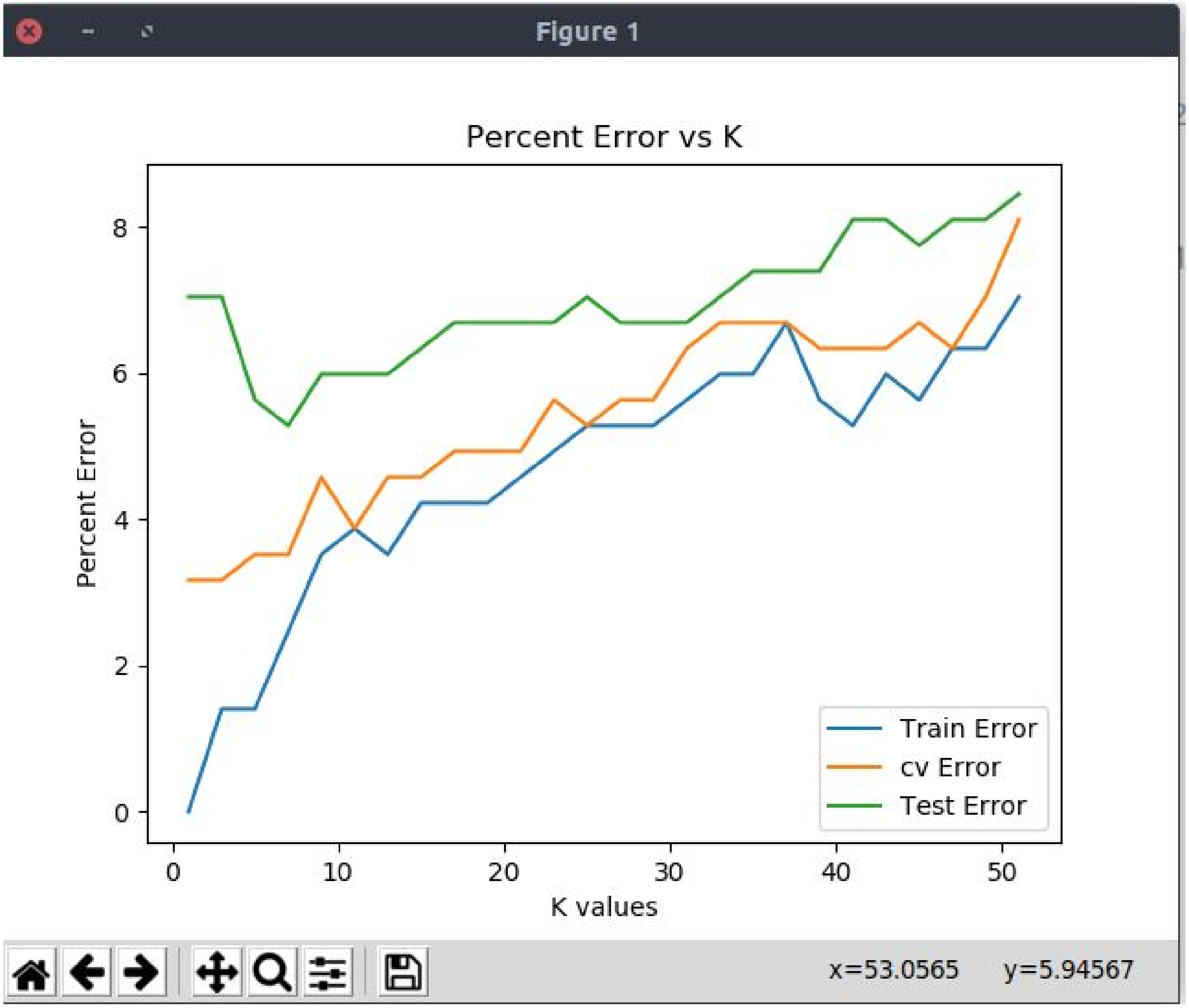
Austin Sanders
Manuel Lara-Navarro

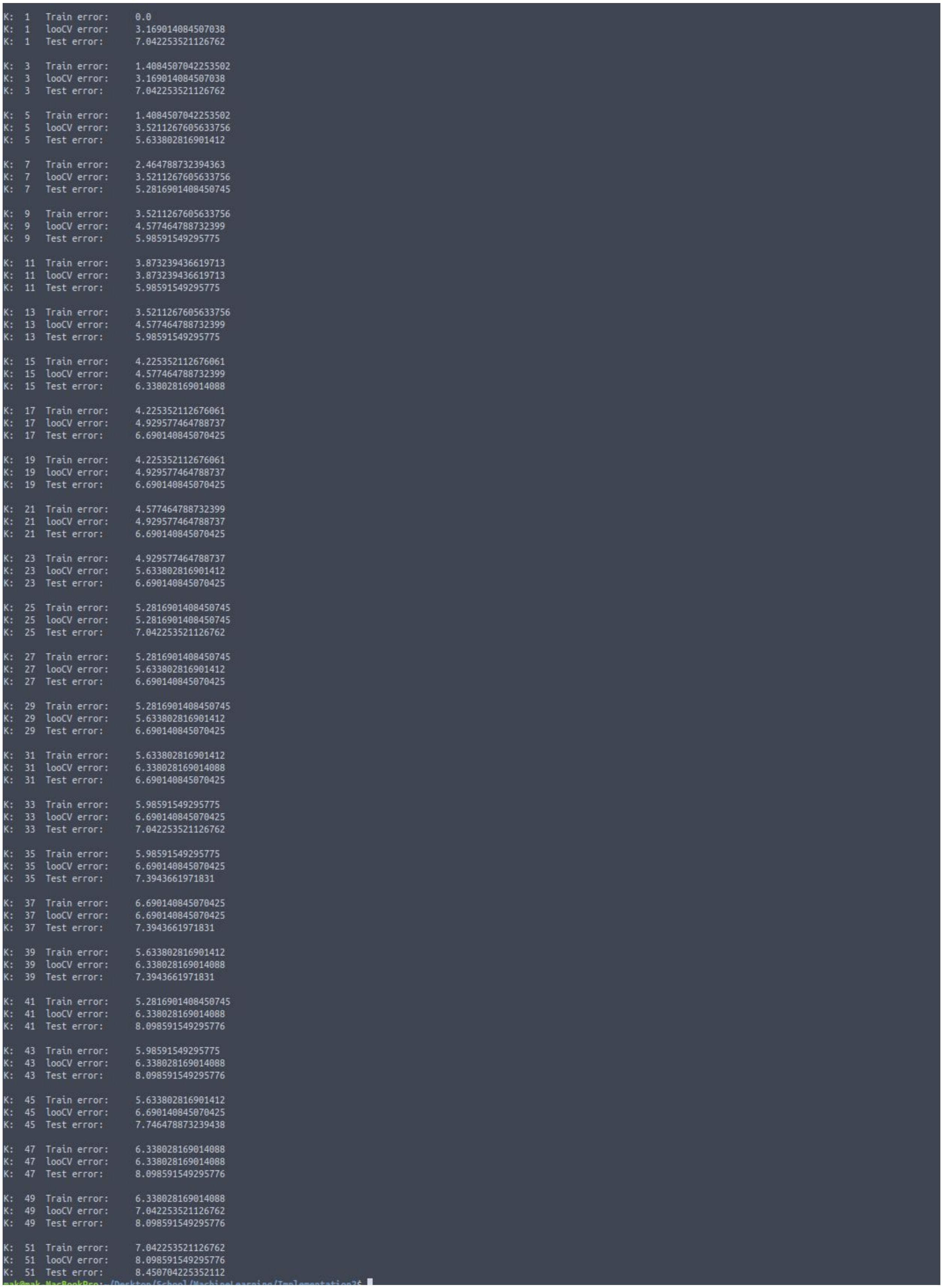
Implementation Assignment Two

1. K Nearest Neighbor

The value we found to be the best fitting with the least amount of error was K = 7. Here is a graph showing the error over time as k was increased:



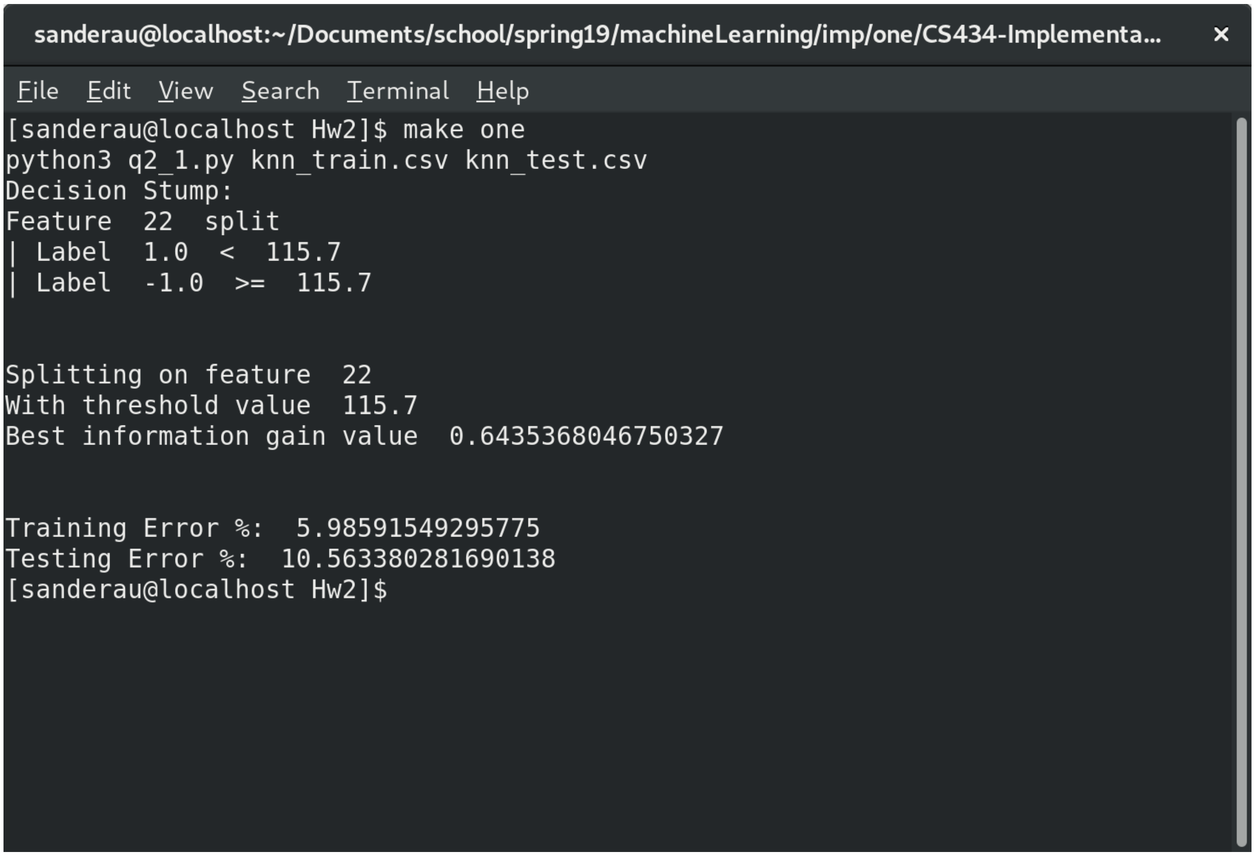
Out of all of the trials we ran the one that had the lowest average error out of all of the categories we were testing was k = 7. What can be observed from the model is that while all of the categories started with relatively different errors as K rose above five the percent error began to rise over time together, and the gap between all three began to close and become much more consolidated. Keeping k = 7 would ensure the highest amount of accuracy with this data set.



```
K:  1   Train error:     0.0
K:  1   looCV error:     3.169014084507038
K:  1   Test error:      7.042253521126762

K:  3   Train error:     1.4084507042253502
K:  3   looCV error:     3.169014084507038
K:  3   Test error:      7.042253521126762

K:  5   Train error:     1.4084507042253502
K:  5   looCV error:     3.5211267605633756
K:  5   Test error:      5.633802816901412

K:  7   Train error:     2.464788732394363
K:  7   looCV error:     3.5211267605633756
K:  7   Test error:      5.2816901408450745

K:  9   Train error:     3.5211267605633756
K:  9   looCV error:     4.577464788732399
K:  9   Test error:      5.98591549295775

K:  11  Train error:     3.873239436619713
K:  11  looCV error:     3.873239436619713
K:  11  Test error:      5.98591549295775

K:  13  Train error:     3.5211267605633756
K:  13  looCV error:     4.577464788732399
K:  13  Test error:      5.98591549295775

K:  15  Train error:     4.225352112676061
K:  15  looCV error:     4.577464788732399
K:  15  Test error:      6.338028169014088

K:  17  Train error:     4.225352112676061
K:  17  looCV error:     4.929577464788737
K:  17  Test error:      6.690140845070425

K:  19  Train error:     4.225352112676061
K:  19  looCV error:     4.929577464788737
K:  19  Test error:      6.690140845070425

K:  21  Train error:     4.577464788732399
K:  21  looCV error:     4.929577464788737
K:  21  Test error:      6.690140845070425

K:  23  Train error:     4.929577464788737
K:  23  looCV error:     5.633802816901412
K:  23  Test error:      6.690140845070425

K:  25  Train error:     5.2816901408450745
K:  25  looCV error:     5.2816901408450745
K:  25  Test error:      7.042253521126762

K:  27  Train error:     5.2816901408450745
K:  27  looCV error:     5.633802816901412
K:  27  Test error:      6.690140845070425

K:  29  Train error:     5.2816901408450745
K:  29  looCV error:     5.633802816901412
K:  29  Test error:      6.690140845070425

K:  31  Train error:     5.633802816901412
K:  31  looCV error:     6.338028169014088
K:  31  Test error:      6.690140845070425

K:  33  Train error:     5.98591549295775
K:  33  looCV error:     6.690140845070425
K:  33  Test error:      7.042253521126762

K:  35  Train error:     5.98591549295775
K:  35  looCV error:     6.690140845070425
K:  35  Test error:      7.3943661971831

K:  37  Train error:     6.690140845070425
K:  37  looCV error:     6.690140845070425
K:  37  Test error:      7.3943661971831

K:  39  Train error:     5.633802816901412
K:  39  looCV error:     6.338028169014088
K:  39  Test error:      7.3943661971831

K:  41  Train error:     5.2816901408450745
K:  41  looCV error:     6.338028169014088
K:  41  Test error:      8.098591549295776

K:  43  Train error:     5.98591549295775
K:  43  looCV error:     6.338028169014088
K:  43  Test error:      8.098591549295776

K:  45  Train error:     5.633802816901412
K:  45  looCV error:     6.690140845070425
K:  45  Test error:      7.746478873239438

K:  47  Train error:     6.338028169014088
K:  47  looCV error:     6.338028169014088
K:  47  Test error:      8.098591549295776

K:  49  Train error:     6.338028169014088
K:  49  looCV error:     7.042253521126762
K:  49  Test error:      8.098591549295776

K:  51  Train error:     7.042253521126762
K:  51  looCV error:     8.098591549295776
K:  51  Test error:      8.450704225352112
mak@mak-MacBookPro:~/Desktop/School/MachineLearning/Implementation2$
```

2. Decision Tree

   2.1.


sanderau@localhost:~/Documents/school/spring19/machineLearning/imp/one/CS434-Implementa...
File Edit View Search Terminal Help

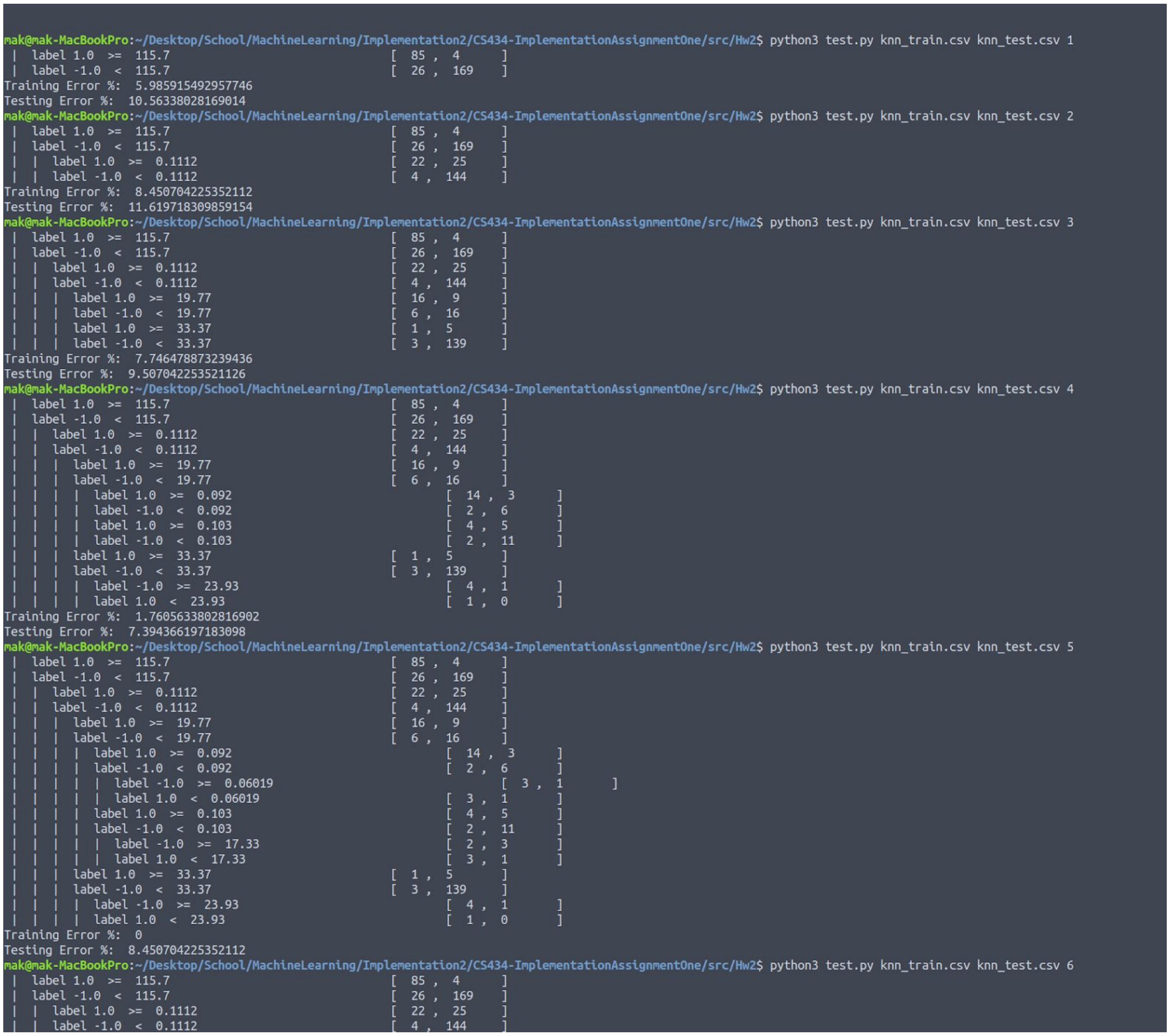
```
[sanderau@localhost Hw2]$ make one
python3 q2_1.py knn_train.csv knn_test.csv
Decision Stump:
Feature  22  split
| Label  1.0  <  115.7
| Label  -1.0  >=  115.7


Splitting on feature  22
With threshold value  115.7
Best information gain value  0.6435368046750327


Training Error %:  5.98591549295775
Testing Error %:  10.563380281690138
[sanderau@localhost Hw2]$
```

We were able to find the decision tree slump by calculating the threshold values for each attribute. We then calculated the information gain for splitting the threshold for that particular threshold.

As you can see in the screen capture above the best information gain we were able to find was 22, and the best split for it was found to be at 115.7. The information gain we got from this was about 0.46.

When we split data at 115.7 we found in the training data we had an accuracy of roughly 94% and in the testing data we found an accuracy of nearly 90%.

2.2.

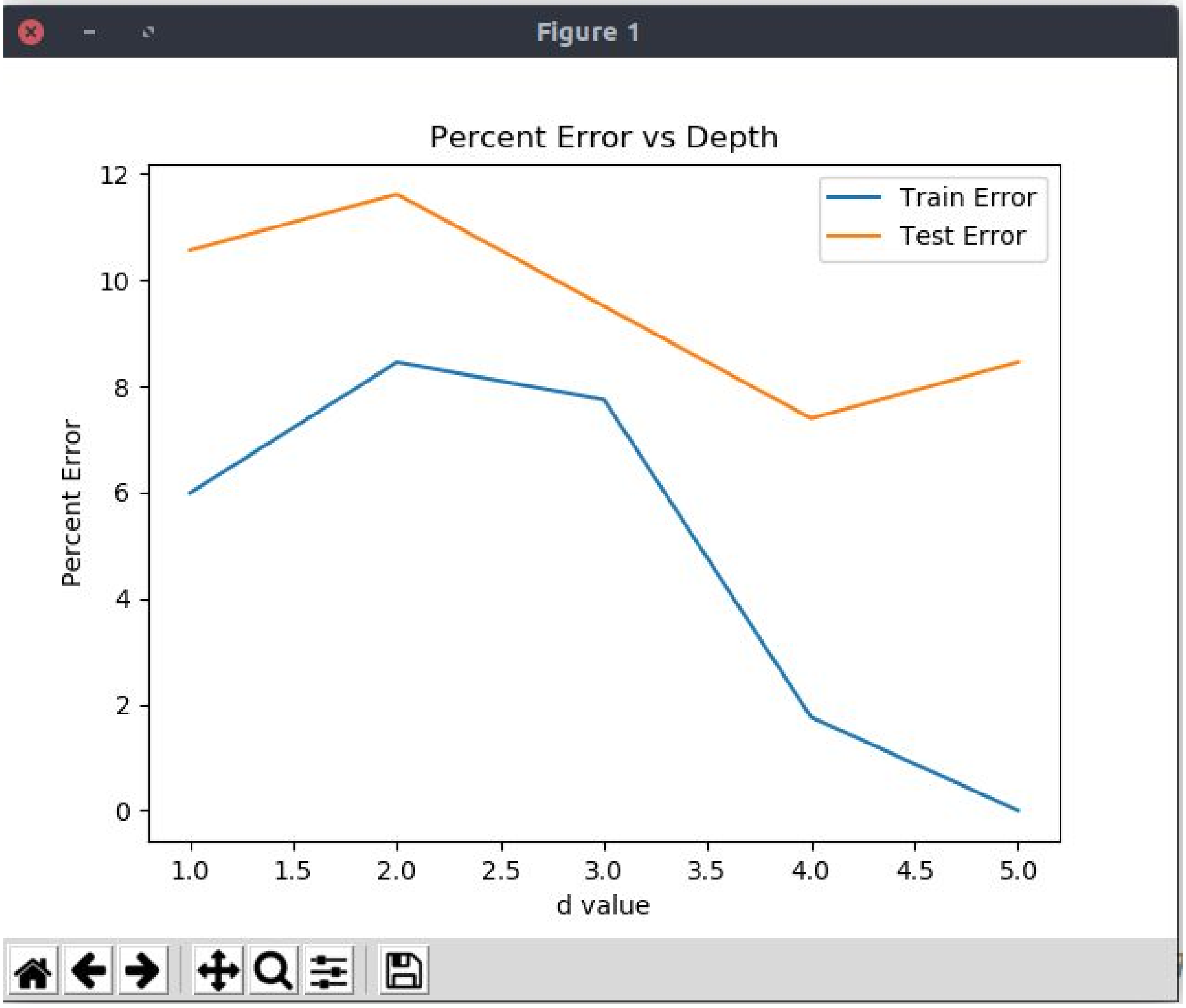


```
mak@mak-MacBookPro:~/Desktop/School/MachineLearning/Implementation2/CS434-ImplementationAssignmentOne/src/Hw2$ python3 test.py knn_train.csv knn_test.csv 1
 |  label 1.0  >=  115.7                                [  85 ,  4      ]
 |  label -1.0  <  115.7                                [  26 ,  169    ]
Training Error %:  5.985915492957746
Testing Error %:  10.56338028169014
mak@mak-MacBookPro:~/Desktop/School/MachineLearning/Implementation2/CS434-ImplementationAssignmentOne/src/Hw2$ python3 test.py knn_train.csv knn_test.csv 2
 |  label 1.0  >=  115.7                                [  85 ,  4      ]
 |  label -1.0  <  115.7                                [  26 ,  169    ]
 |  |  label 1.0  >=  0.1112                            [  22 ,  25     ]
 |  |  label -1.0  <  0.1112                            [  4 ,  144     ]
Training Error %:  8.450704225352112
Testing Error %:  11.619718309859154
mak@mak-MacBookPro:~/Desktop/School/MachineLearning/Implementation2/CS434-ImplementationAssignmentOne/src/Hw2$ python3 test.py knn_train.csv knn_test.csv 3
 |  label 1.0  >=  115.7                                [  85 ,  4      ]
 |  label -1.0  <  115.7                                [  26 ,  169    ]
 |  |  label 1.0  >=  0.1112                            [  22 ,  25     ]
 |  |  label -1.0  <  0.1112                            [  4 ,  144     ]
 |  |  |  label 1.0  >=  19.77                          [  16 ,  9      ]
 |  |  |  label -1.0  <  19.77                          [  6 ,  16      ]
 |  |  |  label 1.0  >=  33.37                          [  1 ,  5       ]
 |  |  |  label -1.0  <  33.37                          [  3 ,  139     ]
Training Error %:  7.746478873239436
Testing Error %:  9.507042253521126
mak@mak-MacBookPro:~/Desktop/School/MachineLearning/Implementation2/CS434-ImplementationAssignmentOne/src/Hw2$ python3 test.py knn_train.csv knn_test.csv 4
 |  label 1.0  >=  115.7                                [  85 ,  4      ]
 |  label -1.0  <  115.7                                [  26 ,  169    ]
 |  |  label 1.0  >=  0.1112                            [  22 ,  25     ]
 |  |  label -1.0  <  0.1112                            [  4 ,  144     ]
 |  |  |  label 1.0  >=  19.77                          [  16 ,  9      ]
 |  |  |  label -1.0  <  19.77                          [  6 ,  16      ]
 |  |  |  |  label 1.0  >=  0.092                               [  14 ,  3      ]
 |  |  |  |  label -1.0  <  0.092                               [  2 ,  6       ]
 |  |  |  |  label 1.0  >=  0.103                               [  4 ,  5       ]
 |  |  |  |  label -1.0  <  0.103                               [  2 ,  11      ]
 |  |  |  label 1.0  >=  33.37                          [  1 ,  5       ]
 |  |  |  label -1.0  <  33.37                          [  3 ,  139     ]
 |  |  |  |  label -1.0  >=  23.93                              [  4 ,  1       ]
 |  |  |  |  label 1.0  <  23.93                                [  1 ,  0       ]
Training Error %:  1.7605633802816902
Testing Error %:  7.394366197183098
mak@mak-MacBookPro:~/Desktop/School/MachineLearning/Implementation2/CS434-ImplementationAssignmentOne/src/Hw2$ python3 test.py knn_train.csv knn_test.csv 5
 |  label 1.0  >=  115.7                                [  85 ,  4      ]
 |  label -1.0  <  115.7                                [  26 ,  169    ]
 |  |  label 1.0  >=  0.1112                            [  22 ,  25     ]
 |  |  label -1.0  <  0.1112                            [  4 ,  144     ]
 |  |  |  label 1.0  >=  19.77                          [  16 ,  9      ]
 |  |  |  label -1.0  <  19.77                          [  6 ,  16      ]
 |  |  |  |  label 1.0  >=  0.092                               [  14 ,  3      ]
 |  |  |  |  label -1.0  <  0.092                               [  2 ,  6       ]
 |  |  |  |  |  label -1.0  >=  0.06019                                 [  3 ,  1       ]
 |  |  |  |  |  label 1.0  <  0.06019                           [  3 ,  1       ]
 |  |  |  |  label 1.0  >=  0.103                               [  4 ,  5       ]
 |  |  |  |  label -1.0  <  0.103                               [  2 ,  11      ]
 |  |  |  |  |  label -1.0  >=  17.33                           [  2 ,  3       ]
 |  |  |  |  |  label 1.0  <  17.33                             [  3 ,  1       ]
 |  |  |  label 1.0  >=  33.37                          [  1 ,  5       ]
 |  |  |  label -1.0  <  33.37                          [  3 ,  139     ]
 |  |  |  |  label -1.0  >=  23.93                              [  4 ,  1       ]
 |  |  |  |  label 1.0  <  23.93                                [  1 ,  0       ]
Training Error %:  0
Testing Error %:  8.450704225352112
mak@mak-MacBookPro:~/Desktop/School/MachineLearning/Implementation2/CS434-ImplementationAssignmentOne/src/Hw2$ python3 test.py knn_train.csv knn_test.csv 6
 |  label 1.0  >=  115.7                                [  85 ,  4      ]
 |  label -1.0  <  115.7                                [  26 ,  169    ]
 |  |  label 1.0  >=  0.1112                            [  22 ,  25     ]
 |  |  label -1.0  <  0.1112                            [  4 ,  144     ]
```

Table

d	Train Error %	Test Error %
1	5.985915492957746	10.56338028169014
2	8.450704225352112	11.619718309859154
3	7.746478873239436	9.507042253521126
4	1.7605633802816902	7.394366197183098
5	0	8.450704225352112

Plot



We can see that error rises in both Training data and Testing data at first. The errors in both data begin to drop steadily after depth 2 and continue to do so. The Training as expected actually reaches 0 error, and Testing flatens out at around 8%.