COMP30027 Assignment_1 Written Report

-Muhan Chu 1346862

2. 1-NN Classification

The accuracy of 1-NN classification is 0.764. In the later figure, it can be seen that within k of 1-20, the prediction for k = 1 is the best. So for this data-set, it is suitable when k = 1.

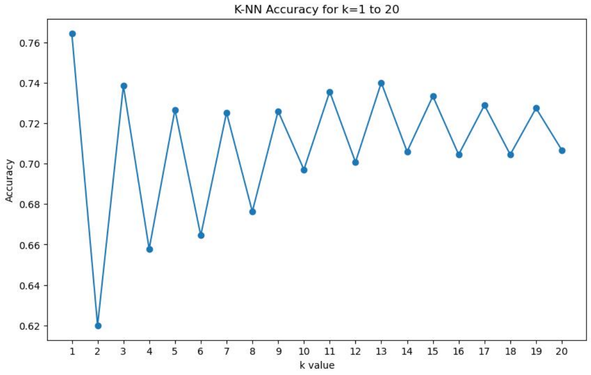

Figure 1. K-NN accuracy for k = 1 to 20

However, there are still many potential problems with 1-NN. For example, if high-quality wines are matched with low-quality outliers or noise, then the prediction results will be wrong. In addition, it can be seen from the scatter plot. The boundary between high quality wines and low quality wines is not clear, and there are even some repetitive points. Therefore, when both red and blue dots appear around the test wine, it is arbitrary to choose only the newest wine as a reference. And in the data set, there are more low-quality wines than high-quality wines, and there is a higher probability of obtaining low-quality wines in the prediction data. And the reason this data set predicts better may be that the data just happens to circumvent all these problems. So it is still important to use different parameters for comparison when making predictions.

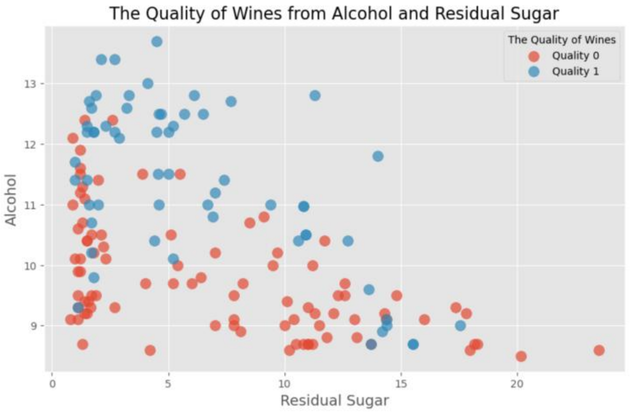

Figure 2. Scatter plot drawn by Alcohol and residual sugar

3. Normalization

By predicting the testing data set, the correct rate of prediction without normalization is 0.764, while the correct rate of prediction after Min-max scale and standardization is 0.850 and 0.864 respectively, which directly indicates that normalization is very meaningful.

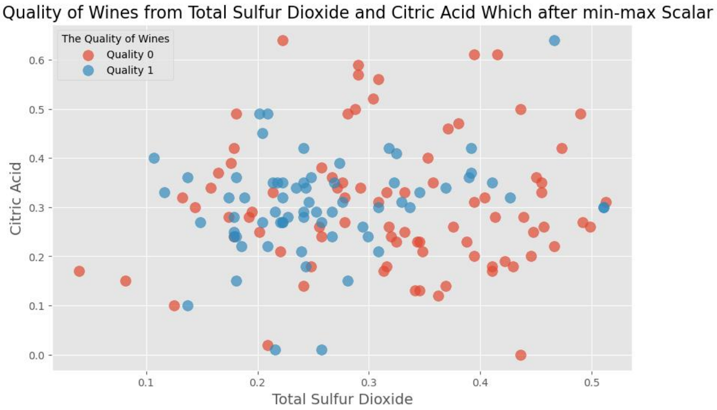

Figure 3. Scatter plot based on min-max scale total sulfur dioxide and citric Acid

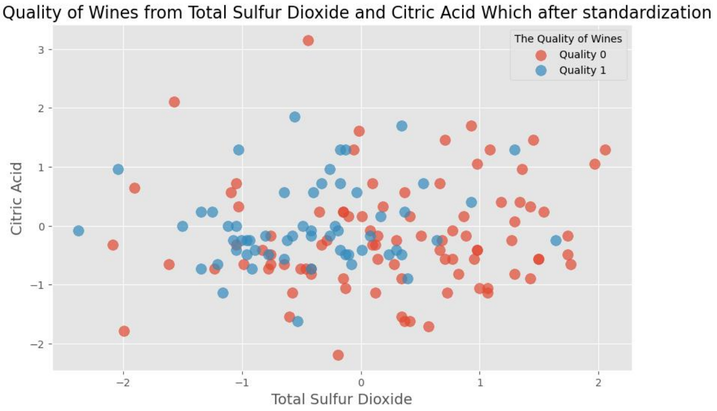

Figure 4. scatter plot based on standardization total sulfur dioxide and citric Acid

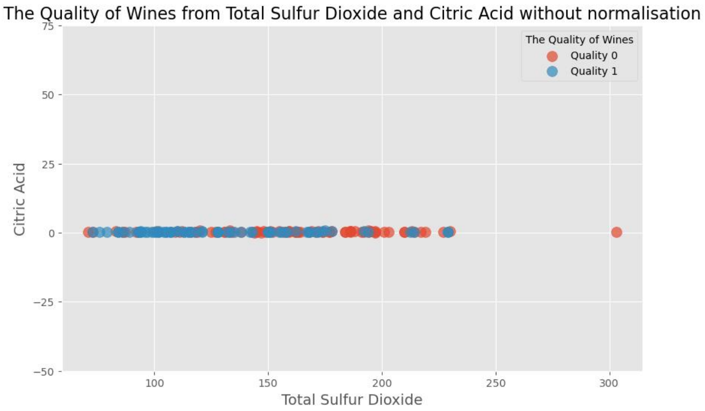

Figure 5. scatter plots without normalization of total sulfur dioxide and citric Acid

The three scatter plots also present a big problem, the distribution of the normalized data is much clearer compared to the um-normalized data, and it can be seen that there is a boundary between the red and blue points. The presentation of the data without normalization becomes very flat and the partitioning is not visible on the graph at all.

The purpose of normalization is to bring the range of the data to a specific range. The Min-Max Scale used here is to reduce all values to a range of 0-1, and standardize is to convert the data to a standard normal distribution, where the difference between the Citric Acid and the interval of Total Sulfur Dioxide is particularly large.

In this case the difference between Citric Acid and Total Sulfur Dioxide interval is particularly large, which leads to the fact that Total Sulfur Dioxide has a very strong influence on the distance when calculating the distance between the wines. The effect of Citric Acid on distance will be very small. As a result, the features with small values become invalid features, which will directly lead to a decrease in the correctness of the 1-NN prediction.The reason standardization performs better here is that Min-Max scale is affected by outlier

4.1 Model extensions: Gaussian naive Bayes model

After the same normalization of the data, the Gaussian naive Bayes model (GNB) predicts correctly 0.774 while 1-NN predicts correctly 0.867. There are 272 data in the testing data set the predictions of these two models are different. Among them 1-NN predicted 199 correctly and GNB predicted 73 correctly.

The potential reason for the relatively low accuracy of the GNB model here is that it assumes that the data are independent of each other and follow normal distribution. But the data here does not follow these two assumptions very well.

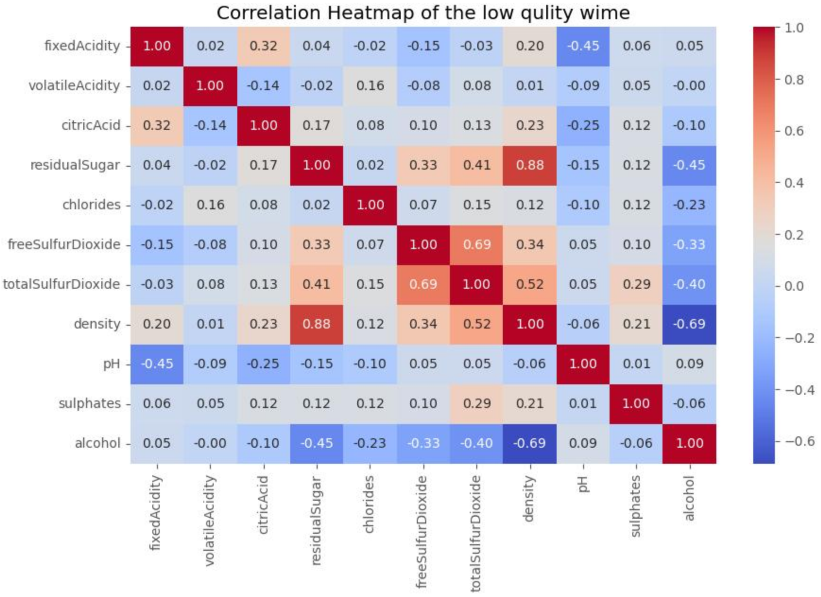

Figure 6. Correlation between features in low quality

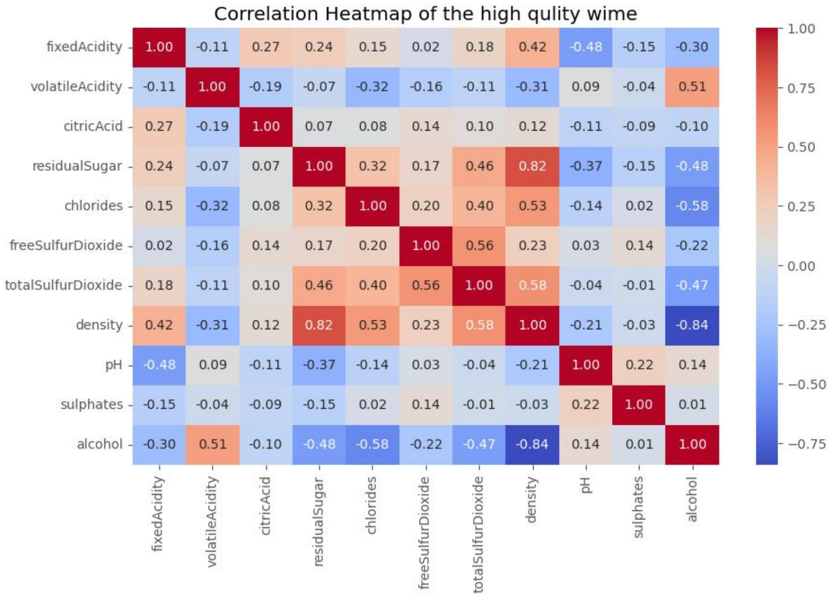

Figure 7. correlation between features in high quality wine

The correlation between feature can be seen in these 2 graphs above, and the correlation between density and sulphate, density and residual sugar is very strong. It is possible that they have a linear relationship or are not independent of each other, which would likely lead to inaccurate model predictions.

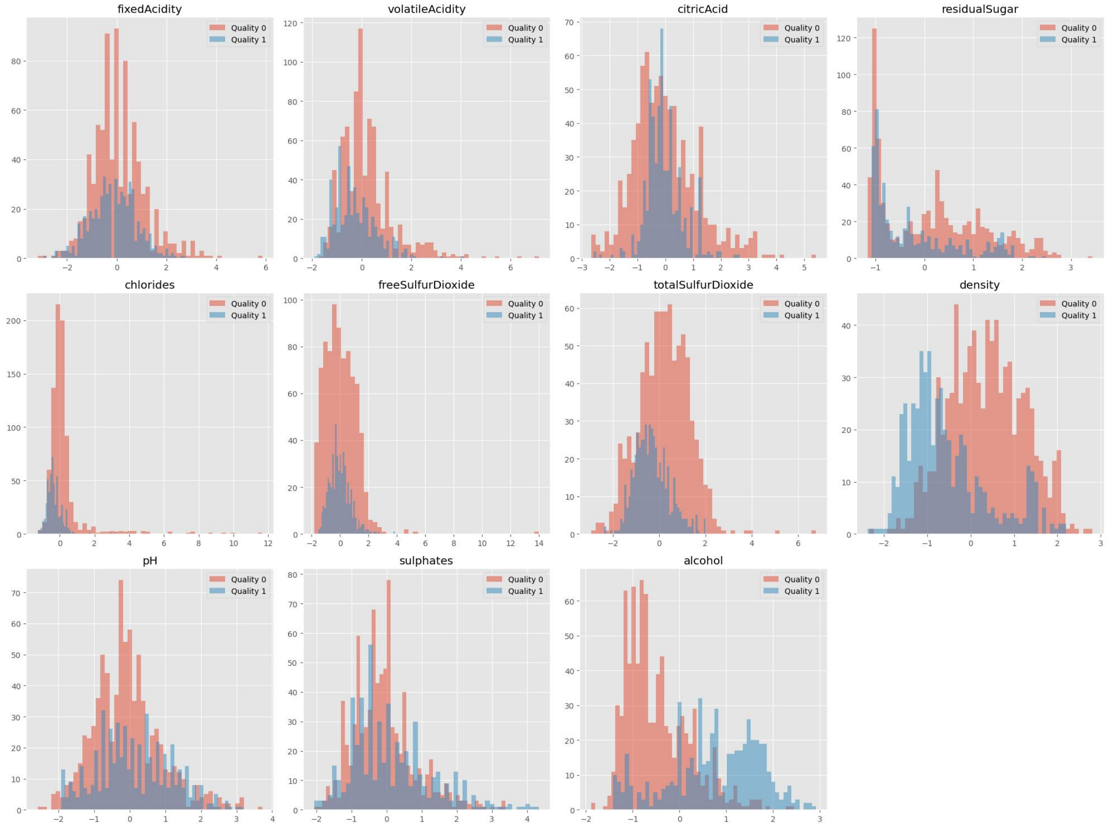

Figure 8. Distribution of each feature

It can also be seen from this picture that not all features follow the normal distribution. From the graph, it seems that only the pH distribution is close to the normal distribution and is symmetrical, while the distributions of Residual Sugar, Free Sulfur Dioxide, and Total Sulfur Dioxide have long tails, which may be caused by the outlier. The other features also have more or less bias or multi-peak problems. This can all lead to inaccurate predictions. But here 1-NN has no requirements on the data and can relatively adapt to different data even if they don't have a good distribution。