Analysis Report

2. 1-NN Classification

The accuracy of 1-NN classifier is 76.44%, and the scatterplots in 2D from a subset of 500 samples are provided as follows:

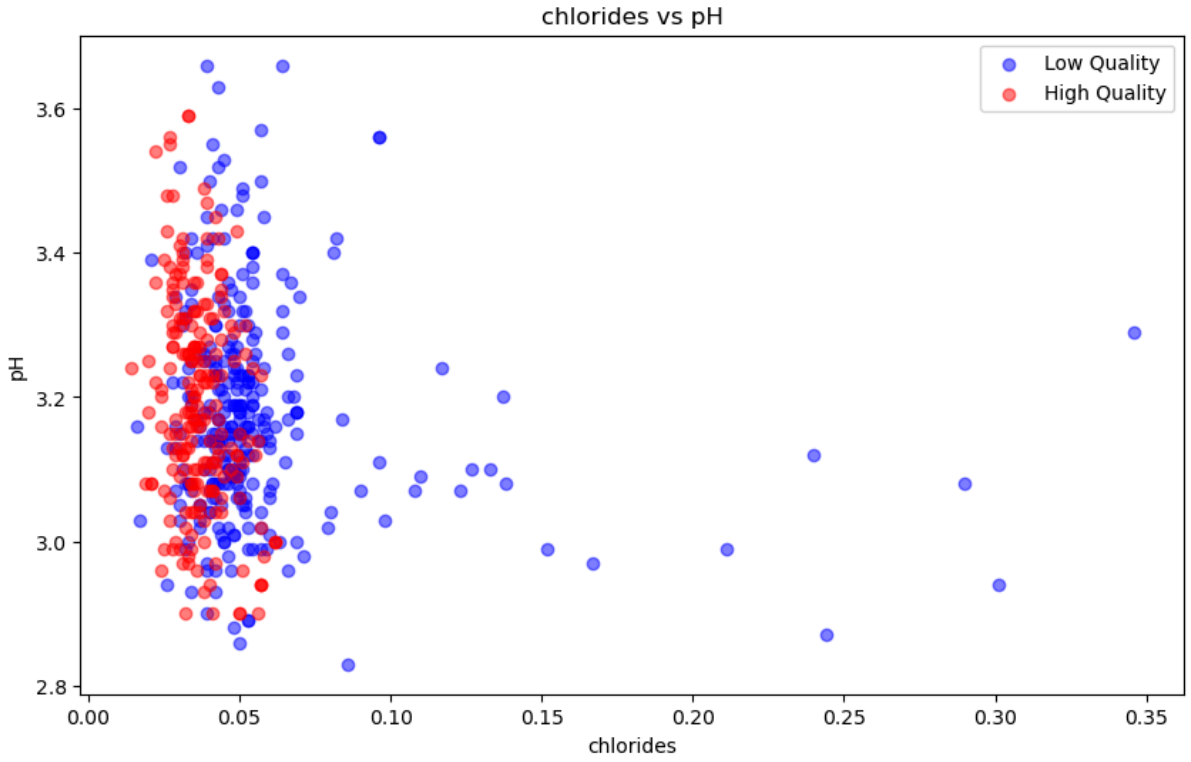

Figure 1

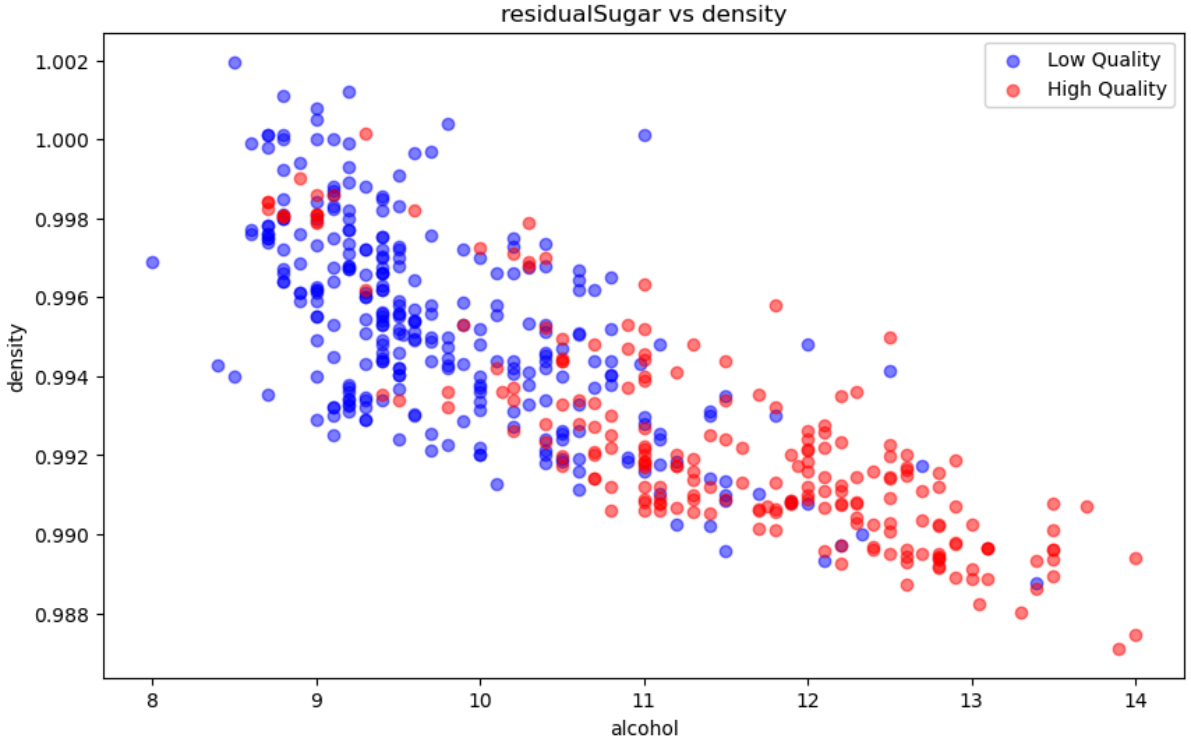

Figure 2

In Figure 1, there are a few “Low Quality” points with low pH and relative high chlorides values that seem to be isolated or distant from the majority of points. These points could potentially be considered as outliers, which can negatively impact the performance of the 1-NN classifier by biasing the classification of neighboring instances. Moreover, as the chlorides values decrease to under 0.1, a notable overlap between pH and chlorides emerges, posing an additional challenge for accurate classification in this region since points from different classes are closely intermingled.

In Figure 2, the classes show some separability, such as instances are mostly classified as "Low Quality" for lower alcohol and high density values. This suggests that the 1-NN classifier could achieve reasonable accuracy by leveraging the proximity of instances within the same class cluster. However, the overall decision boundary separating the two classes is complex and non-linear when alcohol is around 11, despite the apparent separability in this specific region. The 1-NN classifier would need to create a complicated decision boundary to separate the instances, increasing the risk of overfitting.

Moreover, in the actual high-dimensional dataset with 11 attributes, sparsity makes it even hard for a 1-NN classifier to find similar instances. Instances that appear "close" in lower dimensions, may be considerably distant in the high-dimensional space, which might result in misclassifications, making it difficult to attain high accuracy with the 1-NN classifier.

3. Normalization

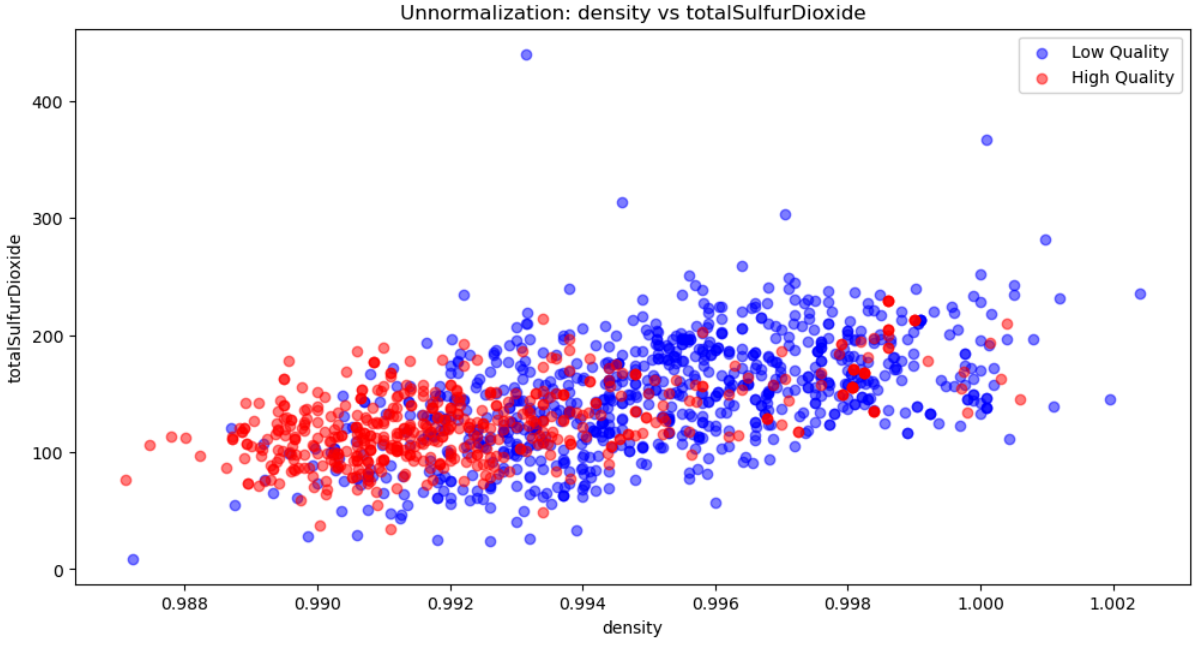

Figure 3

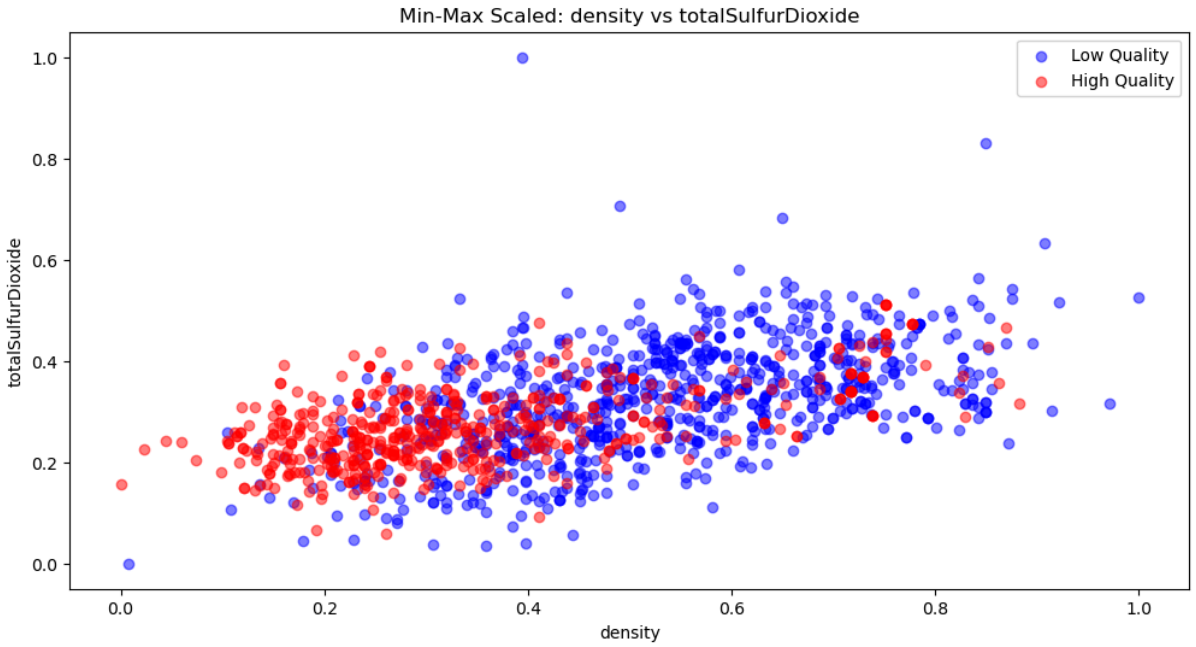

Figure 4

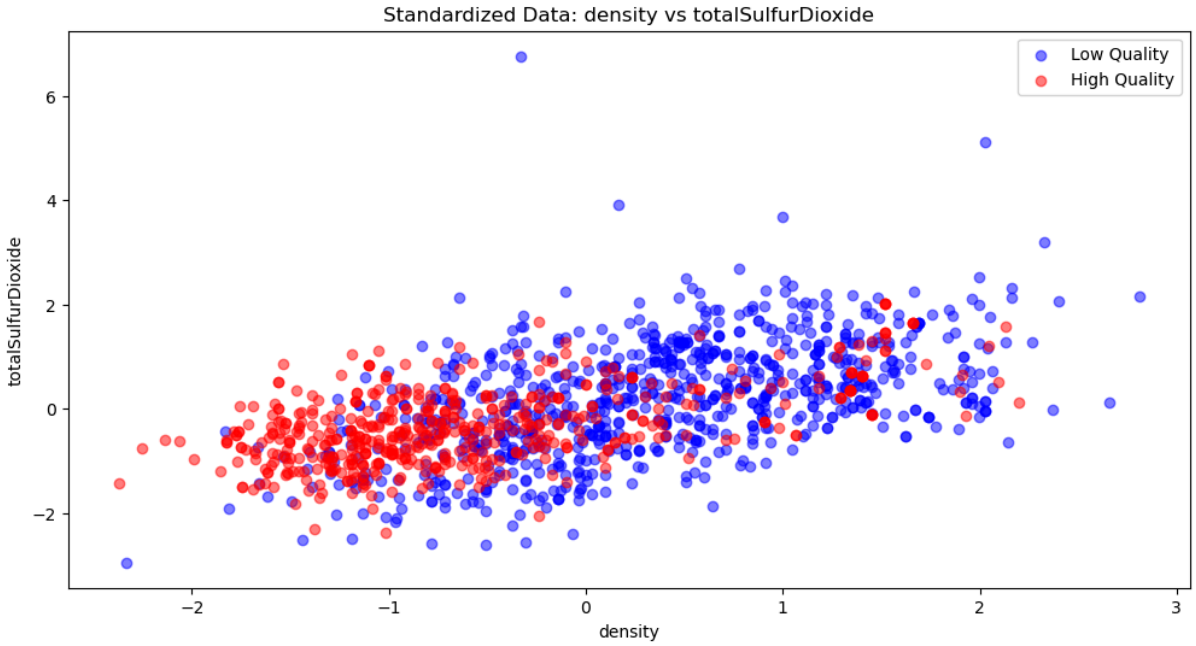

Figure 5

In our dataset, both min-max scaling and standardization significantly improved the accuracy of the 1-NN classifier compared to unnormalized data. The accuracy increased from 76.44% for unnormalized data to 85.04% for min-max scaled data, and 86.74% for standardized data, which is the highest.

Figure 3 displays unnormalized data, where density values vary widely from approximately 0.988 to 1.002. Due to the larger scale of totalSulfurDioxide (up to 400) compared to density, differences in totalSulfurDioxide values may have a greater influence on the Euclidean

distance than density differences, potentially biasing the classifier towards totalSulfurDioxide.

Figure 4 illustrates min-max scaling, which linearly transforms the data points to fit within the range of 0 to 1 for both totalSulfurDioxide and density axes, ensuring a uniform scale across features, preventing features with larger scales from dominating the Euclidean distance. However, this scaling method is relatively more sensitive to outliers, especially when the outliers are the maximum or minimum value, such as the data point (0.4, 1), leading to a slightly lower accuracy compared to standardization.

In Figure 5, standardization is applied to center the data around a mean of 0 and scale it to have a standard deviation of 1. The data points are centered around 0 on the x-axis (density), with most falling within a narrow range, typically around -2 to 2. Similarly, totalSulfurDioxide is centered around 0 with a range from -2 to 2. This ensures that all variables contribute equally to similarity measures, making it less sensitive to outliers.

4.1 Model Extension

The accuracy of the Gaussian Naive Bayes model on the dataset is 77.78%, whereas the best performing 1-NN classifier with standardization achieved 86.74%.These findings indicate that the 1-NN classifier with standardization performs better than Gaussian Naive Bayes model on the dataset.

Moreover, there are 265 instances where the Gaussian Naive Bayes model and the standardization model disagree, with a percentage of 19.63%. For example, the first instance has a Gaussian Naive Bayes of 1, standardization prediction of 0, while the true label is 0.

```
Instances where the GNB model and the standardization model disagree:
[   1   10   19   20   24   27   28   30   36   55   58   61   65   67
   69   72   77   83   84   86   88   94   95   97   98  100  102  108
  110  115  119  122  124  125  142  143  144  145  149  159  179  180
  194  216  217  218  223  230  231  236  238  239  249  252  254  256
  258  259  262  263  267  269  273  277  278  285  295  299  301  318
  321  326  332  334  343  344  349  354  356  359  362  376  380  382
  385  399  400  412  413  414  421  424  430  434  438  442  444  448
  453  464  469  472  479  481  485  486  487  488  495  501  502  515
  527  531  532  539  542  560  561  564  574  575  576  579  580  581
  583  584  586  587  596  603  604  606  614  616  617  620  621  634
  636  637  640  641  648  672  678  680  682  684  697  699  721  725
  730  743  745  746  747  765  768  769  775  776  777  793  794  801
  808  813  815  821  825  827  832  836  838  854  867  881  889  896
  906  907  909  911  912  913  918  920  925  927  928  932  939  941
  942  945  949  953  970  983  989  992 1002 1013 1014 1017 1020 1028
 1029 1030 1032 1036 1037 1064 1065 1066 1067 1068 1069 1070 1071 1072
 1076 1083 1084 1086 1087 1088 1092 1106 1112 1121 1127 1135 1154 1155
 1157 1160 1168 1169 1180 1181 1189 1191 1199 1223 1224 1226 1254 1255
 1265 1266 1271 1274 1276 1280 1285 1302 1303 1310 1313 1316 1334]
```

The Gaussian Naive Bayes model assumes that continuous features are conditionally independent given the class label, which means it considers each feature independently when making predictions. However, in real-world datasets, features can be correlated or dependent on each other, which can lead to poor performance of the Gaussian Naive Bayes model.

Hence, the main reason for the disagreement between the two models is likely due to the violation of the independence assumption in the Gaussian Naive Bayes model. In Figure 6, several feature pairs such as ('residualSugar', 'density') and ('alcohol', 'density') exhibit high correlation, indicating a strong relationship between these variables. Violations of the normality and independence assumptions can impact the model's performance negatively, although Gaussian Naive Bayes tends to be relatively robust to minor violations of its assumptions.

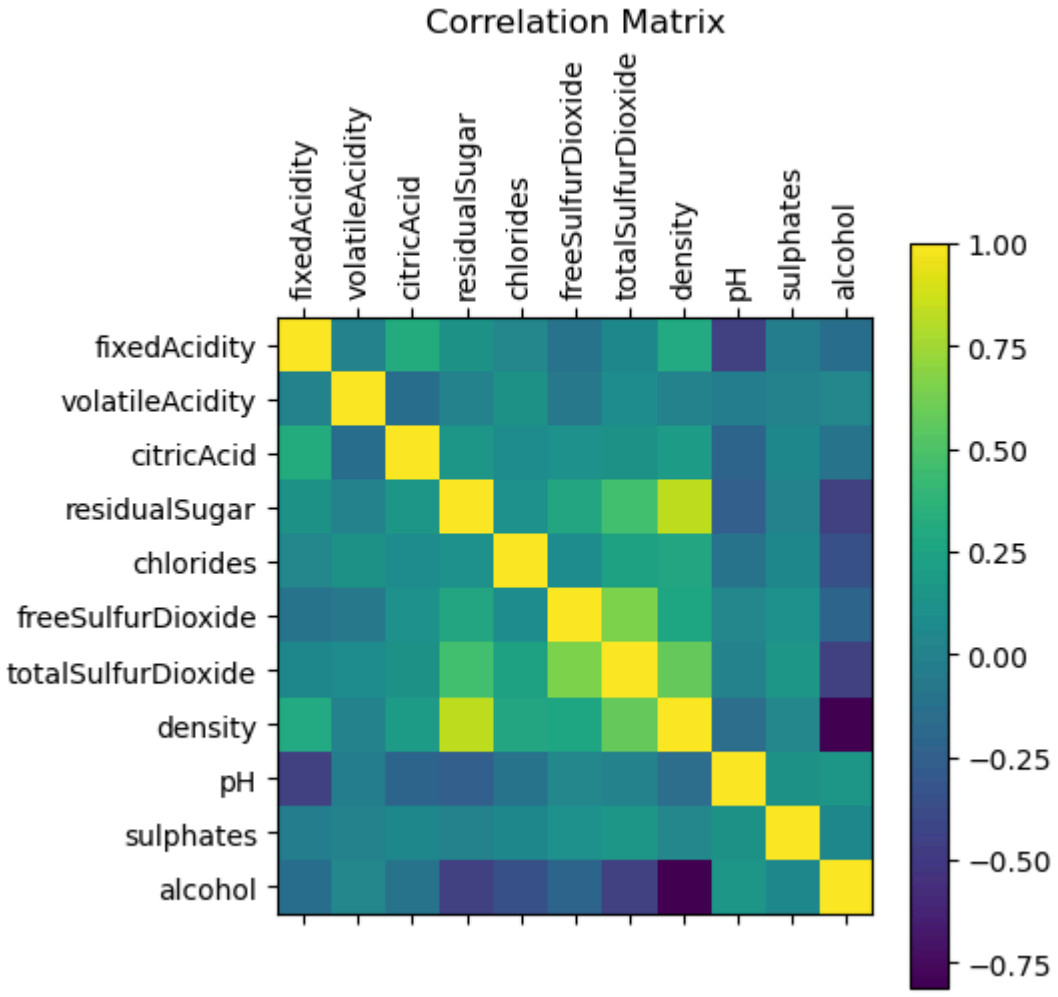

Figure 6

On the other hand, 1-NN with standardization does not make any assumptions about the distribution or independence of features. This makes it more flexible and potentially more accurate in cases where the feature independence assumption is violated, compared to Gaussian Naive Bayes model.