Project 1

k-Nearest Neighbor

**Problem Description:**

Given a dataset representing body length and fin length of two new species of fish (ie. Tiger0 and Tiger1), develop a kNN classification algorithm that will predict whether the fish belongs to Tiger0 or Tiger1 category.

A screenshot of a cell phone

Description automatically generated**Data Description:**

The training data consisted of 300 records representing two features of either Tiger0 fish or Tiger1 fish. Each row has three-tab separated entries. The first is a floating-point representation of the body length in centimeters, second one is the fin length in centimeters and the third column indicates the class of fish it belongs to. A plot of the data is shown in Figure 1 where green is represented by Tiger1 fish and red is represented by Tiger0 fish.

Figure 1. The Initial Data Set

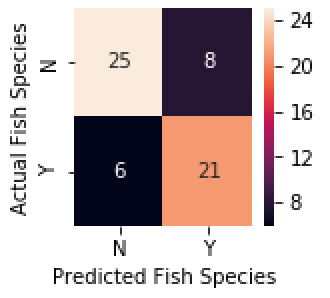
**Training of kNN Algorithm:**

We have developed a k Nearest Neighbor algorithm using a 5-fold Cross Validation. First the length of the input file was calculated, then we shuffled the records to make it random. Out of the total 300 records, 240 records (i.e. 80% of the total) were taken as Training Set and the rest 60 records (i.e. 20% of the total) were taken as Test Set. The Training Set was further divided into 5 smaller parts with 48 records each (i.e. for 5-fold cross validation). The five folds were used to create five smaller training sets of four folds (192 records) each, with the leftover (48 records) in each case used as validation set. Each training set was then executed via kNN with odd values of k of 1 to 30. For each value of k, the number of misclassifications were recorded for all five validation set combinations (Figure 2). From this computed information, the accuracy of cross validation was plotted for each and every value of k from k=1 to k=30. From the figure 3 we can see that the highest accuracy is achieved when the value of k=7.

**A screenshot of a cell phone

Description automatically generated**

Figure 2. Misclassifications for different values of k on 5 sets of training

A close up of a map

Description automatically generated

Figure 3. Accuracy for cross validation set Figure 4. Confusion Matrix

**Results:**

A Confusion matrix for the results of Nearest Neighbor algorithm with k=7 is shown in form of heat map (refer Figure 4). In the test set, 23 record were correctly identified as Tiger0 fish and 25 were correctly identified as Tiger1 fish. The results achieved an accuracy of 0.90 and precision of 0.92. Recall that is how many of the found were correct hits was equal to 0.89. The overall F1 score of Project1 is 0.90.