## 2.1

The accuracy of the 1-NN classifier is 76.44%

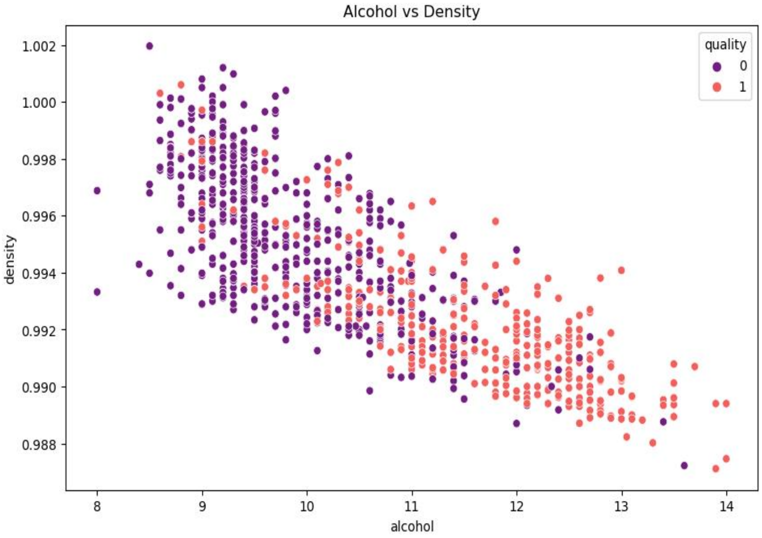


*Figure 1.1 Alcohol vs Density*

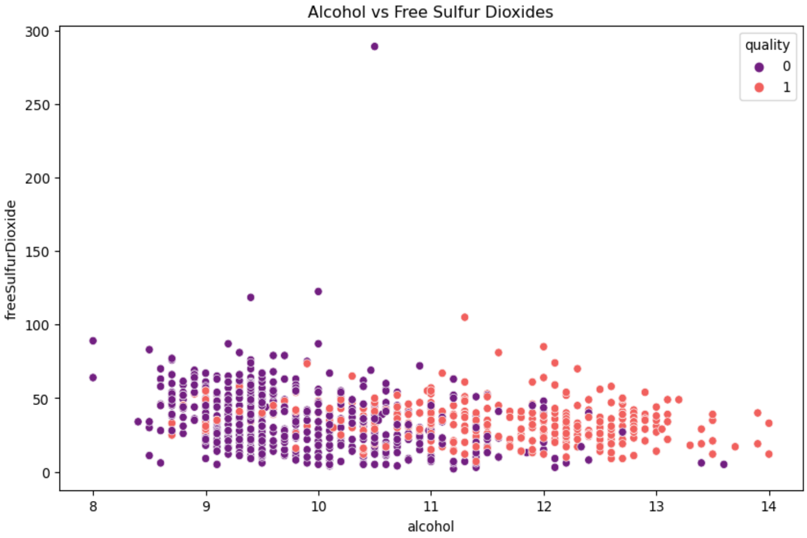


*Figure 1.2 Alcohol vs Free Sulfur Dioxides*

## 2.2

Based on Figures 1.1 and 1.2, the scatterplots show good class separation for quality 1 and 0, with dense clusters within distinct boundaries. Quality 1 instances cluster around alcohol values over 11 while quality 0 clusters are clustered around values below 11. Despite the overall good separation, there are instances where quality 1 appears within the mostly quality 0 cluster and vice versa. These differences could be due to the instances being mislabelled or inherently possessing characteristics of the opposite class which leads to potential misclassification and a decrease in accuracy.

A lack of outliers can also be observed, which is advantageous for K-NN classifiers which rely on distance calculations. Outliers can significantly affect the outcome of these classifications, potentially leading to choosing the wrong nearest neighbour and consequently a misclassification.

Feature scaling is also another important factor in K-NN classification which are sensitive to the range of feature values. In figure 1.2, the scale of free sulfur dioxide which ranges up to 300 while compared to alcohol's range of 8 to 14, could lead to disproportionately influencing distance calculations. Without proper scaling, features with larger ranges could dominate the classification process, overshadowing other important features.

Despite these problems, the dataset demonstrates a relatively high accuracy of 76.44% when doing 1-NN classification, indicating good suitability for it. The presence of strong class separation and dense clustering are important for accurate classification. However, the presence of misclassified instances and the use of feature scaling could be things to be considered that could lead to better classification as well as increased accuracy.

**3**

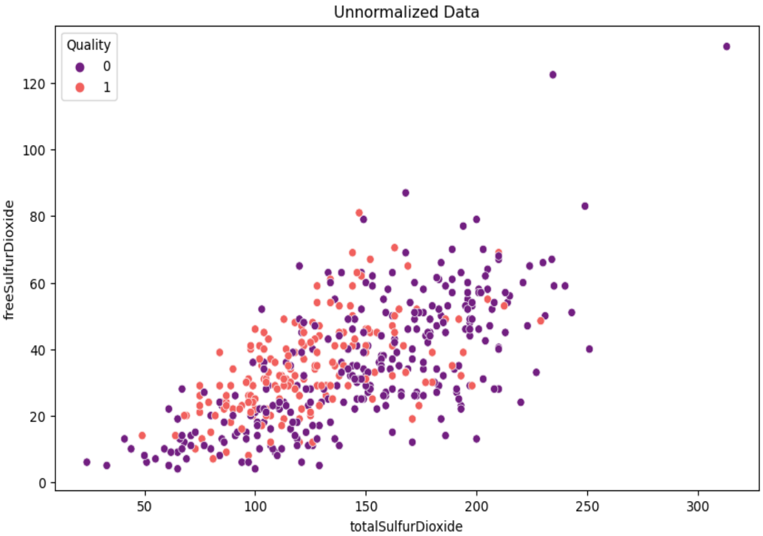


*Figure 2.1 Unnormalized Data*

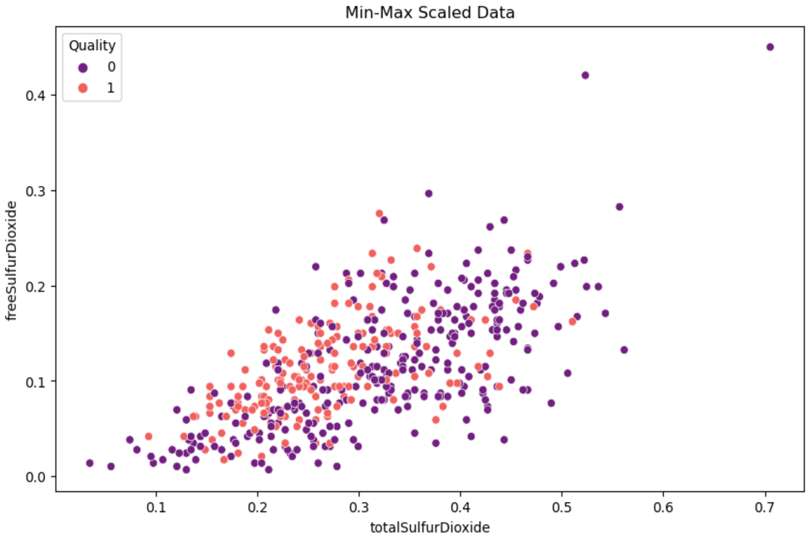


*Figure 2.2Min-Max Scaled Data*

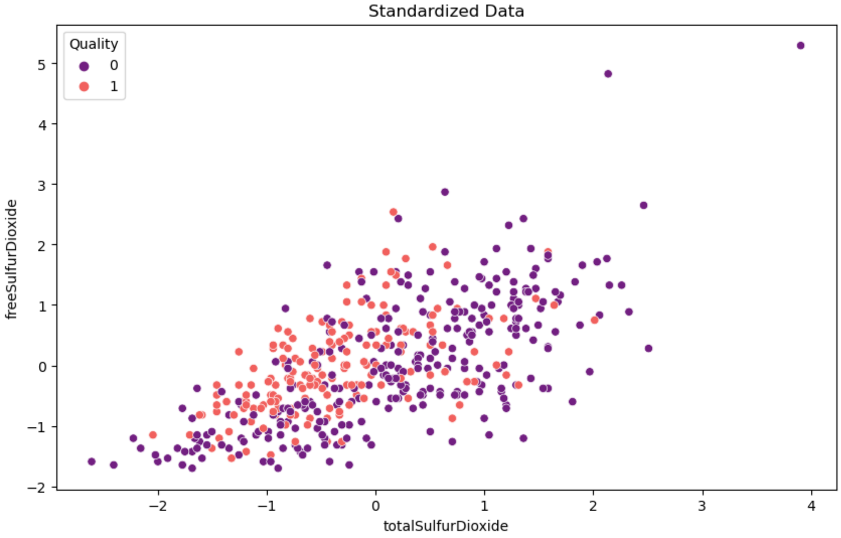


*Figure 2.3 Standardized Data*

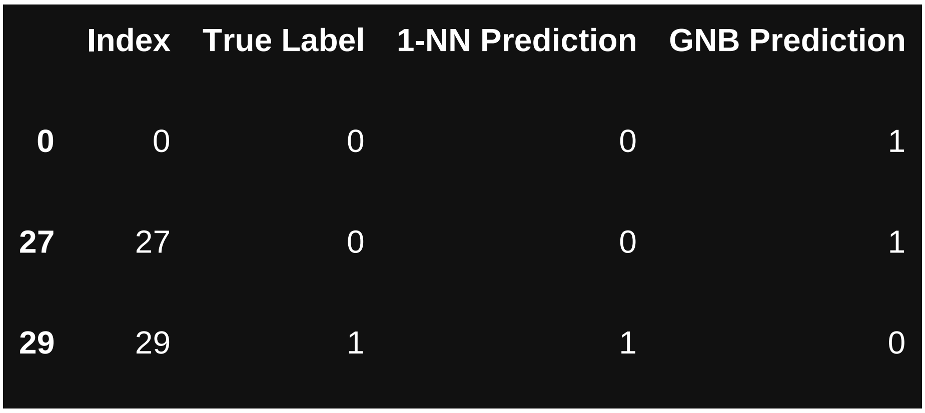
Based on the 1-NN classifiers' performance, we can see significant differences based on the various data normalization done. Initially, the raw, unnormalized data had an accuracy of 76.44% which is relatively high. However, when min-max scaling was applied, all feature values were brought to a uniform range between 0 and 1 and increased the accuracy to 85.04%. Then, we get another slight increase when using standardization, which adjusts features to have a mean of zero and unit variance, resulting in an accuracy of 86.74%.

We can see based on the figures above, while their plots pattern remains relatively similar, the scales of each plot vary with the new normalization. By observing figure 2.1, the unnormalized values of total sulfur dioxide as well as free sulfur dioxide range from 0 to 300 and 0 to 120, while they look to be of consistent range, they would dominate distance calculations when other features with smaller ranges are introduced such as alcohol and density from figures 1.1 and 1.2. This would then lead to potential misclassification of instances due to biased distance calculations. By applying the min-max and standardization normalization methods, we can see that the scales in the next two figures have been brought down to 0 to 1 as well as from -2 to 5, which would allow all features to contribute equally to distance calculations thus providing a more accurate classification.

The increase in accuracy highlights the importance of normalization in distance-based calculations. Without normalization, features with a larger numerical range can disproportionately affect the classifier, which could lead to wrong results. By scaling the features, we ensure that each feature contributes equally, thus leading to a more equal and accurate classification.

**4.1**

The accuracy of the best performing 1-NN is with standardized data at 86.74% while the Gaussian naïve Bayed (GNB) model achieved a 77.78%.



| | Index | True Label | 1-NN Prediction | GNB Prediction |
|---|---|---|---|---|
| **0** | 0 | 0 | 0 | 1 |
| **27** | 27 | 0 | 0 | 1 |
| **29** | 29 | 1 | 1 | 0 |

*Table 1 Disagreements between 1-NN and GNB*

We observed a few instances where the GNB model disagrees from the 1-NN and misclassifies the instance from the true label. While the 1-NN classifies instances through determining the nearest neighbour(s) with distance, the GNB classifies instances by applying Bayes' theorem with the "naïve" assumption that every feature pair is independent and normally distributed.

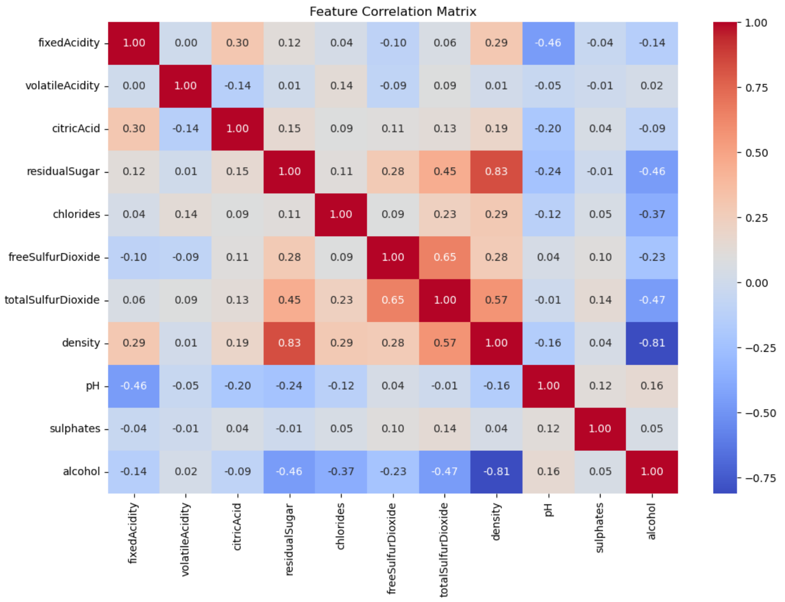
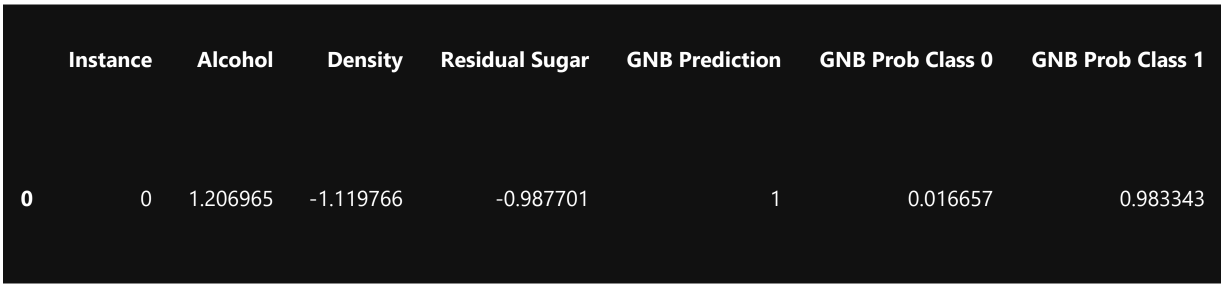


*Figure 3 Feature Correlation Matrix*

Based on Figure 3.1, the matrix shows a few feature pairs with strong correlations such as density/residual sugar at 0.83 and density/alcohol at -0.81. These high correlations between these features could mislead the GNB's probability calculations, while 1-NN is unaffected by this as it does not have any probabilistic assumptions.



| | Instance | Alcohol | Density | Residual Sugar | GNB Prediction | GNB Prob Class 0 | GNB Prob Class 1 |
|---|---|---|---|---|---|---|---|
| **0** | 0 | 1.206965 | -1.119766 | -0.987701 | 1 | 0.016657 | 0.983343 |

Table 2 Feature Values of instance 0 and GNB Prob Estimation of Misclassified Instances

The strong correlation between density/residual sugar (0.83) and density/alcohol (-0.81) would not have been accounted for by the GNB model due to the assumption of independence. This could lead to an overestimation of the probability for class 1 at 98.33% if both the correlated features were to suggest the same class. The 1-NN on the other hand most likely had a nearest neighbour of class 0. As it is mainly skewed by feature values and distance, these high correlations would have less of an effect thus leading to the true label prediction.