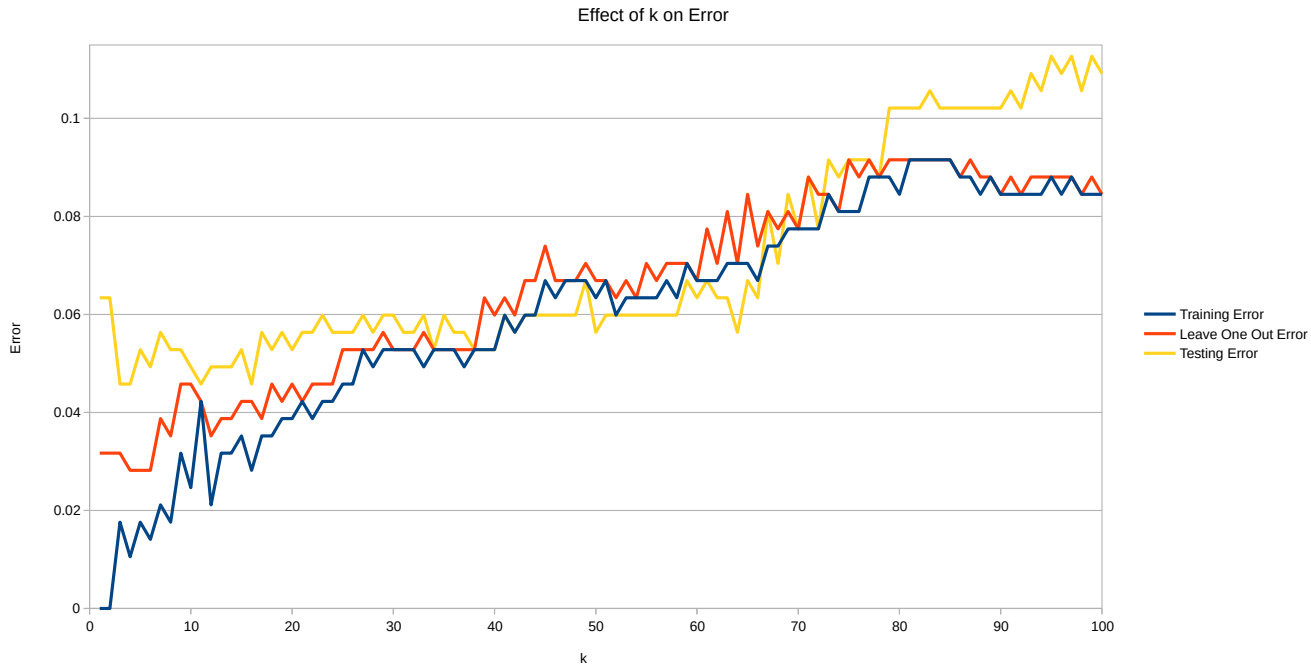
Implementation Assignment #2

Trevor Hammock, Charles Koll

Part 1

1. Implement the K-nearest neighbor algorithm, where K is a parameter.
2. Plot K values 1,3,5,..,51 showing training, leave-one-out cross-validation, and testing error



3. Discuss your observations between the three errors and perform model selection

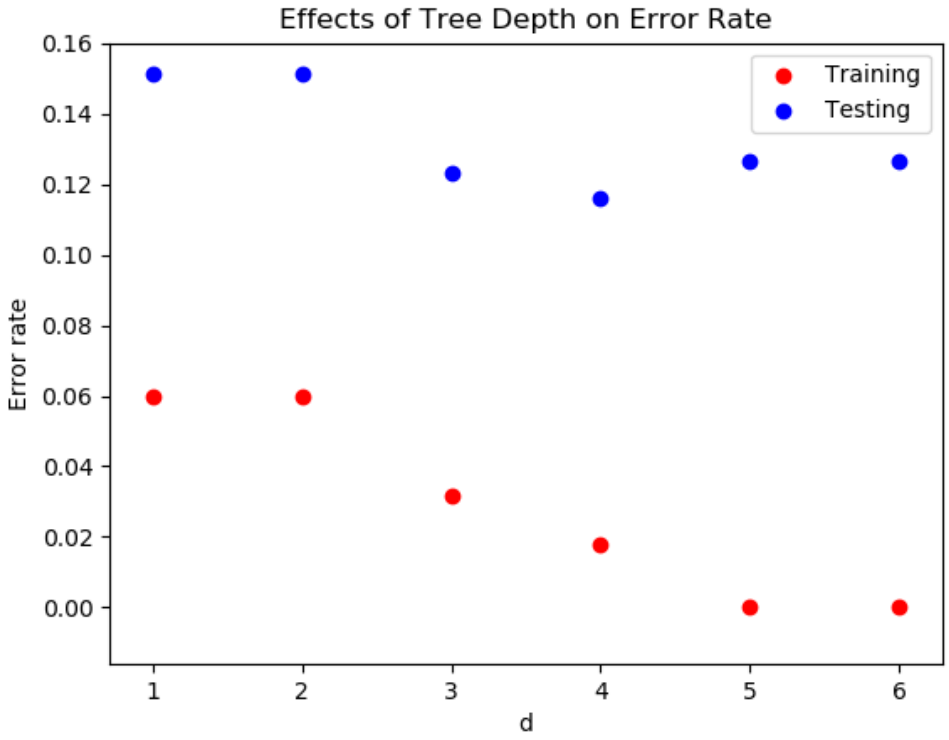
As k increases, generally the three different errors also increase. The trends for leave-one-out cross validation is very similar to the training error. The training error is guaranteed to be equal to or less than the leave-one-out cross validation. The testing error increases very slowly til about k=65, where the slope dramatically increases, which is most likely coincidence. The testing error occasionally can dip below the leave-one-out cross validation and the training error, but this is also a coincidence. For model selection, you generally want to choose k values that produced the lowest cross-validation error. According to our test results, k values between 4 to 6 produced the lowest cross-validation error (2.8%). Additionally for those range of k values, we also observed the testing error to be between 4.6 to 5.3%, which helps reinforce the accuracy of those chosen k values.

Part 2

1. 1) The learned decision stump
   Feature index: 22
   Test boundary: 0.37284675258389693
   Classification of left child: -1
   Classification of right child: 1
   2) The computed information gain of the selected test
   0.6435368046750327
   3) The training and testing error rates (in percentage) of the learned decision stump
   Training error: 6.0%
   Testing error: 15.1%

2. Report in a table the training and testing error rates of the learned decision trees for different d values. Plot the error rates as a function of d. What behavior do you observe? Provide an explanation for the observed behavior.

d	Training error	Testing error
1	6.0%	15.1%
2	6.0%	15.1%
3	3.2%	12.3%
4	1.8%	11.6%
5	0.0%	12.7%
6	0.0%	12,7%



It appears that training error decreases consistently as d increases, whereas testing error decreases as d increases until d = 4, at which point it increases. This is due to overfitting of the data. At depths greater than 4, the testing boundaries fit so closely with the training data that the overarching similarities become lost. This causes the testing data to be less closely matched.