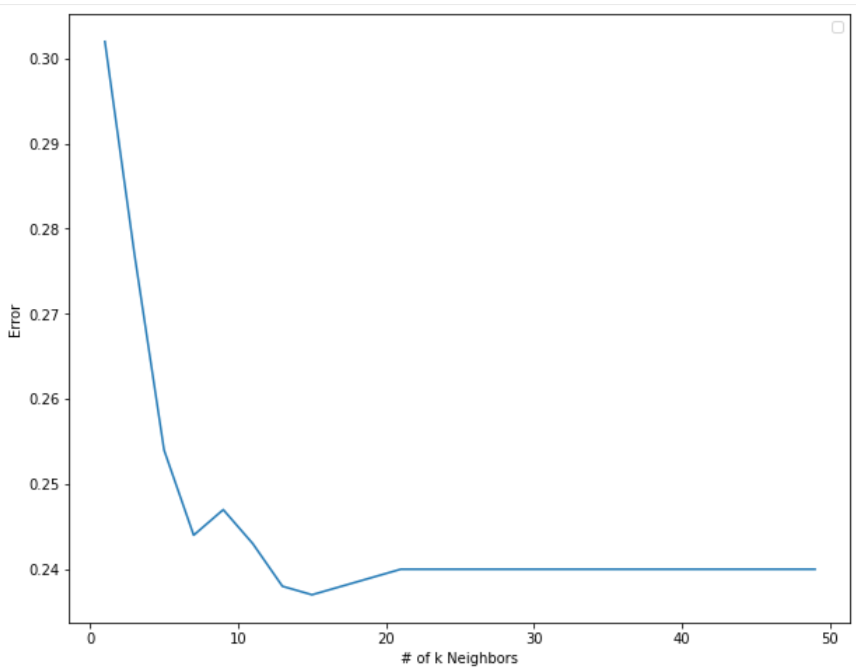
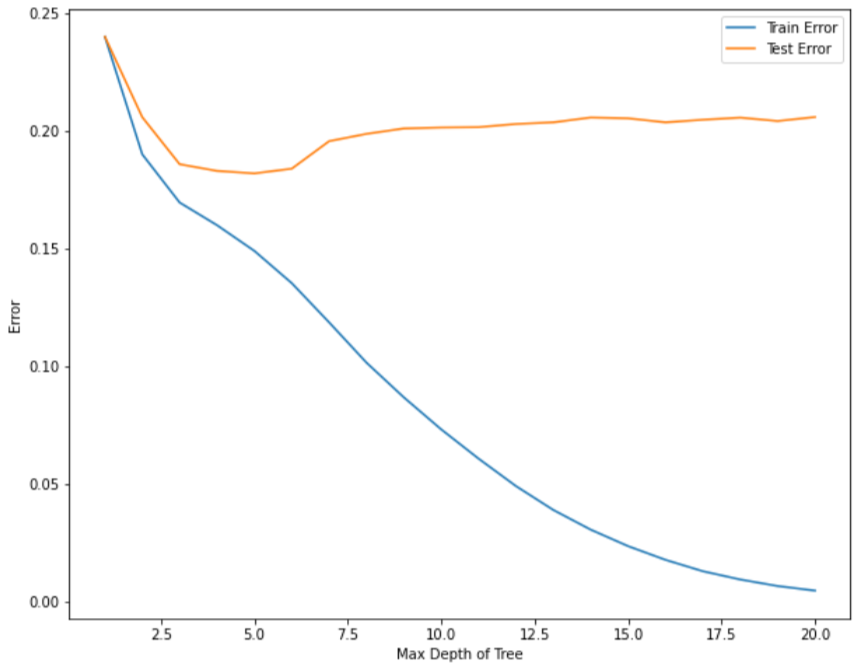
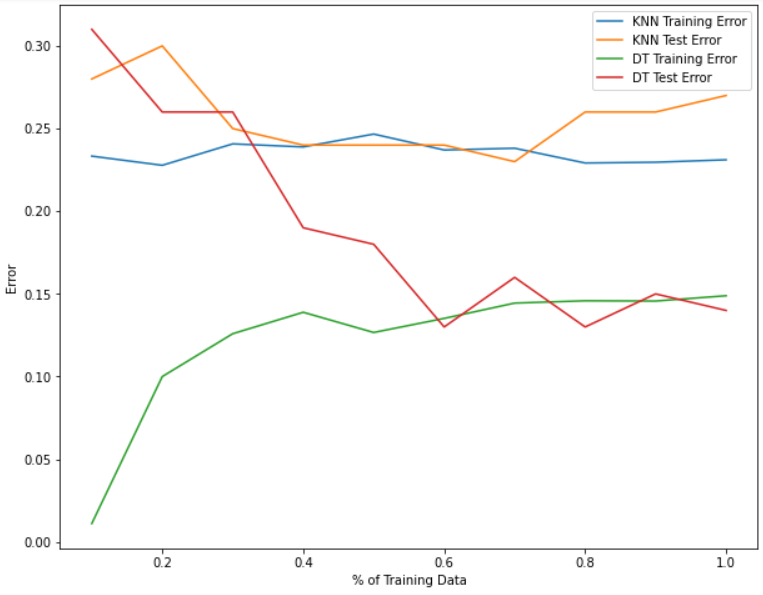
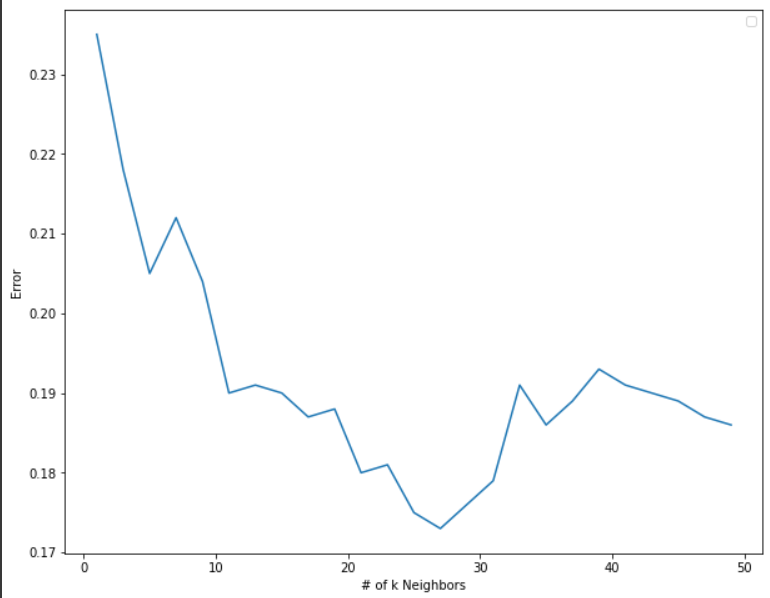
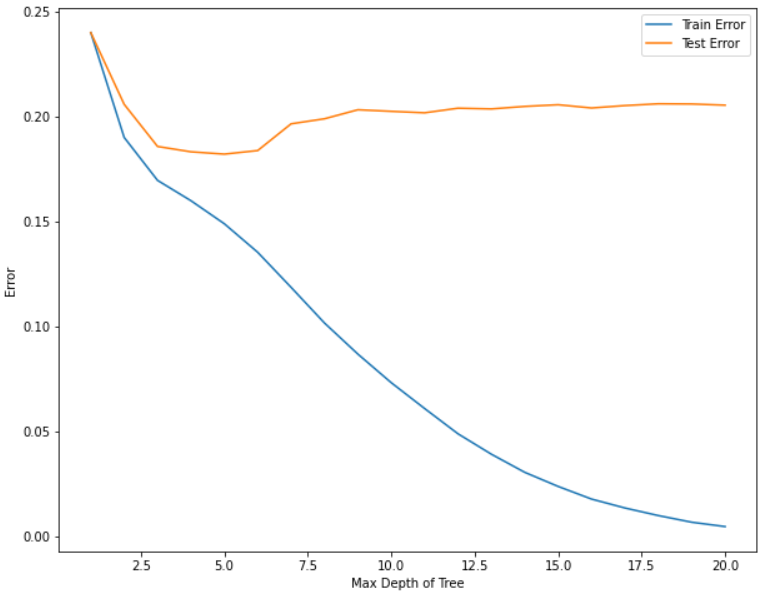
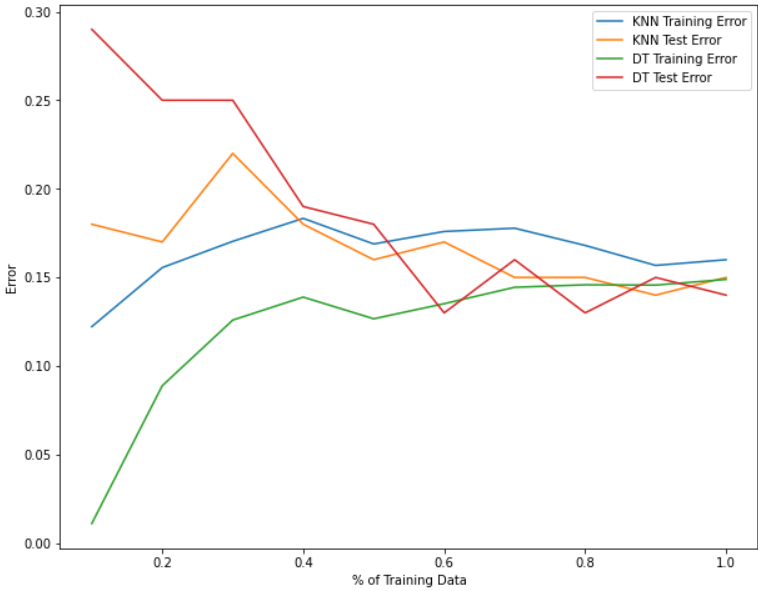
**CSM146 Problem Set 1**

1. 1. Workclass: There isn’t much data in many of the work classes except for number 2. I don’t necessarily see a general trend in the data regarding the salary, however there are much more people in workclass #2 as opposed to any other.
   2. Education: Pretty evenly spread out in terms of the salary. I notice again that there is more data in certain categories, such as 2 and 4 as opposed to the rest.
   3. Marital Status: I notice that marital status is a good feature, as most >50k people are in category #1, while most of <=50k is in category #3. Unfortunately, there is again a lot of missing data in some of the categories.
   4. Occupation: I notice that categories #5 and #6 are pretty evenly split among >50k and <=50k. Categories #1, 2, 3, 8, 9, 10 are very skewed towards <=50k
   5. Relationship: Categories #2, 4, 5, 6 are very skewed towards <=50k, while categories #1, 3 are pretty evenly split I’d say.
   6. Race: The data is very uneven, and there is much more data for category #1. That being said, I’d say it’s pretty evenly split among all categories, but I can’t entirely be sure due to the lack of data.
   7. Native Country: Again, the data is uneven. I can’t make any accurate predictions because of this, however I believe that all >50k are from category #1.
   8. Age: Two clear trends I see. <=50k goes down as age increases. This makes sense as you can’t make >50k when you’re a kid. >50k increases from 30-45, but then decreases from 45-65.
   9. Fnlgwt: Pretty evenly spread out with regard to <=50k and >50k. Not much data 0.6 and above however.
   10. Education Nums: In category #14, <=50k and >50k is evenly split, while category #15 is predominantly >50k. All the rest are predominantly <=50k
   11. Capital Gain: It seems that as capital gain increases, probability of >50k seems to increase as well. However, there is very limited data in the categories besides 0<=x<=5000.
   12. Capital Loss: There does not seem to be much data in the categories besides 0<=x<=500. Because of this, it is not possible to see a general trend.
   13. Hours per Week: It seems there are almost no people who make >50k and work less than 30 hours per week. There may be a slight trend in which people who work more are more likely to make >50k.
   14. Sex: It seems sex 0 is more likely to make >50k while sex 1 is more likely to make <=50k.
3. The training error is 0%. I did not set a max depth for this tree, so it may be overfitting on the training data.
4. 1. With K = 3: Training Error = 0.153
   2. With K = 5: Training Error = 0.195
   3. With K = 7: Training Error = 0.213
5. 1. Majority Vote Classifier:
      1. Training Error: 0.240
      2. Test Error: 0.240
      3. F1 Score: 0.760
   2. Random Classifier:
      1. Training Error: 0.457
      2. Test Error: 0.463
      3. F1 Score: 0.537
   3. Decision Tree Classifier:
      1. Training Error: 0.000
      2. Test Error: 0.205
      3. F1 Score: 0.795
   4. KNeighbors Classifier:
      1. Training Error: 0.202
      2. Test Error: 0.259
      3. F1 Score: 0.741
6.   
   It appears that the best value of k is 15. When the value of k is small, it doesn’t perform well, however as the value of k increases, error seems to drop. However, as the k increases past 15, the error starts going back up again. This is because the model is now behaving like a Majority Classifier at this point, because it’s taking in a lot of neighbors to make its classification.
7.   
   The best depth limit is 5, as the test error is the lowest at this point. There is a lot of overfitting as the depth limit increases, as the training error decreases down to 0 while the test error increases/stays constant. This shows that the deeper decision trees are overfitting on the training data.
8.   
   For the Decision Tree, as the % of training data increases, the Test error increases, resulting in better performance. However, the Training error increases. I believe this is because although we are training our model on more data, we are also increasing the size of the training data we predict on, therefore more chances of error.

For the KNN, as the % of data increases, the Training Error and Test Error more or less stay constant. I believe this is because since KNN simply classifies based on the k nearest neighbors, it doesn’t matter as much the amount of training data it takes given that the training data it *does* receive is a good sample of the overall training data. It basically does Majority classification (or something similar) if k is smaller than or equal to the actual amount of training points received.

1. 1. (b)
   2. (c) The Training error for the Decision Tree is still 0%, which was the same as before. Clearly, the Decision Tree is still overfitting to the Training Data.
   3. (d) All errors of k=3, 5, and 7 are lower than before. Standardizing the Data helped and lowered the error!
      1. With K = 3: Training Error = 0.114
      2. With K = 5: Training Error = 0.129
      3. With K = 7: Training Error = 0.152
   4. (e) The Errors of Majority Vote, Random Classifier, and Decision Tree Classifier stayed the same, so standardizing the data had no effect. However, for KNN classifier, the Training Error and Test Error went down, while the F1 score went up. The performance is better (going by Test Error), so standardizing the data helped.
      1. Majority Vote Classifier:
         1. Training Error: 0.240
         2. Test Error: 0.240
         3. F1 Score: 0.760
      2. Random Classifier:
         1. Training Error: 0.457
         2. Test Error: 0.463
         3. F1 Score: 0.537
      3. Decision Tree Classifier:
         1. Training Error: 0.000
         2. Test Error: 0.205
         3. F1 Score: 0.795
      4. KNeighbors Classifier:
         1. Training Error: 0.133
         2. Test Error: 0.209
         3. F1 Score: 0.791
   5. (f) The best value of k seems to be k=27. This is a very different k value before (k = 15). With standardizing, it looks like for some reason the best k value increases.  
      
   6. (g) The best max depth seems to still be 5. This is the same max depth as before, so standardizing does not seem to have much of an effect on Decision Trees.  
      
   7. (h)  
      The biggest difference is that as % of training data increases, the Error is very similar for all KNN Training Error, KNN Test Error, Decision Tree Training Error, and Decision Tree Test Error. Clearly standardizing helped the KNN classifier reach the performance of Decision Tree. The Decision Tree Performance stayed the same however, and standardizing the data did not negatively or positively affect the Error.