Ethan Wong
305319001
Dis 1B

CM146 Problem Set 1

1. a) Any sample X will only be labeled 0 ($Y=0$) if $X_1=0$, $X_2=0$, and $X_3=0$. After the initial first decision, there will be $n-3$ more features and therefore 2^{n-3} options (2 options per feature). $2^n - 2^{n-3}$ is on the same order as 2^n. This means the decision tree should predict 1 for every input as it is the most probable result. This leads to 2^{n-3} mistakes. Mistakes will be made $\frac{2^{n-3}}{2^n}$ times, or $\frac{1}{8}$ times.

b) No, it is impossible to reduce the error rate as it will just stay at 1/8. Say that we split on X_1, X_2, or X_3. This will create a leaf that only has 1's and another leaf where there is a 3/4 proportion of 1's. This means that the tree predicts 1 in the two leaf options. Splitting on a larger X_n such that n is greater than 3 also yields a similar result. There will be a 7/8 proportion of 1's in both of the leaves, which means the tree again predicts 1 in both leaf options. Overall, the additional split still has the same error rate (1/8) as the original single-leaf tree.

c) $H(x) = -\frac{1}{8}\log_2(\frac{1}{8}) - \frac{7}{8}\log_2(\frac{7}{8}) = 0.5436$

The entropy is 0.5436 bits.

d) Yes, the entropy of the output Y can be reduced by a non-zero amount by splitting with X_1, X_2, or X_3.

$$H(Y|X_1) = \frac{1}{2}H(Y|X_1=1) - \frac{1}{2}H(Y|X_1=0)$$
$$= \frac{1}{2}[0] + \frac{1}{2}\left[\underbrace{\frac{1}{4}\log_2(\frac{1}{4})}_{Y=0} + \underbrace{\frac{3}{4}\log_2(\frac{3}{4})}_{Y=1}\right]$$
$$= 0 + 0.406$$
$$= 0.406$$

Note: 0 entropy b/c we are certain if $X_1=1$, then $Y_1=1$

The resulting conditional entropy is 0.406 bits.

2. $B(q) = -q \log q - (1-q)\log(1-q)$

$H(S) = B\left(\frac{p}{p+n}\right)$

a) $\frac{\partial B}{\partial q} = -\log q - \frac{q}{q} - \left[-\frac{(1-q)}{(1-q)} - \log(1-q)\right]$

$= -\log q + \log(1-q)$

$= \log\left(\frac{1-q}{q}\right)$

Find the critical point

$\log\left(\frac{1-q}{q}\right) = 0$ when $q = 0.5$

$H(S \mid q=0.5) = -\frac{1}{2}\log_2\left(\frac{1}{2}\right) - \frac{1}{2}\log_2\left(\frac{1}{2}\right) = 1$ [This is the maximum, from critical pt]

There are also critical points at $q=0$, $q=1$ due to them having undefined results

$H(S \mid q=0) = 0 - \log_2(1) = 0$

$H(S \mid q=1) = -\log_2(1) - 0 = 0$

These are the minimum values.

The max is 1 and the min is 0, so $0 \le H(S) \le 1$.

When $p = n$, we see that:

$H(S) = B\left(\frac{n}{2n}\right) = B\left(\frac{1}{2}\right)$

$B\left(\frac{1}{2}\right) = 1$ (seen above), so we know $H(S) = 1$ when $p = n$.

b) $\frac{p_k}{p_k + n_k}$ is the same for all k

$p = \sum_k p_k \qquad n = \sum_k n_k$

$\frac{p_k}{p_k + n_k} = \frac{p}{p+n} \longrightarrow q = \frac{p}{p+n}$ for all k

$H(S) = B\left(\frac{p}{p+n}\right)$

Weighted average entropy:

$$\sum_{i=1}^{K} \frac{p_i + n_i}{n+k} B\left(\frac{p_i}{p_i + n_i}\right) = \frac{1}{p+n} \sum_{i=1}^{K} (p_i + n_i) B\left(\frac{p}{p+n}\right)$$

$$= \frac{p+n}{p+n} B\left(\frac{p}{p+n}\right) = B\left(\frac{p}{p+n}\right)$$

$$\text{Gain} = B\left(\frac{p}{p+n}\right) - B\left(\frac{p}{p+n}\right) = 0$$

Entropy − weighted entropy

Gain = 0

3. a) Using K=1 will minimize the training set error. Using K=1 essentially overfits the training data as each instance gets assigned to itself (each instance is its own neighbor). This causes zero training set error as the overfitting guarantees perfect results for this particular dataset. This is not a reasonable estimate of test set error, though, because using K=1 means that our results are only accurate with this specific data set. Thus, we will not see a good generalization to other data sets such as the test data set.

b) Using K=5 or K=7 will minimize the leave-one-out cross-validation error for this dataset. When using these values for K, the only points that will be misclassified are the bottom two circles or the top two asterisks. The error is thus 4/14. Cross-validation is generally a better measure of test set performance because it uses a test set that is completely different than the training data set. This minimizes risk of "memorization" of results, and therefore increases the odds of ~~being~~ being able to generalize to other data sets such as the test set.

c) Lowest: K=1 Error: 10/14
Using K=1, the only points that will be correctly classified are the outer two circles on the top left, and the outer two asterisks on the bottom right. All the other points are closest neighbors with points of the other type, so there will be error.

Highest: K = 13 Error: 14/14

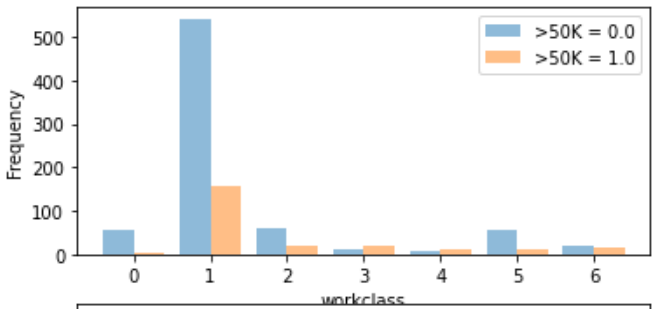
Using K = 13, none of the points will be classified correctly. There are exactly 7 circles and 7 asterisks. When you leave one circle or asterisk out, there will be 6 points that are the same and 7 of the other point. Thus, every single point will be misclassified as there will always be more of the other type.

Using too large of a K value will lead to underfitting, while using too small of a K value will lead to overfitting. Both of these are serious issues that will negatively affect test set accuracy. It is better to use a K-value that is not too small and not too large.

4.1

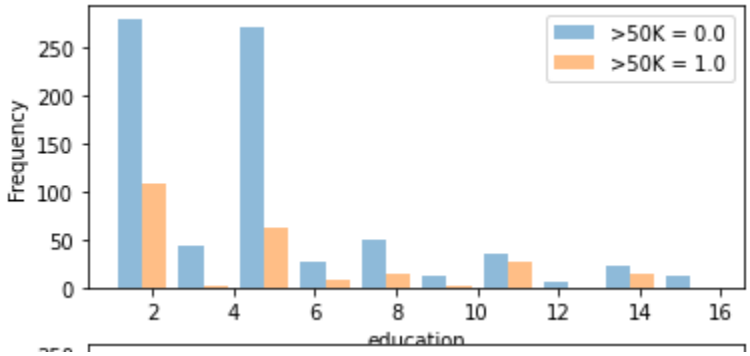
a)

1) Workclass



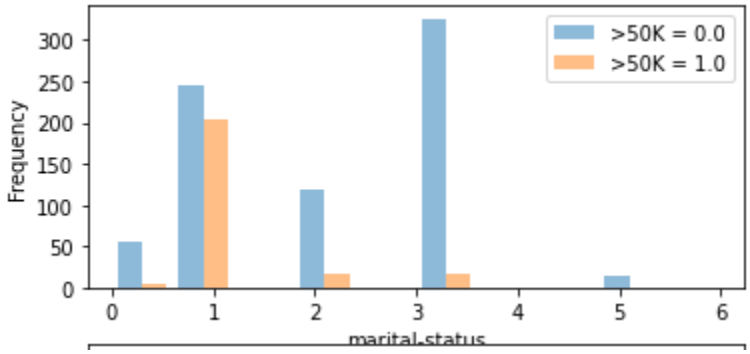
For both classes, the greatest frequency occurred at x=1, where there is more frequency than all the other workclasses combined. The majority of the workclasses have ≤ 50k.

2) Education



For both classes, the highest peaks are observed at x=2 and 4. Much more people have ≤ 50k, though the difference decreases as you go to the right. (2 and 4 are approximations, the data does not seem to fall exactly on those values but they are the closest discernible values)

3) Marital Status

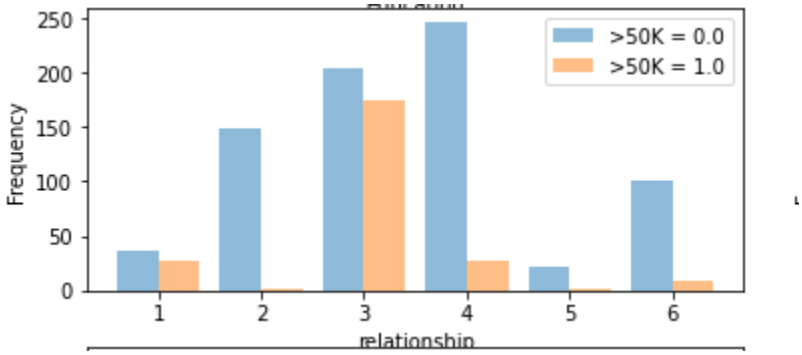


The highest peak for > 50k is at x=1. The highest peak for ≤ 50k is at x=3.

4) Occupation

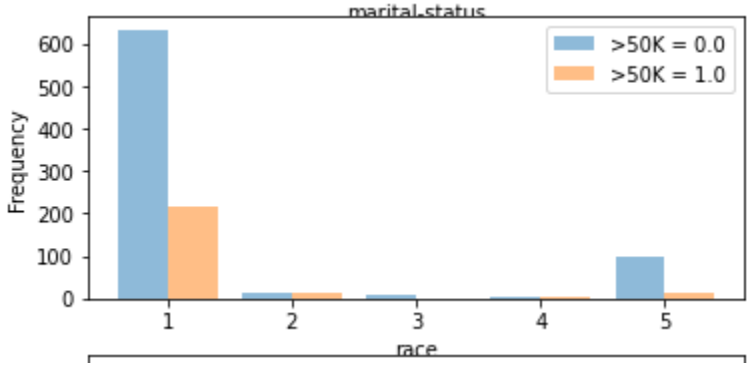
Both classes are close to resembling a normal distribution; much more so than any of the previous plots. The highest peak for > 50k occurs at around x=5 and the highest peak for ≤ 50k is at around x=4. (These x-values are approximations, the data does not seem to fall exactly on those values but they are the closest discernible values)

5) Relationship



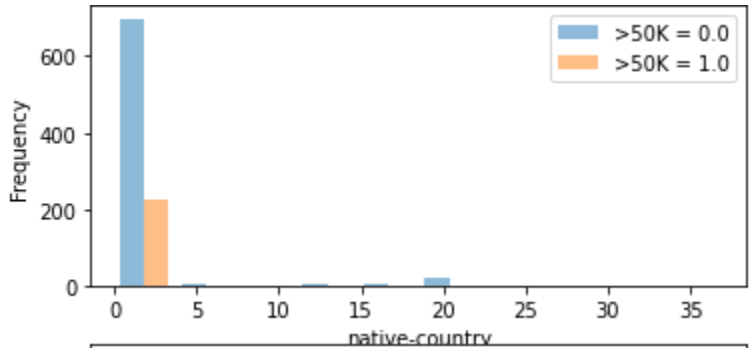
For both classes, there is a normal distribution as the edges are low and the center is a peak. The highest peak for > 50k occurs at x=3 and the highest peak for ≤ 50k is at x=4.

6) Race



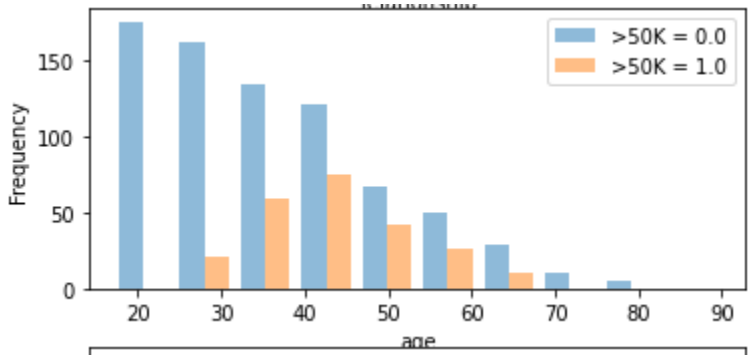
The data for both classes seems very lopsided as the peaks are at the edges but there is very little data in the middle. The highest peak for both > 50k and ≤ 50k occurs at x=1, and they are far taller than any of the other points.

7) Native Country



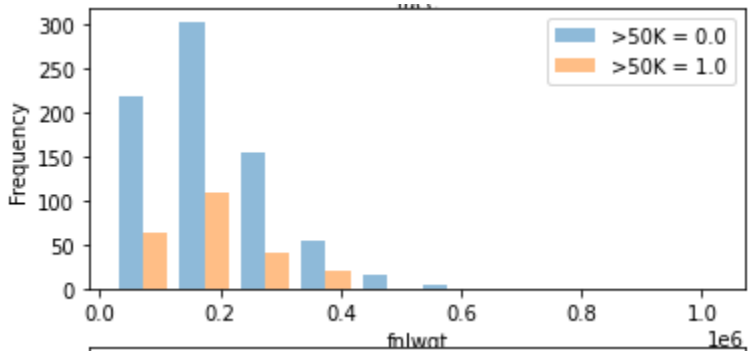
This plot is extremely skewed, with almost all of the data being at x=1 for both > 50k and ≤ 50k. The frequency for ≤ 50k is over three times greater than the frequency for >50k.

8) Age



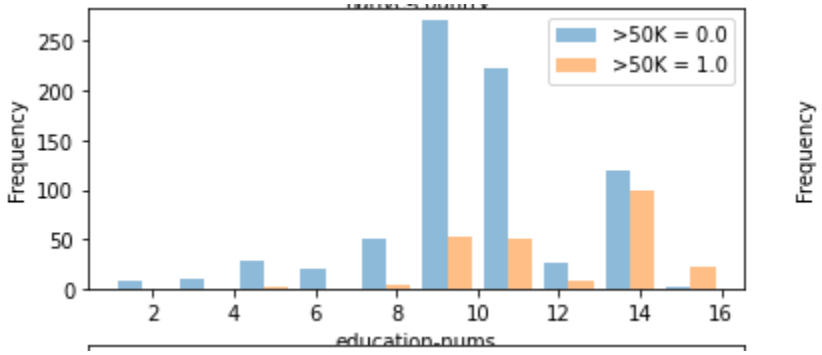
The data is an almost linear downward slope for the age groups. Younger age groups tend to be ≤ 50k, and middle aged people have more that are > 50k, presumably because they have been working for some time now.

9) fnlgwt



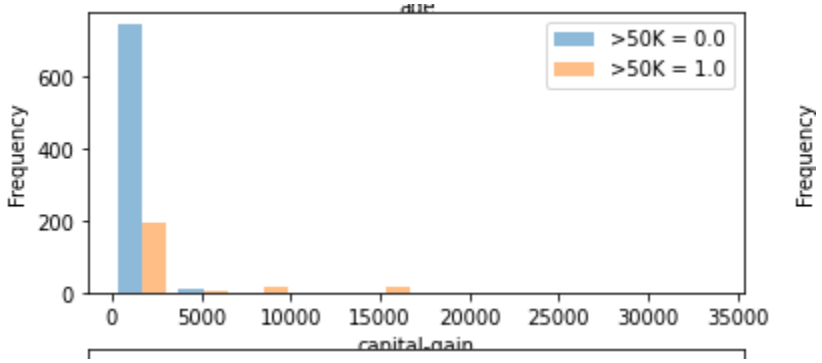
The data is almost all on the left side. The highest peak for > 50k occurs at x=0.2 and the highest peak for ≤ 50k is also at x=0.2.

10) Education nums



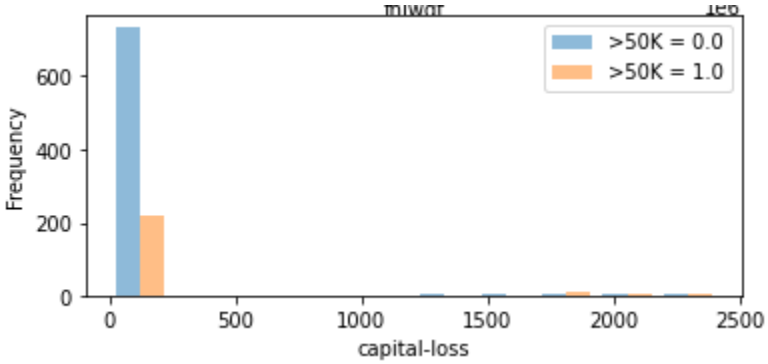
The data is the opposite of fnlgwt, where the data is almost all on the right side instead. The highest peak for > 50k occurs at x=14 and the highest peak for ≤ 50k is at x=9.

11) Capital Gain



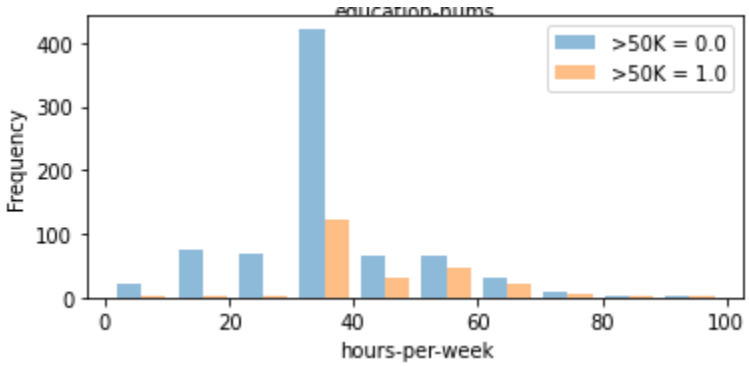
This plot has almost all of the data in one spot, just like the 'Native Country' plot. The highest peak for both > 50k and ≤ 50k occurs at x=1000, and they are far taller than any of the other points.

12) Capital Loss



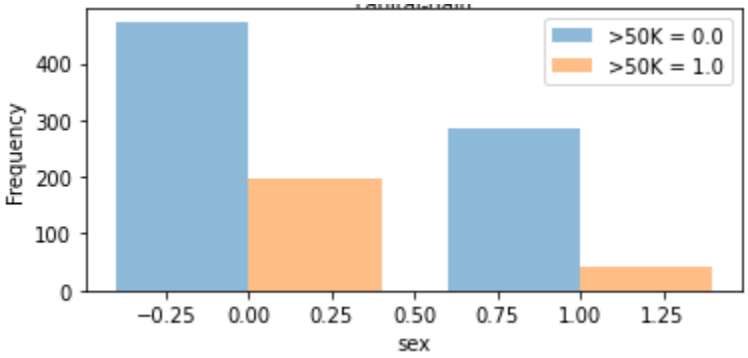
This plot has a nearly identical shape as the 'Capital Gain' one above, but the x-axis is scaled down by a factor of 10. The highest peak for both > 50k and ≤ 50k occurs at x=100.

13) Hours per Week



Both sets of data in this plot resemble a normal distribution, although the center peak is extremely high. The highest peak for both > 50k and ≤ 50k occurs at x=35.

14) Sex



There is a downward slope for both sets of data in this plot. In both instances there are much more in the ≤ 50k category than there are in the >50k category.

4.2

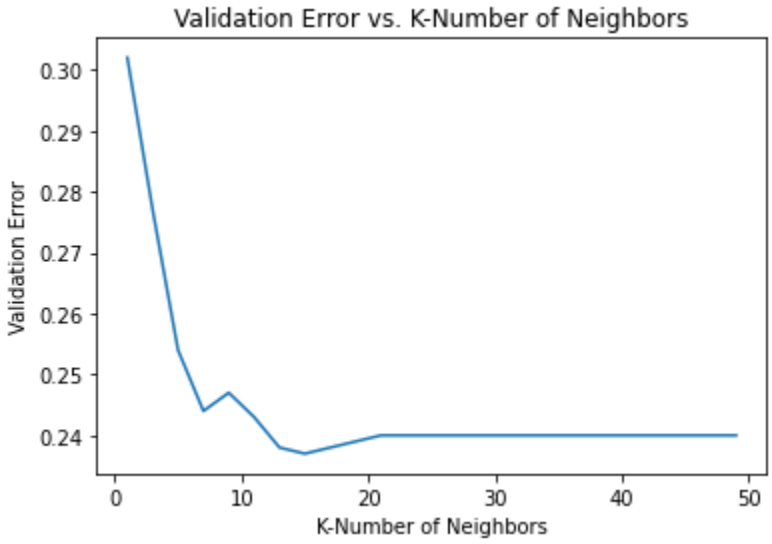
b) The RandomClassifier Training Error was 0.374, while the MajorityVoteClassifer Training error was 0.240.

c) The training error for the DecisionTreeClassifier is 0.000.

d) For the k=3, the training error is 0.153. For the k=5, the training error is 0.195. For the k=7, the training error is 0.213.

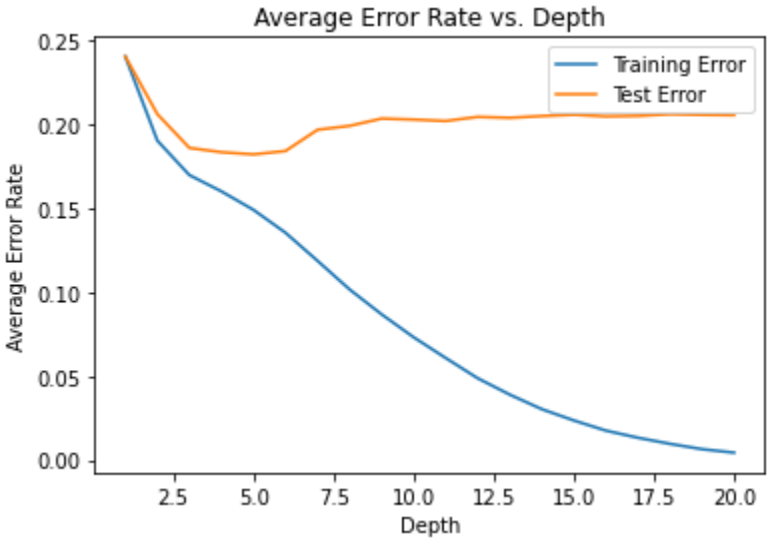
e) Majority Vote Classifier: Training Error: 0.240, Test Error: 0.240, F1 Score: 0.000
Random Classifier: Training Error: 0.375, Test Error: 0.382, F1 Score: 0.251
Decision Tree Classifier: Training Error: 0.000, Test Error: 0.205, F1 Score: 0.569
5-Nearest-Neighbor Classifier: Training Error: 0.202, Test Error: 0.259, F1 Score: 0.160

f)



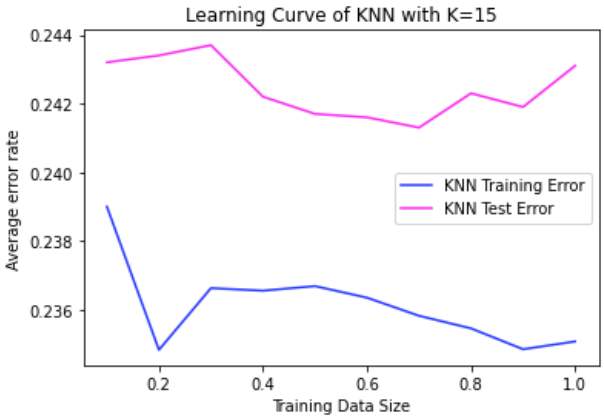
The best k-value for this particular instance was k=15, meaning we should check the 15 Nearest Neighbors for the best results. The worst possible value of k we could have chosen was k=1, where there was an error of 0.30. At k=15, there was an error of 0.235. Once we increase k past 15, the error levels off at approximately 0.24 for the rest of the possible k-values.

g)



According to the testing, the best depth is depth=5. The test error is at a minimum at depth 5, which kind of matches what we learned in class. The decision tree should not be too shallow nor too deep. This is evidenced by the fact that once the depth exceeds 5, the training error steadily decreases until it reaches 0 at a depth of 20. These depths are not good as they are overfitting the data; the training error decreases while the test error increases.

h)



This data was obtained using k=15 for the K-Nearest-Neighbors and depth=5 for the Decision Tree, as these were found to be the optimal values in the previous questions. For the K-Nearest-Neighbors, there does not seem to be any particular trend that the learning curve is following even as more data is being added in to be analyzed. This suggests that the model is not learning from the training data all that well. The learning curve for the Decision Tree, on the other hand, seems to be converging. This is a good sign for the model because it shows that the model is learning the data better over time and is getting more accurate as well, at least for the test data. Test error is decreasing while training error increases.

i)

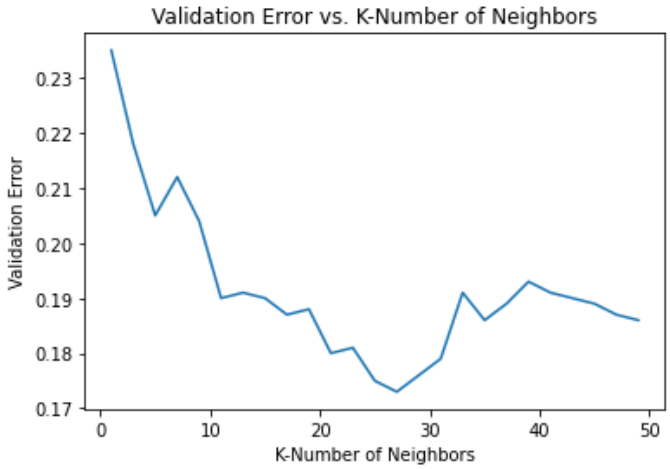
Values before standardization

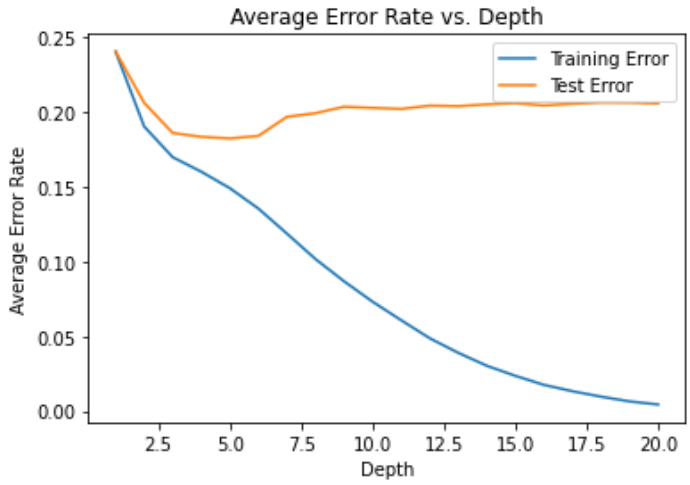
```
Classifying using Majority Vote...
        -- training error: 0.240
Classifying using Random...
        -- training error: 0.374
Classifying using Decision Tree...
        -- training error: 0.000
Classifying using k-Nearest Neighbors...
        -- 3-NN training error: 0.153
        -- 5-NN training error: 0.195
        -- 7-NN training error: 0.213
Investigating various classifiers...

        -- Majority Vote Classifier
         training error: 0.240
         test error: 0.240
         F1 score: 0.000


        -- Random Classifier
         training error: 0.375
         test error: 0.382
         F1 score: 0.251


        -- Decision Tree Classifier
         training error: 0.000
         test error: 0.205
         F1 score: 0.569


        -- 5-NN Classifier
         training error: 0.202
         test error: 0.259
         F1 score: 0.160
```

Values after standardization

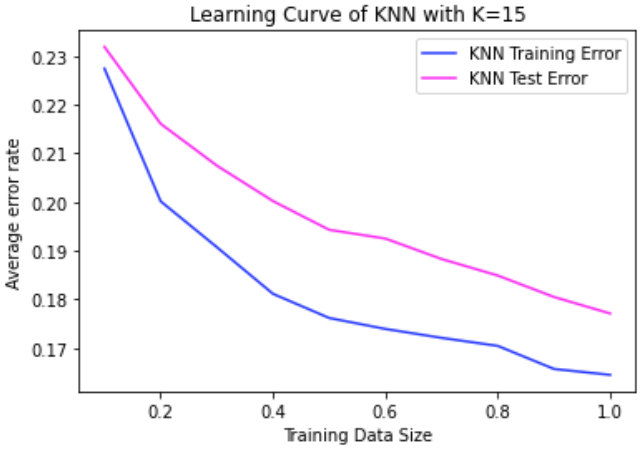
```
Classifying using Majority Vote...
        -- training error: 0.240
Classifying using Random...
        -- training error: 0.374
Classifying using Decision Tree...
        -- training error: 0.000
Classifying using k-Nearest Neighbors...
        -- 3-NN training error: 0.114
        -- 5-NN training error: 0.129
        -- 7-NN training error: 0.152
Investigating various classifiers...

        -- Majority Vote Classifier
         training error: 0.240
         test error: 0.240
         F1 score: 0.000


        -- Random Classifier
         training error: 0.375
         test error: 0.382
         F1 score: 0.251


        -- Decision Tree Classifier
         training error: 0.000
         test error: 0.205
         F1 score: 0.569


        -- 5-NN Classifier
         training error: 0.133
         test error: 0.209
         F1 score: 0.520
```

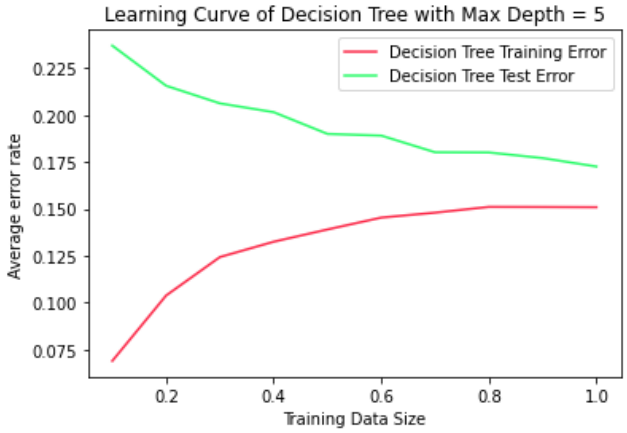
It seems that standardization only had an effect on the error values for K-Nearest-Neighbors. The training error and test error decreased, whereas the F1 score increased.





As we saw previously, the standardization affected the error results for K-Nearest-Neighbors. The standardization also affected the optimal k-value. The new optimal value for k is now k=27, instead of k=15 as it was before.
The decision tree model looks the same even after the standardization. The optimal depth for the decision tree is still 5.





The plot for the K-Nearest-Neighbors changed again due to the standardization. It now looks much better than before, as now the training error and test error are steadily decreasing over time. The plot for the decision tree looks the same as before, which is to be expected because the standardization doesn't seem to affect the decision tree results.