CSCI218 Week 2 Lab

In lab 2 a dataset was given containing measurements of three different types of flowers. In total 150 data elements were recorded with 50 data elements for each flower. The data should then be divided into two sets of an equal number of elements of each flower. This was supposed to be done randomly, I chose to divide the data sets by even and odd indices. The training set is then used to compare new flower measurements to. So, if a new flower measurement is closer to a greater amount of the Setosa flower compared to the others, it will be categorized as a Setosa flower. This will be done by determining the k-nearest-neighbour. So, for a given flower the number of neighbours within a k-radius will determine which flower group it will be assigned as seen in Figure 1.

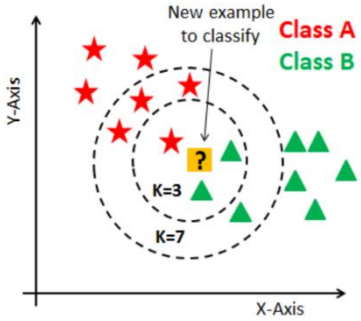

Figure 1: K-nearest neighbour

To implement the k-nearest-neighbour algorithm the Euclidean distance is calculated as a measure for the distance to a neighbour. This is calculated for each flower in the training set and compared to the given k value. If a flower is within k, it is noted. After looping through all flowers, a list is obtained with the amount of neighbours for each flower. Now the new flower can be classified as the flower with the most neighbours. After determining the k-nearest flower it is compared to the actual flower and hereby the accuracy of a given k-value can be determined.

By plotting different k values with their respective accuracy with matplotlib, it can be seen how k changes the accuracy.

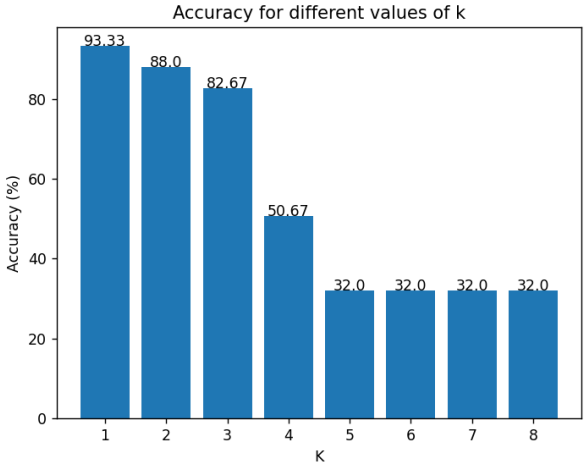

Figure 2: Accuracy for k values from 1-8

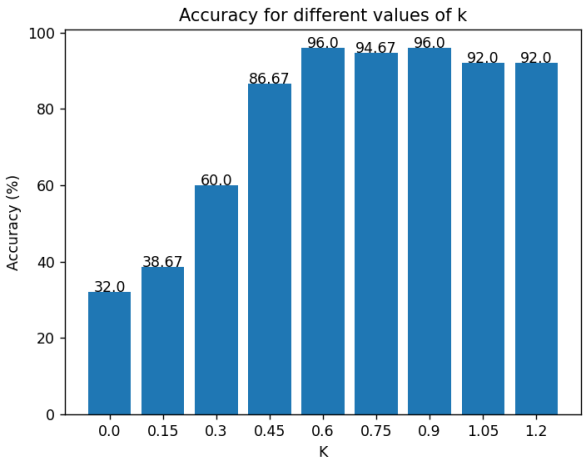

Figure 3: Accuracy for k values from 0.0 - 1.2

It can be seen in Figure 3 that a k value of 1 is by far the best performing one. Therefore, a closer look is taken on k values near 1. Here it can be seen that the best performing k values ist just under 1. However, it cannot be said for certain which k value would perform best if the training data was randomized differentl.

A way to improve this further would be to find the mean of the best performing k values where the training and data sets were randomized differently.