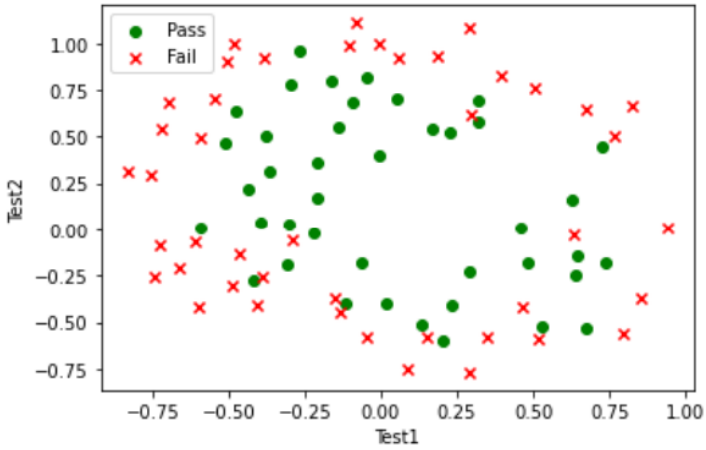
Problem Description

This program predicts whether capacitors from a fabrication plant pass quality control based on values from two different tests.

Data Description

The given data has 118 examples of labeled data points. I used a 5-Fold Cross Validation strategy for this program, so I split the data into 5 sets with each set being divided randomly into a 1:4 ratio of test to train data. The plot of the whole data set with respect to both of the test scores where the green represents the passed examples and the red represents the failed examples is shown in the figure to the right.

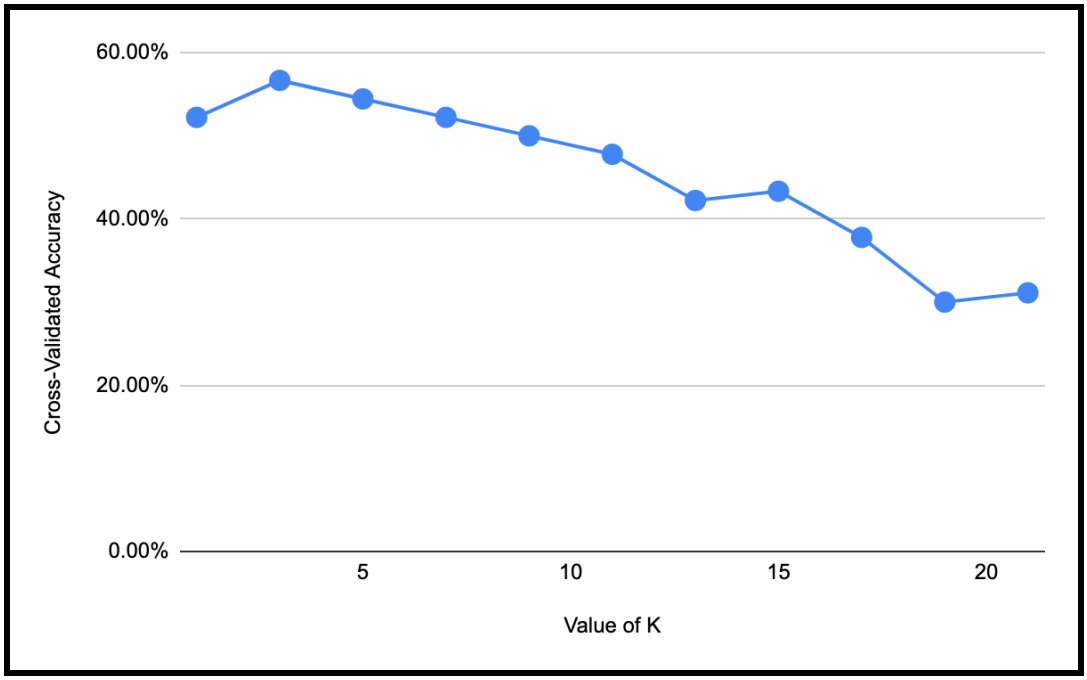


Training a KNN Algorithm

For each of the 5 folds described above, I checked the accuracy (number of incorrectly identified examples) with different K-values. A K-value determines the number of nearest neighbors to the specified test data point that are considered when classifying the data point. I tested K-values from 1-21. I picked the K-value with the lowest total incorrectly identified examples over all the training sets (Folds 1-5) to use in the final program. The number of total incorrectly identified examples is plotted in the table below.

K	1	3	5	7	9	11	13	15	17	19	21
Test 1	6	6	6	6	6	6	6	6	6	11	11
Test 2	9	8	8	8	8	8	8	8	8	8	8
Test 3	11	6	6	6	6	6	6	6	6	6	6
Test 4	9	9	9	9	9	9	9	9	9	9	9
Test 5	7	7	7	7	7	7	10	7	10	10	7
Total	43	39	41	43	45	47	52	51	56	63	62

This table shows how a K-value of 3 is shown to be the most accurate with the least total of incorrectly identified examples. This accuracy is represented as a percentage correct of the total test examples in the graph below.



Results

A Confusion Matrix for a KNN value of 3 is shown to the right.

In addition to the TN, FP, FN, and TP values, the algorithm produced an accuracy of 0.9, a precision of 0.95, a recall of 0.9, and a total f1 score of 0.92.

		Predicted Passed	
		N	Y
Actual Passed	N	TN = 12	FP = 1
	Y	FN = 2	TP = 18