sulfur dioxide is a measure of the amount of free sulfur dioxide and bound sulfur dioxide in wine.
* Density: slightly left skewed, indicating that the data is more concentrated towards the higher values. This could be due to the fact that density is a measure of the weight of a substance per unit volume, and so wines with higher densities are more likely to be viscous.
* pH: slightly left skewed, indicating that the data is more concentrated towards the higher values. This could be due to the fact that pH is a measure of the acidity of a substance, and so wines with higher pH values are more likely to be acidic.
* Sulphates: moderately right skewed, indicating that the data is more concentrated towards the lower values. This could be due to the fact that sulphates are a natural component of wine, and so wines with higher levels of sulphates are more likely to be bitter.
* Alcohol: slightly right skewed, indicating that the data is more concentrated towards the lower values. This could be due to the fact that alcohol is a product of fermentation, and so wines with higher levels of alcohol are more likely to be alcoholic.
* Quality: slightly left skewed, indicating that the data is more concentrated towards the higher values. This could be due to the fact that quality is a subjective measure, and so wines with higher quality ratings are more likely to be rated highly.

The skewness of the data could have a number of consequences. For example, a dataset with a high degree of skewness could be more difficult to fit a model to, as the model may not be able to accurately capture the distribution of the data. Additionally, a dataset with a high degree of skewness could lead to biased results, as the model may be more likely to overfit to the data in the tails of the distribution.