Question 1

As shown in codes.

Question 2.1

Accuracy= (Number of correctly labeled test instances)/(Total number of test instances)
=(652(TN)+380(TP))/1350=**76.44%**

Question 2.2

1. 1-NN classifies the test set only based on the one nearest point in the training set, which is more sensitive to noise and causes overfitting. Looking at the scatterplot of two features, residualSugar and density (**Figure 1 left**), points in the **top right (the red circle)** with both high density and residualSugar are more likely to be the class 0, while there are also some points of class 1 around the same location (possibly noise). If a class 0 instance of the test set had both high residualSugar and density but was more closed to a class 1 point, it would be predicted as the class 1 by mistake. The same situation also applies for the scatterplot of residualSugar versus alcohol (**bottom right, the red circle,Figure 1 right**).

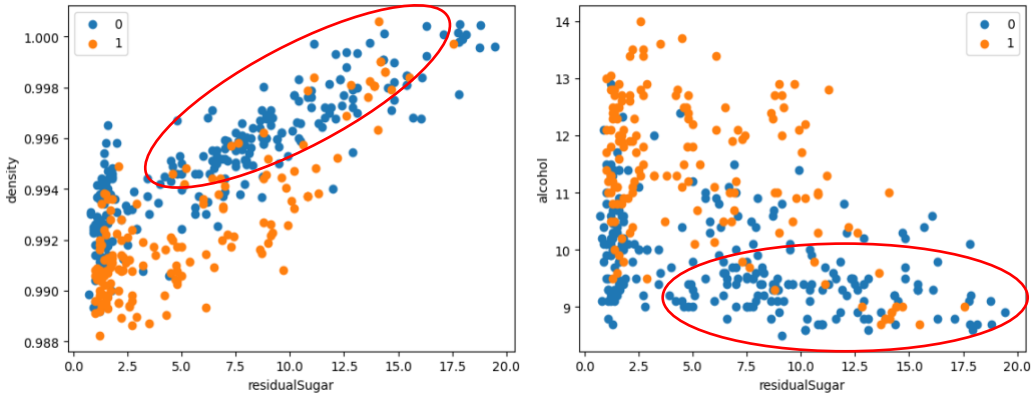

Figure 1 The scatterplots of residualSugar vs density (left) and residualSugar vs alcohol (right)

2. Given that 1-NN together with no normalization, features with large scale will dominate the calculation of the distance and the nearest point. However, two features, **freeSulfurDioxide** and **totalSulfurDioxide**, with large scales, cannot separate the class 1 and 0 very well (**Figure 2**), class 0 and 1 points clustering together, thus making this 1-NN less powerful.

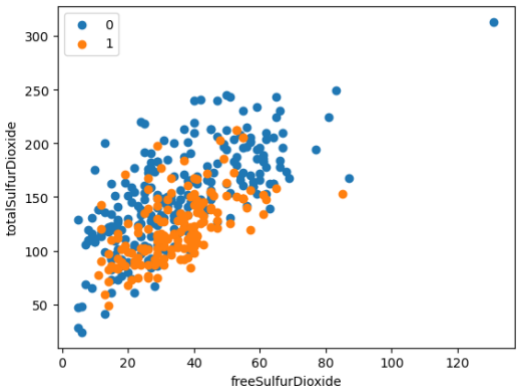

Figure 2 The scatterplot of freeSulfurDioxide vs totalSulfurDioxide

Question 3

After **min-max normalization**, the accuracy is **85.04%**; and after **standardizing all attributes**, the accuracy is **86.74%**, both higher than the accuracy achieved without normalization (76.44%). Normalization aligns all attributes to a similar scale, allowing attributes with smaller scales to contribute more effectively to the prediction process.

Looking at one feature, **density**, with a small scale, when no normalization, the difference of density between two classes is not clear compared with **freeSulfurDioxide** (the large-scale feature) (**Figure 3 left**), while after normalization, the difference becomes clear, where high density for class 0 and low density for class 1 (**Figure 3 middle and right**). Thus, the accuracies after normalization become higher.

Comparing the two normalization methods, the **standardizing method** is less susceptible to outliers. There being outliers for **freeSulfurDioxide**, the scale of freeSulfurDioxide for the main cluster is around 0 to 0.2 (smaller than the scale of density, 0.1 to 0.9) for min-max normalization (**Figure 3 middle**) while for the standardizing method, it is around -2 to 2 (comparable with the scale of density, -2 to 2) (**Figure 3 right**). It is a possible reason why there is a slightly **higher** accuracy for the **standardizing method** than min-max normalization.

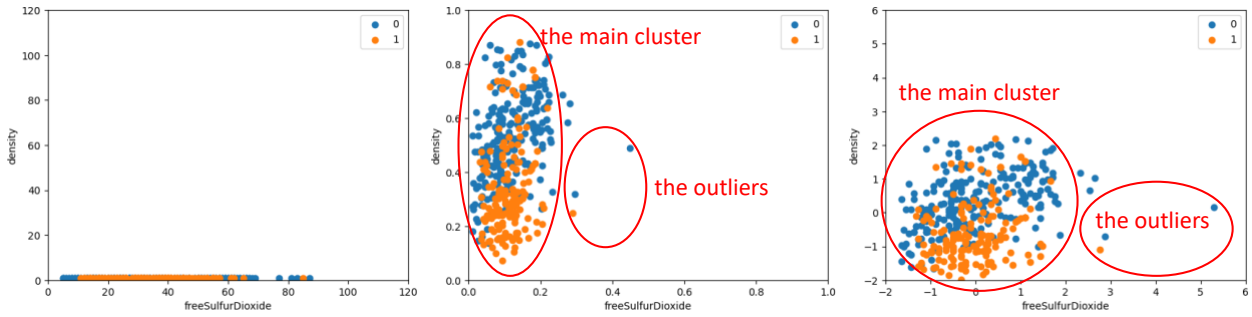

Figure 3 The scatterplots of freeSulfurDioxide vs density without normalization (left), after min-max normalization (middle), and after standardizing method (right)

Question 4.1

1. The accuracy of GNB model is **77.78%**, lower than the best 1-NN (**86.74%**). There are **265** instances where the two models disagree, among which, 153 are class 0, and 112 are class 1; and 1-NN predicts **193** correctly, while GNB predicts **72** correctly. Thus, the best 1-NN performs better.
2. Among the **265 instances**, **residualSugar**, **totalSulfurDioxide**, and **alcohol**, have reversed trends in mean values when compared with **all instances in the training set** (as shown in the table). For example, in the training set, **class 0** has **higher residualSugar** than class 1, whereas for those 265 instances, **class 0** has **lower** residualSugar.

`residualSugar mean:`

	`All instances`	`265 instances`
`class 0`	`7.091707317073167`	`4.108823529411765`
`class 1`	`5.25339622641509`	`8.548660714285717`

`totalSulfurDioxide mean:`

	All instances	265 instances
class 0	148.79756097560977	121.83006535947712
class 1	125.56981132075472	146.45089285714286

alcohol mean:

	All instances	265 instances
class 0	9.823126016260156	10.596732026143792
class 1	11.397182389937106	10.131696428571427

3. In the training set, **residualSugar**, **totalSulfurDioxide**, and **density** have negative correlations with class 1, but in the 265 instances, they have positive correlations with class 1. (**Figure 4**)
4. In the training set, **alcohol** has positive correlations with class 1, but in the 265 instances, it has negative correlations with class 1. (**Figure 4**)
5. The calculation of GNB probabilities is by multiplying of each feature while the calculation of 1-NN distances is by the addition of each feature. Thus, these reversed trends or correlations will impact the probability calculation more than the distance calculation, potentially resulting in a lower accuracy for GNB.
6. Further, naïve Bayes model assumes that each feature is independent, but for the training set, many pair-wise features have high correlations, like **density** and **residualSugar** with a correlation coefficient of 0.83. These correlations will not influence 1-NN. Therefore, GNB has poorer performance. (**Figure 4**)

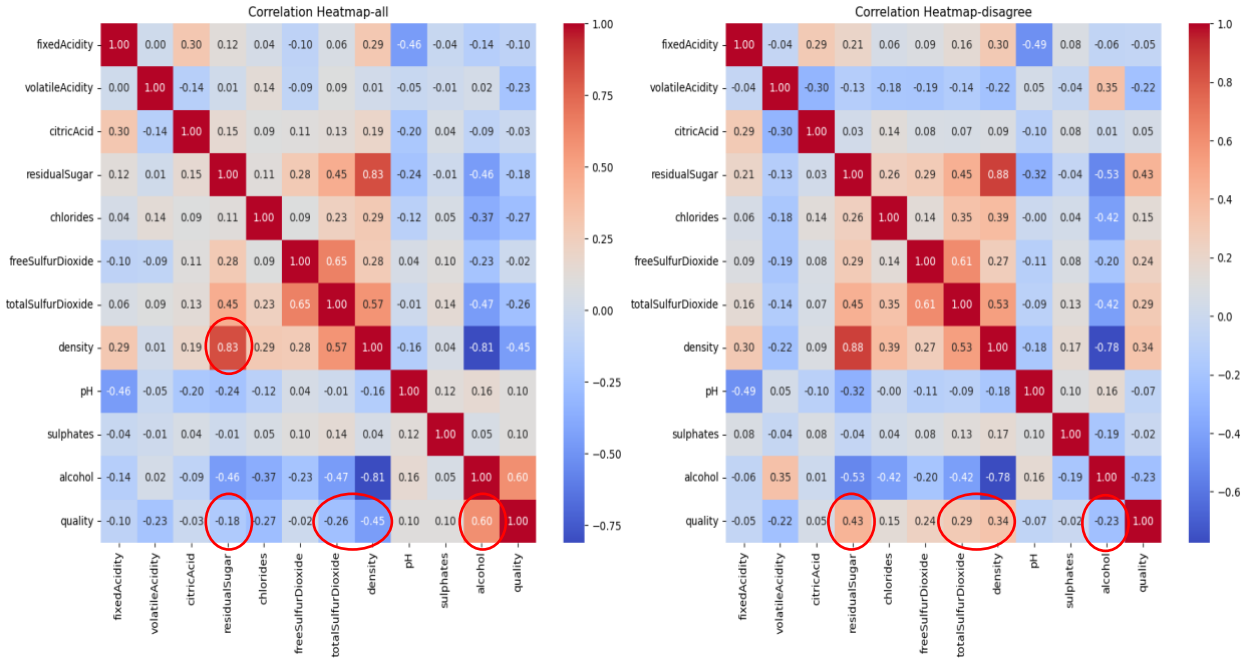

Figure 4 The correlation coefficients between features and between features and the class of quality for the training set (left) and the 265 instances (right).