**Summary of Classification Methods on Wine Quality**

**Explanatory Data Analysis: Cleaning and Transformation**

1. I have looked at the summary statistics of the data, which includes the average, minimum value, maximum value, distribution of each wine characteristics. I found that there are a few missing values in some data columns, so I dropped these.
2. I plotted the data for each wine characteristics. First, look at the distribution. Then, check for any outliers (extremely small or big values) and remove if any.
3. Calculate the correlation of each wine characteristics to the outcome variable (wine quality), which shows how closely related are the characteristics to quality.
4. Convert ‘quality’ to a binary variable, where the value is 1 if quality is greater than 6, and 0 otherwise.

**Implement Gaussian/Bernoulli Naïve Bayes Classifier**

1. Split the data into train/test dataset in 80:20 ratio. 80% of the dataset will be train data and 20% will be test data. All wine characteristics are used for the classifiers.
2. I started with implementing Gaussian classifier. Then, I created a confusion matrix to illustrate the results. The model predicts whether a wine is considered good quality, based on the wine characteristics. The model has a 49% rate of correctly predicting whether a wine is good quality. Considering there are only 12.9% wines are good quality in the dataset provided for this assignment, I believe 49% rate is fairly good. The overall accuracy is 86.2%.
3. Then, I implemented Bernoulli model for comparison. The result shows that it has 0 chance of correctly predicting a good quality wine.
4. Furthermore, I removed a few features that have low correlations to quality, and re-implemented Gaussian model. The result shows that it has a 45.5% rate of correctly predicting whether a wine is good quality. The overall accuracy rate is 84.8%.
5. Lastly, I used cross-validation techniques on all three attempts to test the robustness of the models. The results support the first Gaussian model based on all wine characteristics, which on average, has a 43% rate of correctly predicting whether a wine is good quality and a 94% rate of correctly predicting whether a wine is bad quality.
6. I recommend my client to use Gaussian Naïve Bayes Classifier to predict high quality wines (based on all wine characteristics).