COMP30027 assignment 1 2024

1266382

Q2:

1. The accuracy of 1-NN classifier on the wine dataset is approximately 0.764.
2. The dataset is quite suited for 1-NN analysis due to its entirely numerical nature, which lends itself to meaningful distance calculations. With binary labeling, the 1-NN algorithm benefits from a straightforward decision-making process, needing only to choose between two distinct options. Moreover, many features of the data, despite lacking explicit boundaries, somewhat divided into two groups across its features, makes the 1-NN process more accurate.

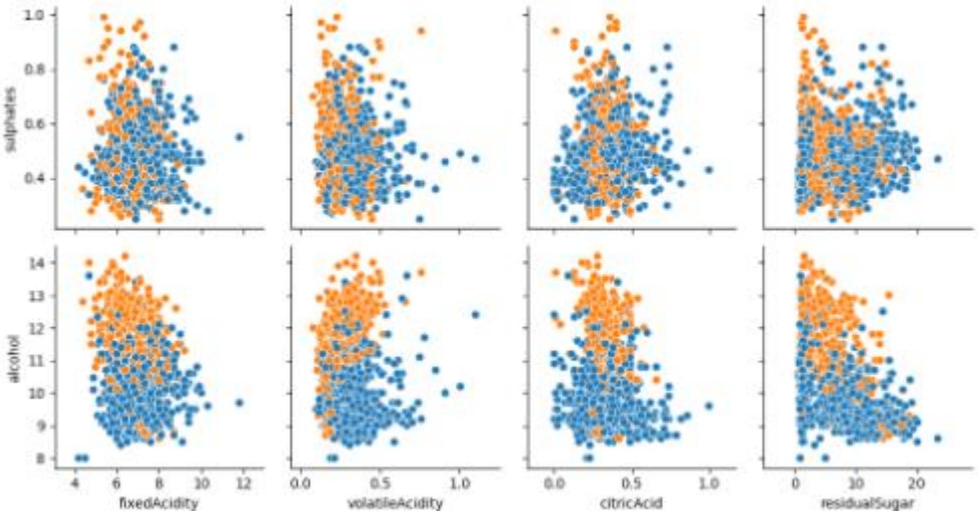

Figure 1: Part of the pairplot of training dataset.

With the help of dimension reduction method PCA, we can see the data has a 2 part characteristic.

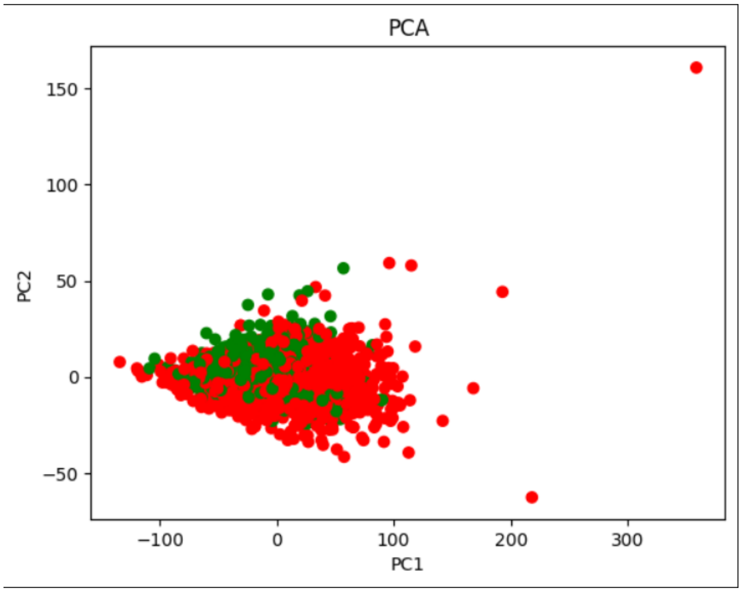

Figure 2: PCA of the training dataset without normalization

However, the problem of not having a clear boundary might cause some wrong predictions to the 1-NN, as new data that is close to the boundary of 2 class might get classified into wrong class, this is more of a problem for K > 1, as the overall density difference of the 2 class, one data might be place to the wrong class even it’s surrounded by the correct class.

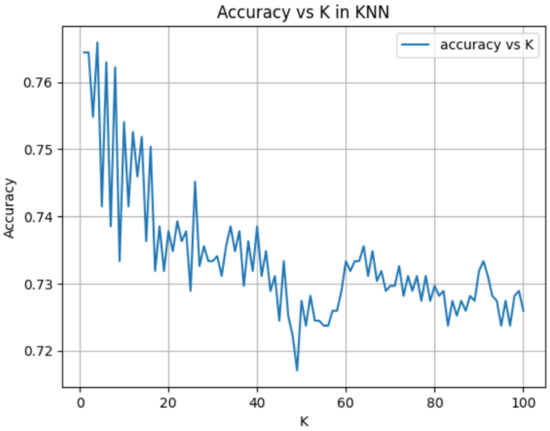

Figure 3: Accuracy vs choice of K in KNN classification.

Q3:

1. Comparing the accuracy of 1-NN on original data vs normalized data, the min-max normalization achieved accuracy of approximately 0.850 and standardization achieved accuracy of approximately 0.867. the normalization process does make the 1-NN performed better.

   The features of the wine dataset are on different scales; for example, the 'volatileAcidity' feature values range from 0 to 1, while the 'totalSulfurDioxide' feature values extend into the hundreds. This could mean that when implementing a KNN, which calculates the 'distance' between data points, a feature with larger values will weigh more than a feature with smaller values, regardless of their actual importance. Therefore, the normalization step is crucial to ensuring that the distance calculation considers all features equally.

   The following PCA also shows a clearer division of class after normalization.

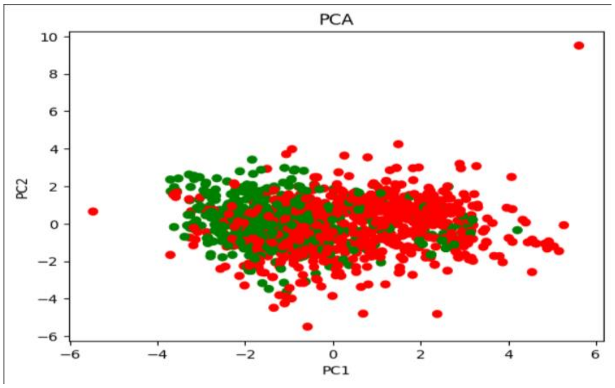

Figure 4: PCA after standardization.

This, however, would not change the overall distribution of the relation between the features, as the relation of features would remain the same after normalization, only the scale will change. Comparing figure 1, and following figures.

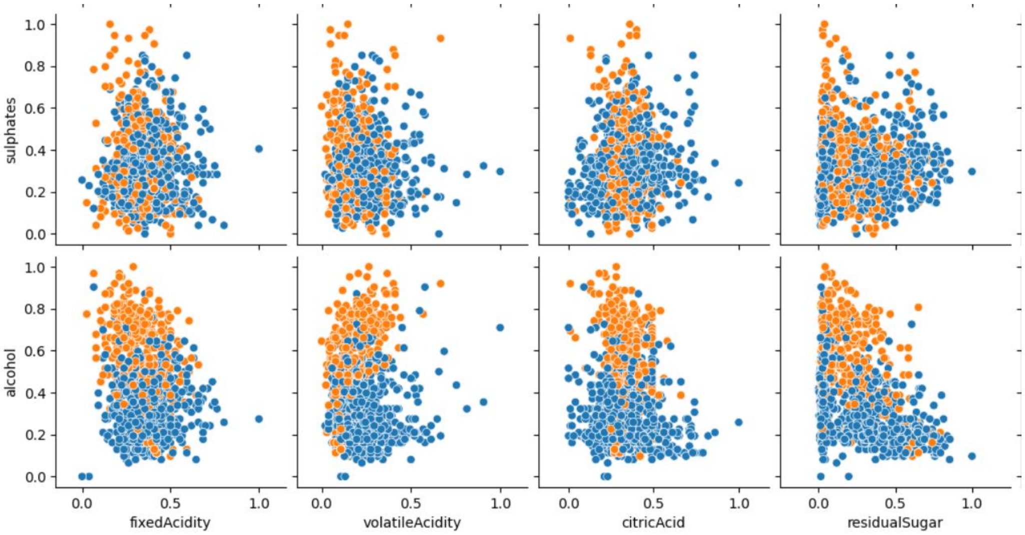

Figure 5: Part of the pairplot of training dataset after min-max normalization.

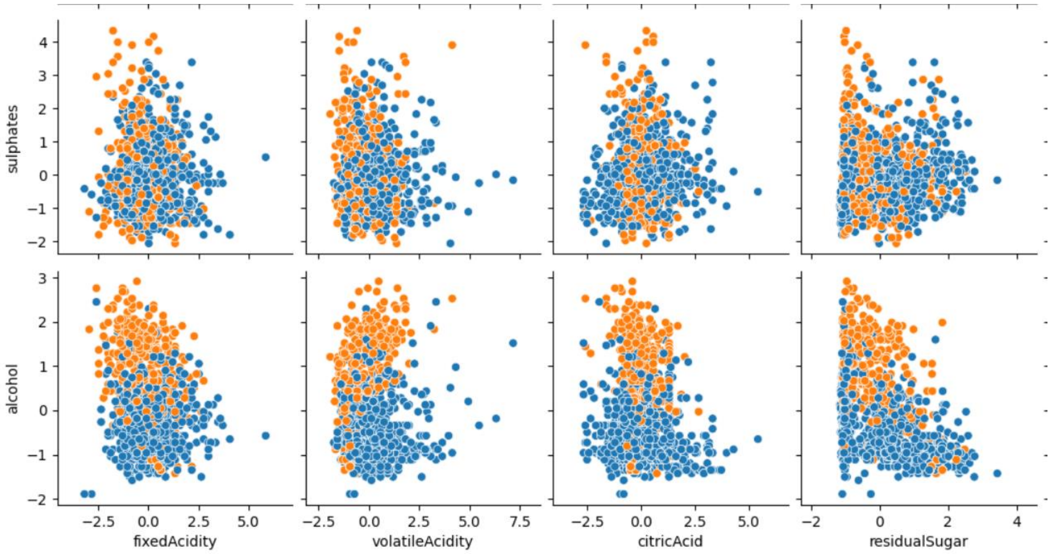

Figure 6: Part of the pairplot of training dataset after standardization.

Q4.4:

1. The training dataset contains 820 class ‘0’ which indicates low quality and 530 ‘1’ which indicates high quality. The choice of K in the KNN classification in this case, would not be very helpful if K is quite large, especially if K > 1060, as there are more class ‘0’, all unseen data will be predicted to class ‘0’.

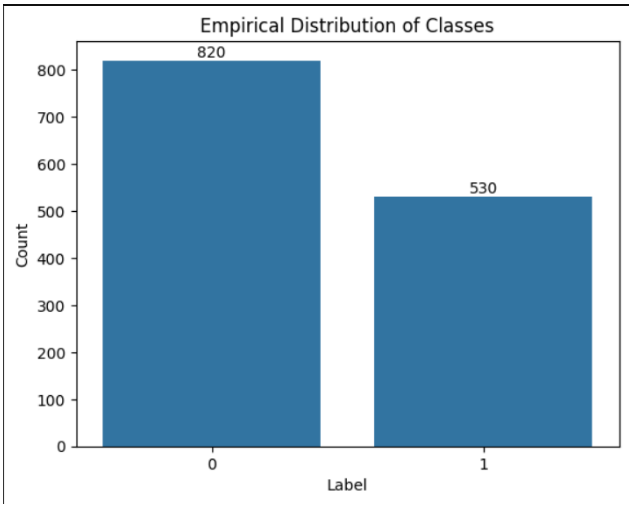

Figure 7: Frequency of each class in training data.

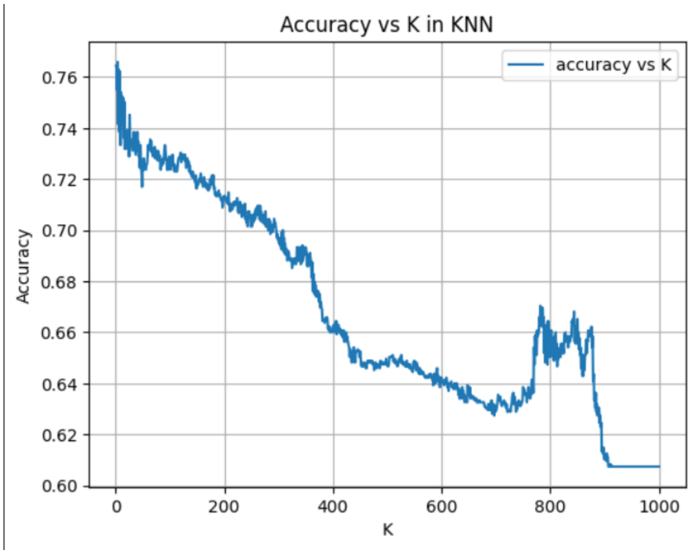

Figure 8: Accuracy of KNN model vs choice of K, unnormalized.

The trend of accuracy going down as increase of K indicates that the data are very close and no clear boundary or cluster pattern in the dataset, the spike in accuracy when K approaches the number of data points of the majority class(class ‘0’), is possibly due to testing dataset also contains majority of class ‘0’, as the model would correctly predict the majority class but not accurately predict the minority class.

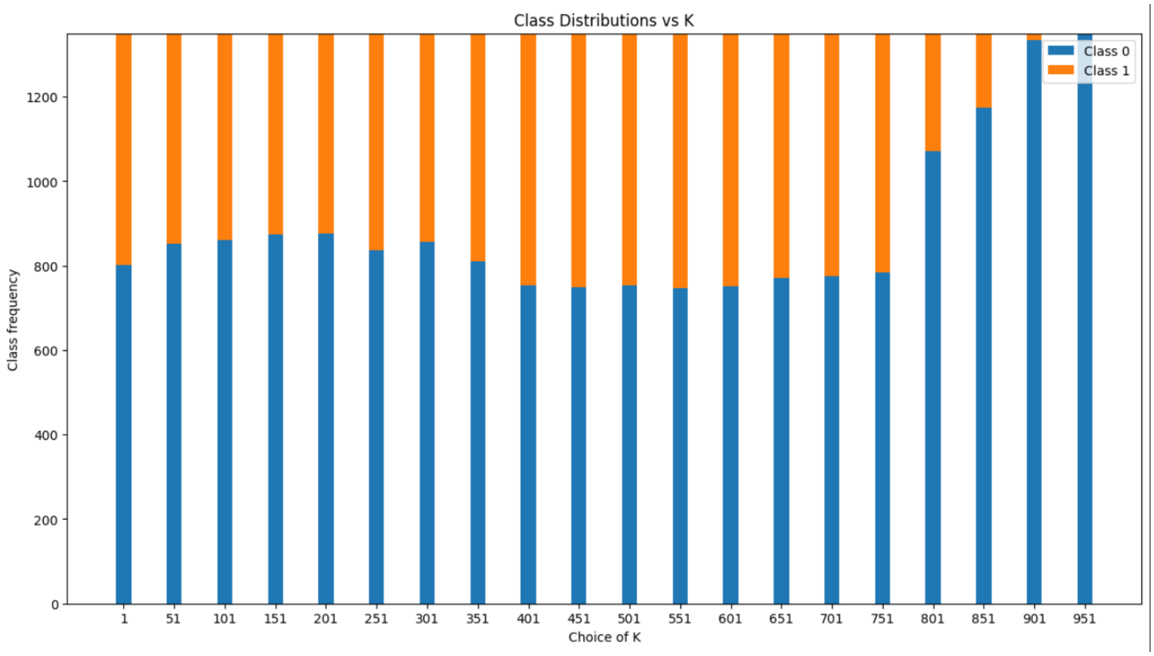

Figure 9: Distribution of class in different choice of K.

This can be further supported by the distribution chart, when K at small value, the class distribution of the predicted label matches the class distribution in training data. However, as K approaches and exceeds the number of class ‘0’ in training set, the model gives more prediction label of class ‘0’, due to the number of class ‘0’ dominates class ‘1’ in the training set, giving out these false prediction.