Question 2

The accuracy of my 1-NN classifier was 0.7644, which indicates an acceptable performance for a KNN model. The data provided was suitable for a 1-NN classifier as all the attributes were numeric and the attribute distribution for the classes showed good decision boundaries for some attributes. In Figure 1, we see that for the majority of scatterplots, there is a separation between low-quality alcohols and high-quality alcohols making 1NN classification simple. Furthermore, looking at the kernel density estimate plots we see that in the alcohol and density attributes, there are strong indications that there is a correlation between the values of the attribute and the quality of wine showing good decision boundaries.

However, the dataset also had some harder-to-distinguish boundaries for some attributes, seen in Figure 1 with fixedAcidity and residualSugar. This acts on the caveat of 1-NN in which it is susceptible to noise and may have led to wrong classification in some instances. In addition to this, as the dataset was large, a greater k-value may have been advantageous as it could reduce the impact of noise. Another aspect that made 1-NN difficult was the lack of normalisation, as the main measure was Euclidean distance this may have caused a bias towards one attribute over another due to greater variation in numeric values. Ultimately, the dataset was suitable for 1-NN classification but a greater k-value may have resulted in better accuracy.

Question 3

The normalised dataset improved on the accuracy of 1NN compared to 1NN with no normalisation. The accuracy for using the min-max normalised dataset was 0.8348, which was 7.04% better than no normalisation and the accuracy for the standardised dataset was 0.8622, 9.78% higher than no normalisation.

It was expected that using the two normalisation methods would improve the overall accuracy of the 1NN classifier as they are able to ensure that all attributes have the same scale. This heavily reduces the problems with bias towards certain attributes that have greater numeric variation thus leading to a more balanced model. In Figures 2 and 3, we can look at the axis and see the changes from the non-normalised data. The weighting between the different attributes is now equal across all of them while before each of the scatterplots had different axis values, another noticeable aspect is that the overall shape of the scatterplots is generally the same which shows that the information from the original dataset is still maintained. Furthermore, normalisation not only improves the decision boundaries when classifying points it also enhances the robustness to noise and outliers.

Standardisation performed marginally better than min-max normalisation with this dataset, this may be due to how standardisation preserves variation better than min-max normalisation as it does not limit the values to be between 0 and 1.

Question 4.3.

From the table below, first looking at cosine similarity compared to Euclidean distance, cosine similarity performed poorly compared to Euclidean distance. In contrast with the increase in accuracy when normalisation techniques were implemented with Euclidean distance, there was a 2.22% and 20.07% decrease in accuracy with min-max normalisation and standardisation respectively. This drop in accuracy may be due to how cosine similarity produces a greater value the more similar the vectors are and thus cosine distance (1 - cosine similarity) may have been a better distance measure as my implementation of KNN denotes the smallest distance values as the closest neighbours. The significant drop in accuracy with standardisation may also be due to this reason, as standardisation would make all the attributes more “similar” in general, this would increase the cosine similarity and ultimately increase the distance value when finding nearest neighbours.

Following this, we see Mahalanobis distance has a drastic improvement with no normalisation of data compared to Euclidean distance, this may be due to how Mahalanobis distance calculates the distance between a point and a distribution which may have resulted in less bias towards certain attributes in the non-normalised data. Min-max normalisation showed a decrease in accuracy this could be a result of how this normalisation does not preserve distribution well. Standardisation showed about the same level of accuracy as no normalisation and is comparable to Euclidean distance when standardising data, this is probably due to how standardisation preserves the distribution of the values.

	Accuracy for each distance measure		
Input data	**Euclidean distance**	**Cosine Similarity**	**Mahalanobis Distance**
No Normalisation	0.7644	0.6074	0.8570
Min-max Normalisation	0.8348	0.5852	0.7489
Standardisation	0.8622	0.4067	0.8548

Table 1. Accuracy of the 1-NN model for each of the three normalisation options and distance measures

Appendix

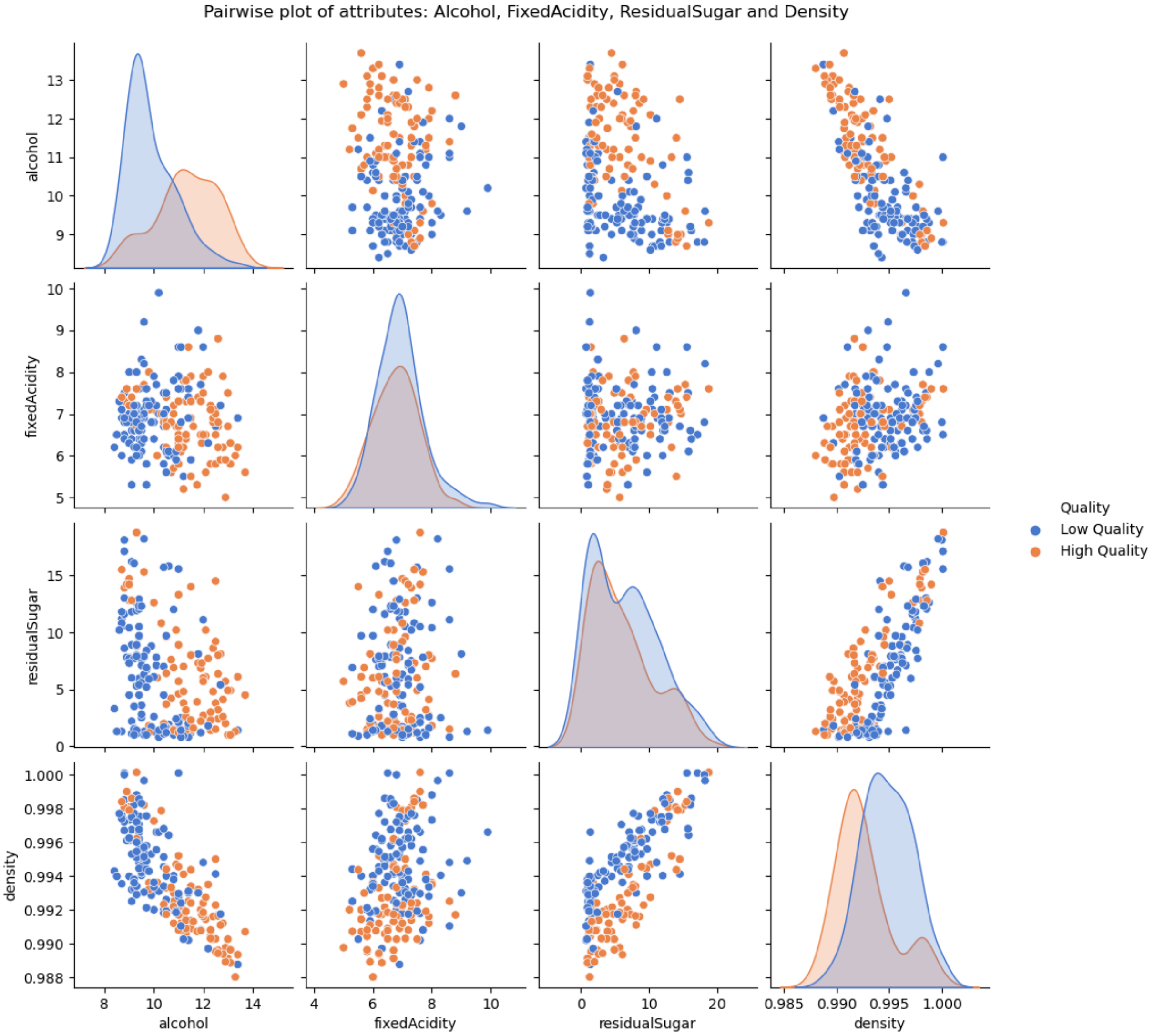

Figure 1. default_data_pairplot.png

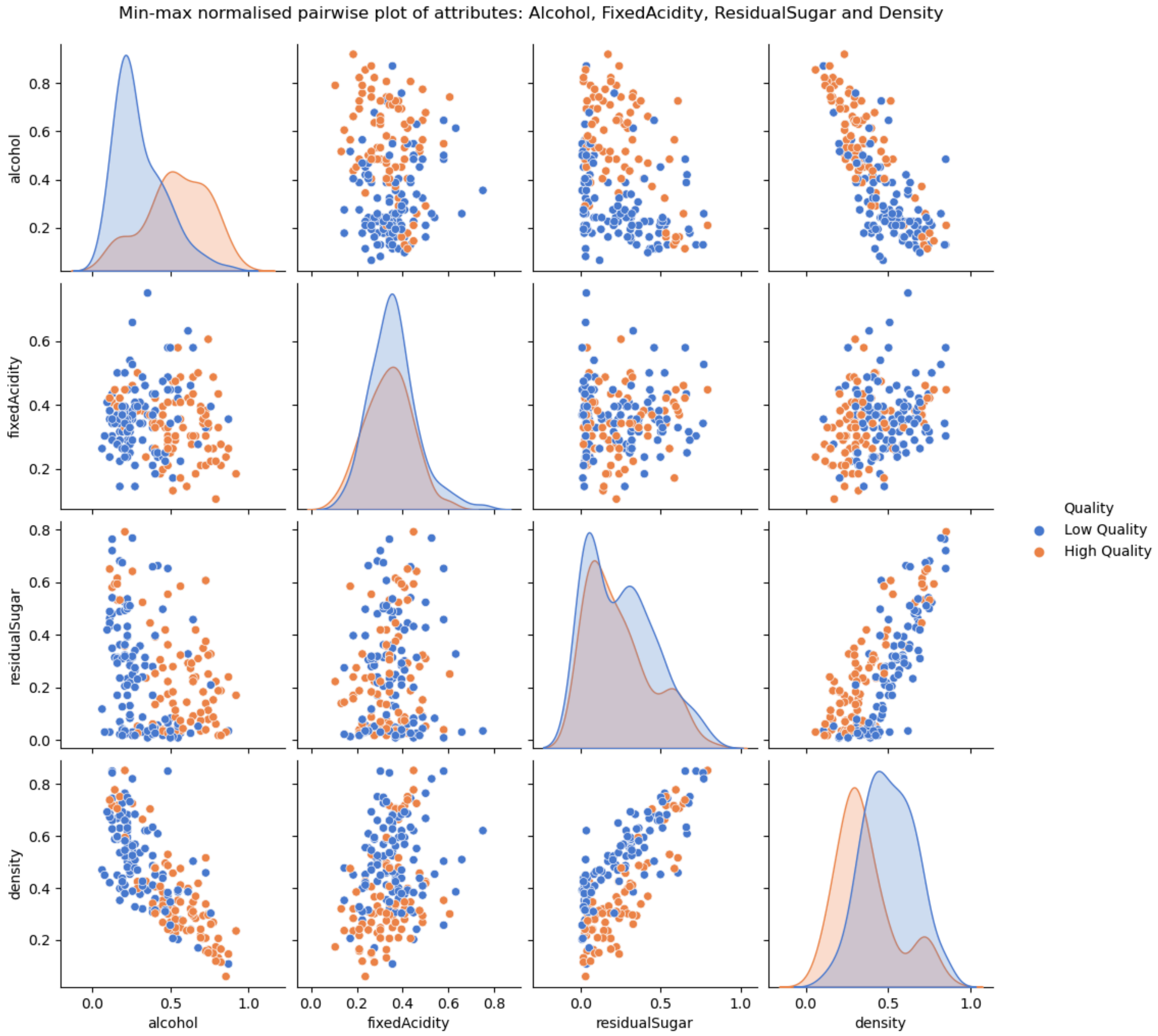

Figure 2. min_max_data_pairplot.png

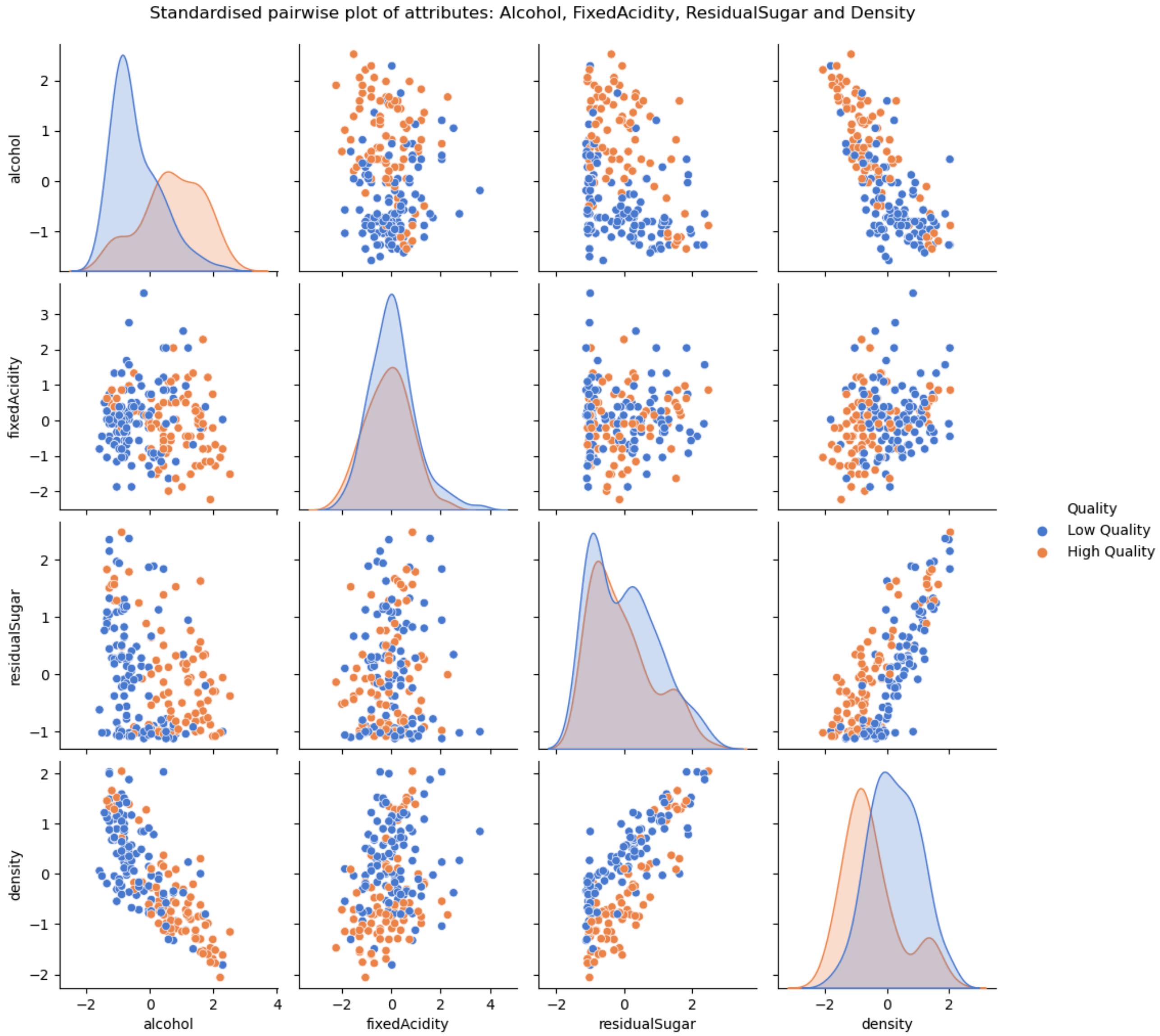

Figure 3. standardised_data_pairplot.png