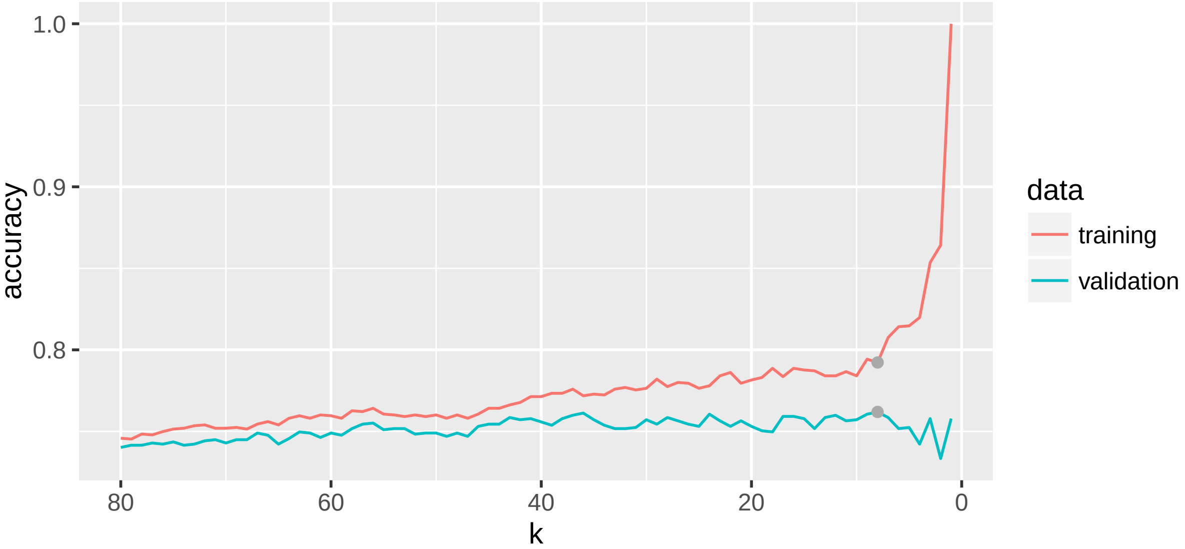
Assignment 1 Group 1 Question 2a

Group 1: Siow Meng Low, Louise Fallon, Nikhita Venkatesan, George Pastakas, Cecilia Nok Sze Cheung, Steven Locorotondo

To separate the data set into the training, validation and test sets, we first split into a training set (1959 observations, ~40%) and an others set (2939 observations, ~60%), using a random sample of the observations, then we split the others set half and half into a validation set (1470 observations, ~30%), and a test set (1469 observations, ~30%).

The training and validation set data is used to provide the parameters (means and sample standard deviations) to the z-score function, which will then be used to scale the training, validation and test data.

When we train and assess the model for k between 1 and 80, we get the following results:



Assuming that true positives and true negatives are equally weighted in this case, we use accuracy as the measure for comparing models. If detecting true positives were more important, we would have used sensitivity and if true negatives were more important we would have used specificity. We can see that using this measure, the best k is 8, with an accuracy of approximately 76% on the validation set.

The model with the best k is then retrained with the data from both the training and validation set, this will more accurately reflect the out of sample error than if only the training set was used, because we will eventually train the model on the full sample of data to maximise its performance. The estimated out of sample performance of this model (estimated using test set) can be seen in the table below:

Table 1: Generalisation Error

k	accuracy	sensitivity	specificity
8	0.7603812	0.8655031	0.5535354

The generalisation error can also be represented using estimation misclassification rate (i.e. 1 - accuracy): 0.2396188

The confusion matrix of the test result is shown below:

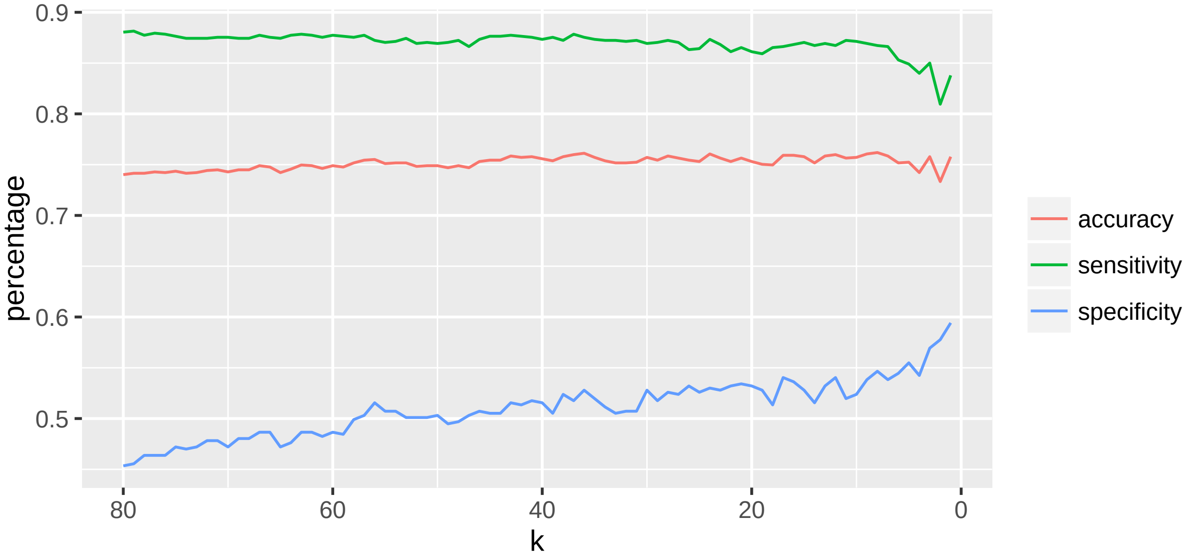
Table 2: Confusion Matrix

	Predicted.Good	Predicted.Bad
ActualGood	843	131
ActualBad	221	274

As a final step, we then train the model with the best k with all the data (training, test and validation), to be used for predictions of future data.

How do you judge whether the classifier is well-suited for the data set?

- From the graph we can see that for 5 < k < 80, the accuracy seems relatively constant, so the choice of k does not make a huge difference. This result is robust to different seeds and therefore different test/validation sets. A different choice in seed generally leads to a different optimal k, and to a similarly flat line for 5 < k < 80. This implies that the selection of k may not have a huge impact on accuracy of prediction for future data points.
- A naïve (majority-based) predictor would predict a good quality wine in all cases, because the proportion of wines of good quality is approximately 66% in the training set. If this predictor was used, then it would have an accuracy of 66% (estimated using test set). The estimated out of sample error for our knn classifier, using the best k, is 76% (in terms of accuracy), even though this is a definite uplift, it is still not a highly accurate predictor.
- Our classifier is more sensitive (predicting true positives given a true result), than it is specific (predicting true negatives given a false result), with specificity decreasing with k, so if it were true negatives that we were more interested in, we may want to choose a lower k.
- The knn classifier is useful for prediction, but is not useful for inference, as it doesn't provide any insight on mechanisms or inputs that have an impact on the target value, i.e. it doesn't tell us what kind of impact the different properties have on what makes a wine of good quality.


percentage
k
accuracy
sensitivity
specificity