Project 2: Classification with K-Nearest Neighbors

May Neelon

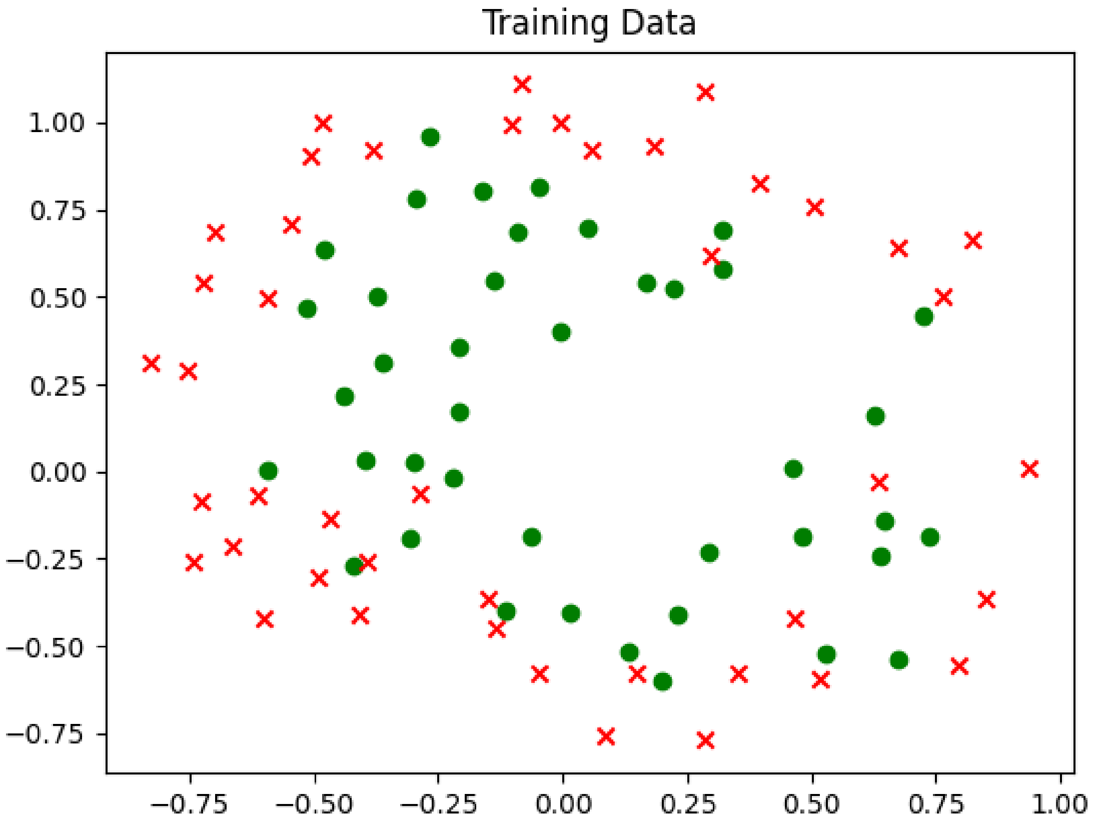
Problem Description

For this project you will develop a kNN program to predict whether capacitors from a fabrication plant pass quality control based (QC) on two different tests. To train your system and determine its reliability you have a set of 118 examples. The plot of these examples is show below where a red x is a capacitor that failed QC and the green circles represent capacitors that passed QC.

Description of the Data Set

The data set is a collection of capacitors described by two features. Some of the capacitors passed inspection, while others failed. A plot of the data shows that the successful capacitors are all clustered in the center, while the failed capacitors are on the outside.

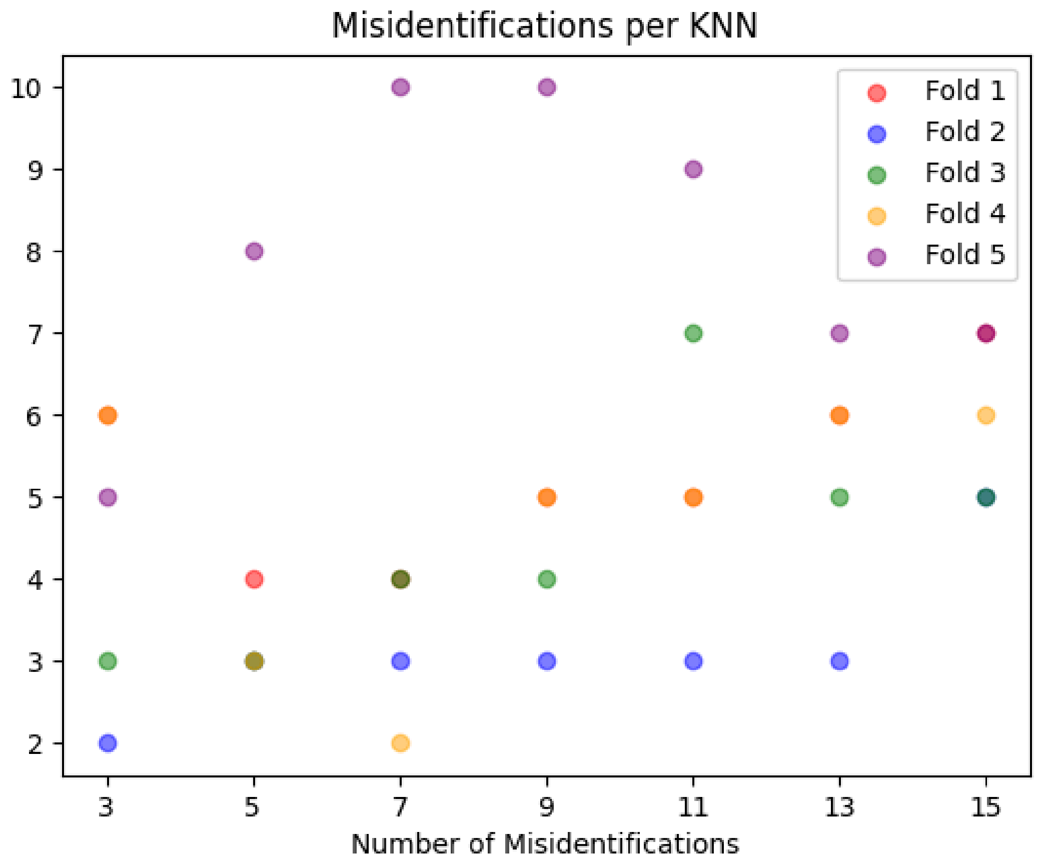
A plot of the training data:



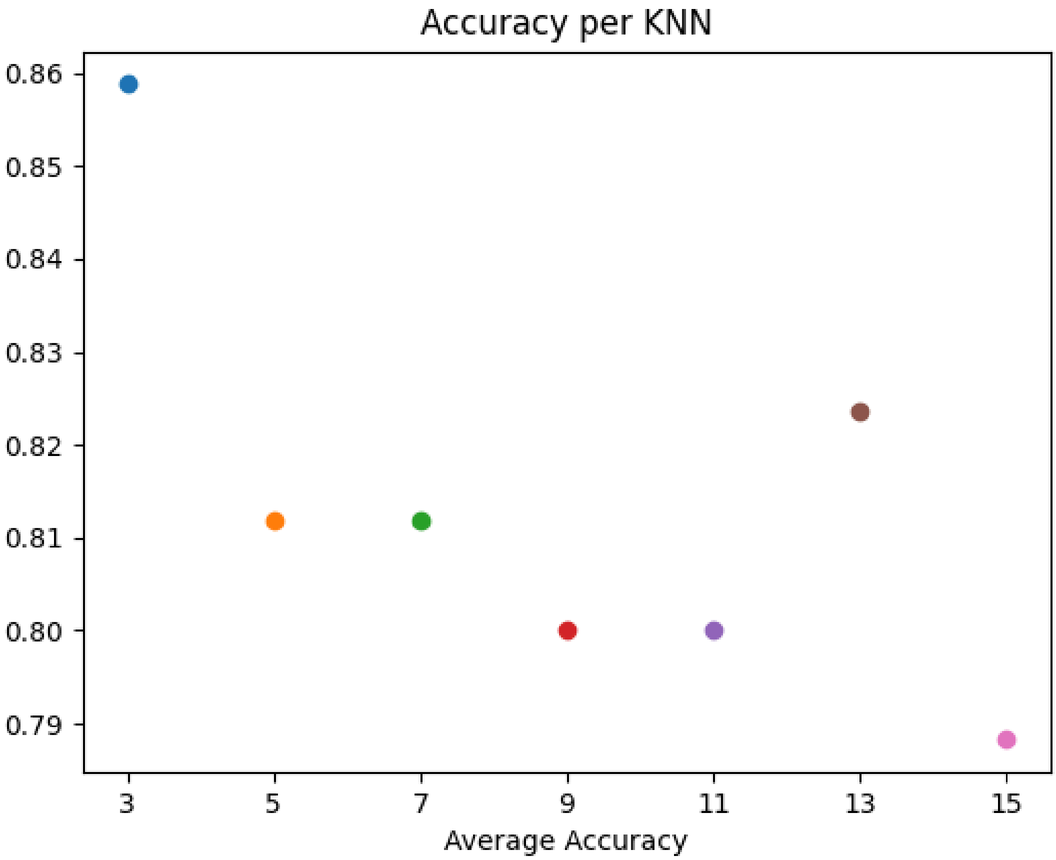
Procedure

My procedure was to run the KNN algorithm for k=3 to k=13, focusing on the F1 scores for the full training and testing set. This led me to conclude that the most optimal choice is k=3, as it gives the highest results for both of the features. Obviously, more attention could be given to the individual metrics, but I think it's better to look at the overall picture to make the final decision.

Misidentifications per K:



Accuracy per K:



Confusion Matrix Metrics for k=3: Accuracy: 0.636, Precision: 0.647, Recall: 0.6478, F1: 0.647

Final Results

Overall, based on my accuracy results, I chose k=3 for my model. Running it on the test data set gives metric values of:

Accuracy: 0.636, Precision: 0.647, Recall: 0.6478, F1: 0.647

I believe that this is the most optimal choice for K given the datasets