Project 3: Classification with k-Nearest Neighbor

Problem Description:

Given a data set of 118 examples to predict whether capacitors from a fabrication plant pass quality control based on two different tests. The kNN classification algorithm was used to determine if the capacitor has passed the quality control test or not.

Data Description:

The dataset has been split into training set with 85 examples and testing set with 33 examples with both results from the two tests by quality board. Each record had three tabbed separate entities depicting the test 1 result, test 2 result and the result of whether it has passed or failed in that respective test (row). The data of the test 1 and test 2 values are in the datatype of float (%0.5f).

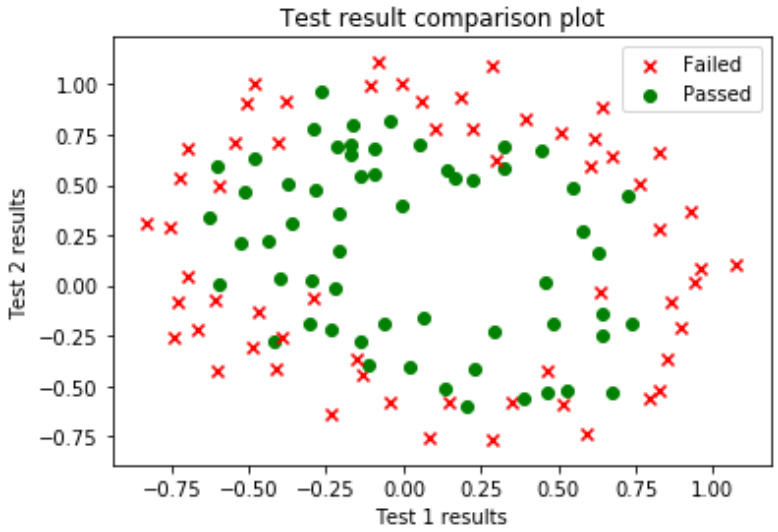

Figure 1 Plot of Initial dataset

Training a kNN algorithm:

A k Nearest Neighbor algorithm was developed using 5-fold Cross Validation. The data given was randomized and the last column which shows the result of whether the test has passed or not was also given. 85 examples were split into five folds and these were used to create five smaller training sets of four folds (68 records) each, the leftover fold (17 records) in each case used as the validation dataset. Each training set was then executed via k-NN with odd values of k of 1 through 23. For each value of k, the number of misclassifications were recorded for all five training and validation set combinations. It can

k	1	3	5	7	9	11	13	15	17	19	21	23
Test1 errors	8	6	4	4	5	5	6	7	6	6	8	8
Test2 errors	7	2	3	3	3	3	3	5	7	7	7	6
Test3 errors	7	3	3	4	4	7	5	5	7	8	7	8
Test4 errors	9	6	3	2	5	5	6	6	8	7	11	11
Test5 errors	8	5	8	10	10	9	7	7	8	9	10	11
Total	39	22	21	23	27	29	27	30	36	37	43	44

Figure 2 Misclassifications for different values of k on the five training sets

be observed from the accuracy values that the result accuracy reduces k = 15, since the data is less and the k- value increases which tries to underfit. From this data, the cross-validated accuracy was plotted for each value of k. k = 5 provided the best accuracy and was chosen for the kNN to use on test dataset.

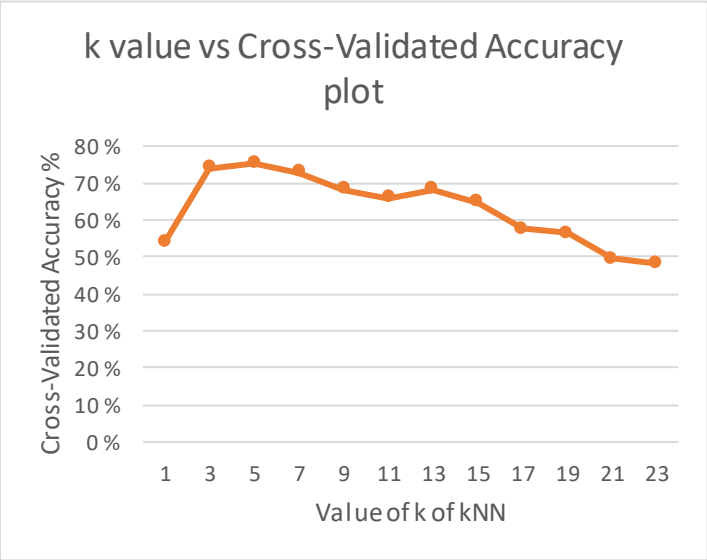

Figure 3 Average accuracy for different values of k

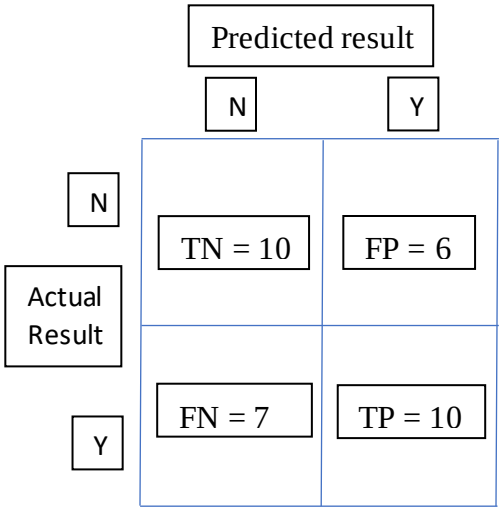

Figure 4 Confusion Matrix

Results:

A confusion matrix was created for the results of the Nearest Neighbor algorithm with k = 5 and is represented above. The test dataset consisted of 33 records representing test 1 and test 2 data records with the result of either passed or failed in the quality board testing. Out of the 33 data records in testing dataset, 20 records were correctly identified for an accuracy of 0.606. Precision was equal to 0.625 which means that the classifier algorithm was correct 62.5% of the time during testing. The recall value was 0.582; seven passed test records were misidentified as negative result. The overall F1 score was 0.606.