Lab 2

The wine quality dataset has the number of various chemicals in wine and the wine quality score. This dataset led me to wonder if wine manufacturers could predict wine quality by using a classification model. Since wine production is supposed to be a very tedious process, predicting wine quality may help manufacturers cut costs and increase profit. Through data mining, wine manufacturers can understand the relationship between different chemical factors and wine quality. Also, they can predict wine quality to avoid waste of resources and save costs. In this lab, I will create a confusion matrix and generate a classification report for wine quality's logistic regression model.

In the dataset, the "quality score" data is scored between 0 to 10. However, I would like to use the model to predict whether the wine quality is either good or bad. Therefore, I set the score below average quality score as "0", indicating that the wine is of poor quality. I then converted the above-average quality score into a "1" for good quality wine. Next, I created some data visualizations to view and understand the data more precisely. In the count plot, there is more good quality wine than bad. In the box plot, it is obvious that the better quality wines have higher alcohol content. In the scatter plot, we can see that the reason for the poor wine quality may be due to more total sulfur dioxide in the wine.

Next, I will use this dataset to generate a confusion matrix. Since this dataset did not contain any data that was null. I did not have to get rid of any useless data except for the id column that will disturb the classification model. I set the chemical elements data in wine to X and the wine quality to y. Then, I split the dataset to test and train data in a ratio of 3:7. In the confusion matrix results, the number of wine quality predicted to be good and actually good is 122 (True Positive). The number of wine quality predicted to be bad but actually good is 37 (False Negative). The number of wine quality predicted to be good but actually bad is 40 (False Positive). The number of wine quality predicted to be bad and actually bad is 144 (True Negative). These confusion matrix results help me generate classification reports.

According to the results of the classification report, I can evaluate the accuracy results of the wine quality prediction model. The precision result with good quality (class 1) is 78% while that of wines with the poor quality precision result (class 0) is 77%, so the precision accuracy of class 1 is higher than that of class 0. In the recall score of the wine quality prediction model, the model's ability to find all positive instances of poor wine quality was 75%, and the model's ability to find all positive instances of good wine quality was 80%. Therefore, the model's ability to find all positive instances of good wine quality was better performed. Moreover, the F1 score of the two classes is not close to 1, indicating that the accuracy of optimistic prediction is not perfect. However, this number is for comparing classifier models, not data accuracy. The support result shows that the number of occurrences of these two classes in the dataset is relatively balanced. As a result, the content of various chemical substances has a close influence on the quality of red wine. Also, we can predict the quality of wine through data mining and a classifier model.