Dorcas Munoko
Big Data and Forecasting
2/3/2022

Predicting Top Quality Wine

Introduction

The aim of the project was to provide a classifier that can predict top-quality wines based on the observable characteristics. The data contains 1599 rows and 12 columns. I created data visualizations to know the data and patterns.

Data Preparation and Visualization

My first step was to clean the data to prepare it for analysis. I checked data type and missing values in dataset. Finally, I checked for outliers in each column in the dataset. I created a count plot to determine the number of values for each quality. The quality value starts from 3 to 8, but the number of values is greater for wine qualities 5 and 6. I also created bar plots to determine which values are related to quality. The output results indicate that volatile acidity and chloride are inversely proportional to quality, while citric acid, alcohol, and sulphates are directly proportional to quality. Fixed acidity, residual sugar, free chloride dioxide, and total sulfur dioxide show no significant relationship with quality. I ran a correlation analysis of the variables against quality. The highest correlation is between alcohol and quality (0.48), volatile acidity and alcohol (-0.2), sulphates (0.25), citric acid (0.23), total sulfur dioxide (-0.19), density (0.019), chlorides (-0.041), fixed acidity (-0.021), pH (-0.0011), and free sulfur dioxide (-0.049).

Algorithms, Methodology, and Results

I began by separating out quality data to be used in machine learning. I used the Naïve Bayes classifier technique to classify wine quality. This technique holds an assumption of independence. It assumes that a class feature is unrelated to any other feature. The accuracy results for the three models are 54.3% (Gaussian), 43.6% (Multinomial), 41.5% (Bernoulli). The 10-fold cross-validation score for the models is 56.8% (Gaussian), 44.7% (Multinomial), and 43.7% (Bernoulli). In conclusion, the Gaussian model did great in predicting wine quality. It has the highest accuracy rate (54.3%) and score (56.8%). It is also the best model for continuous data unlike Bernoulli which is best for binary data, and multinomial which is best for categorical data.