Jacques Gueye

005327762

CM146, Winter 2023 Problem Set 1

1.

a. The best 1 leaf would be just assuming Y=1. In this case, the only mistakes are the ones where X_1 v X_2 v X_3 = 0. The number of mistakes would depend on n>=4, where each number above 3 increases the mistakes exponentially, or 2^{n-3} mistakes.

b. If I can only consider 1 attribute, then it's not possible to reduce the amount of mistakes as it is necessary to ask if the first 3 attributes are 0 to make a deterministic split. There's still not enough information.

c. $2^{n-3}/2^n$ = 1/8

H=-⅛ log(⅛)-⅞ log(⅞)=0.5436

d. Yes. If any of the first 3 Xs are used to split them, this immediately tells us half the data is Y=1 if X=1. The entropy is then:

H=½ [-¼ log(¼)-¾ log(¾)]=0.4056

2.

a. $0 \ \leq \frac{p}{p+n} \leq 1$ since is probability

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

Consider B'(q)= log(1-q) - log(q)=0

q=½ , implies max at B(q)=½

$B(1/2) \ =- \ 1/2 \ log \ 1/2 \ - \ (1/2) \ log \ (1/2) \ = \ 1$

Therefore $B \ \leq \ 1$.

Also consider $B(\frac{p}{p+n}) \ =- \ (\frac{p}{p+n})log(\frac{p}{p+n}) \ - \ (\frac{n}{p+n})log(\frac{n}{p+n})$

log(x) when 0<x<1 is negative=> each term is positive and only addition

Implies sum of all terms >0.

=> $0 \leq B(\frac{p}{p+n}) = H(S) \leq 1$

If p=n, $H(S)$=$B(\frac{n}{n+n})$=B(½)

$B(1/2) =- 1/2\, log\, 1/2 - (1/2)\, log\, (1/2) = 1$

Thus H(S)=1 when p=n

b. If the ratio is the same for all k, the ratio for the whole S is that ratio ($\frac{p}{p+n}$).

In this case, H(S_k)= $B(\frac{p}{p+n})$ = H(S). If gain = H(S) -H(S_k), then gain is 0.

3.

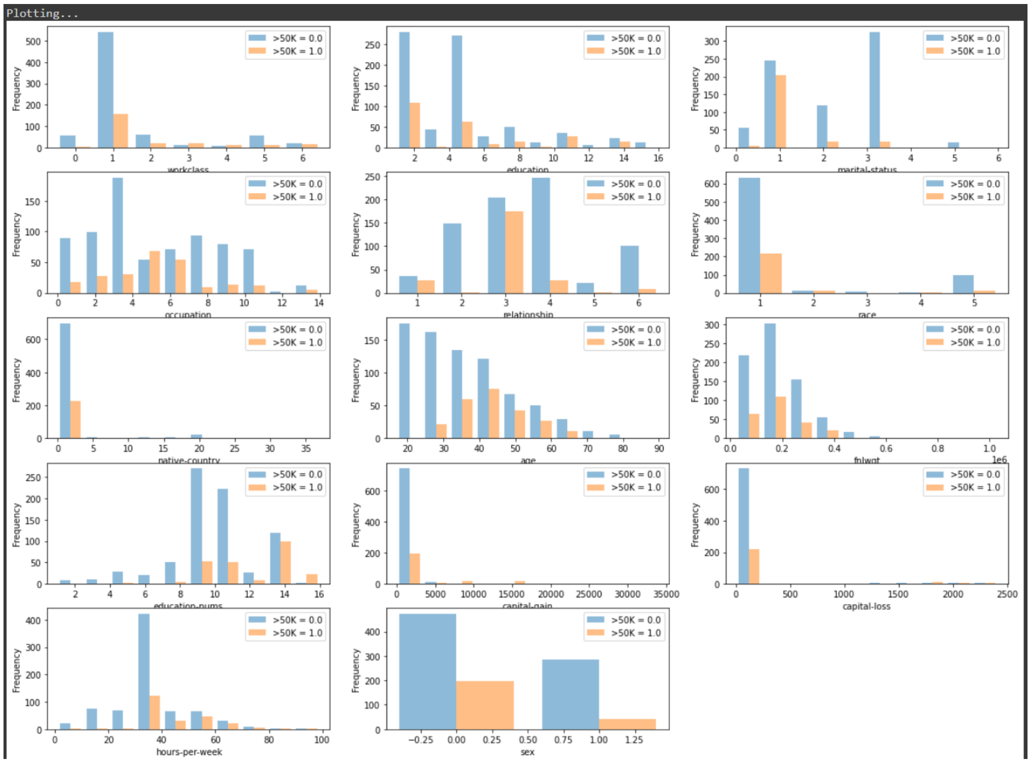
a. A k value of 0 (look at itself) minimizes the error. The training error would be 0 errors for 14 points. The training set error is not a reasonable estimate because we can’t use what we know about the point in test data.

b. A k value of 7 minimizes the error. The error would be 4 errors for 14 points. CV is better so we don’t get overfitting.

c. For k=1, error is 10/14. For k=13, error is 14/14. Too large leads to way too much misclassification. Too little leads to underfitting.

4.

a.



```
Plotting...
```

Workclass: Most people surveyed are part of the type 1 class and making under 50k.
Education: Most people had little education.
Marital Status: If part of type 3 class, very likely to make <50k. Most likely to make over if in class 1.
Occupation: If part of type 4 class, very likely to make <50k
Relationship: If class 4, very likely to make <50k. Most likely to make over if in class 3.
Race: Most surveyed were white, ¼ of them >50k
Native Country: Nearly everyone surveyed native country is US.
Age: The younger the person, the more likely they make <50k
Fnlwgt: The ratios for each class are fairly similar.
Education nums:The higher the num, the more likely they make >50k
Capital gain: Most people have a gain than 5000.
Capital loss: Most people have a loss than 500.
Hours per week: If works 40+ hours a week, more likely to make >50k
Sex: Females are more likely to make >50k than men.

b. Error= 0.374

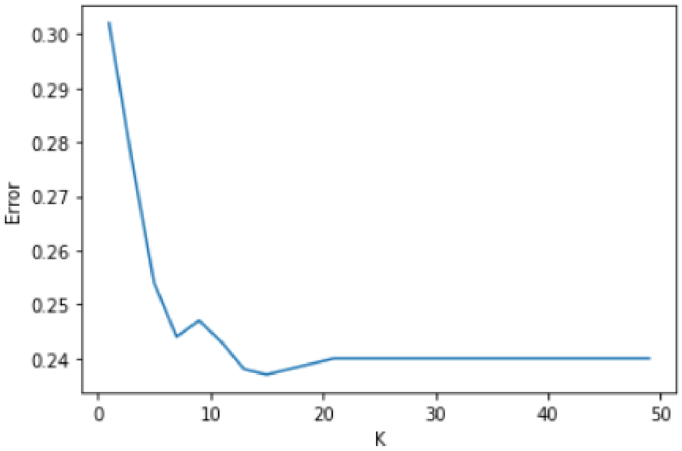
c. 0.0

d. k=3: 0.153
k=5: 0.195
k=7: 0.213

e. -- Majority training error: 0.23999999999999996

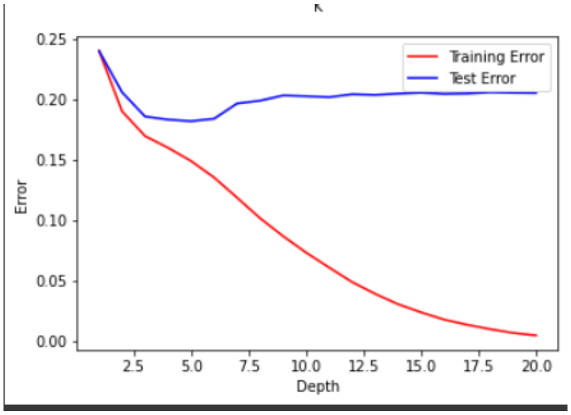
```
-- Majority test error: 0.23999999999999996
-- Majority F1 score: 0.7600000000000002
-- Random training error: 0.37477500000000014
-- Random test error: 0.38199999999999984
-- Random F1 score: 0.6180000000000002
-- DT training error: 0.0
-- DT test error: 0.20475
-- DT F1 score: 0.7952500000000002
-- KNN training error: 0.20167499999999997
-- KNN test error: 0.25915000000000005
-- KNN F1 score: 0.7408499999999998
```

f.



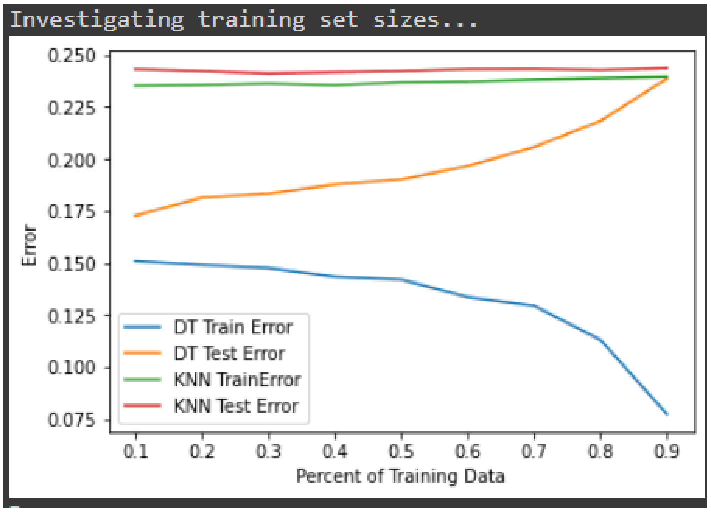
It looks like the error decreases up to k=15, then rises a little to stay constant. The best value of k is 15.

g.



We can see clear signs of underfitting and overfitting in the graph. The best depth limit seems to be around 5 depth as the graphs are close while test error is low. There is clear overfitting as the depth increases, as shown by the training error getting further away from the test error.

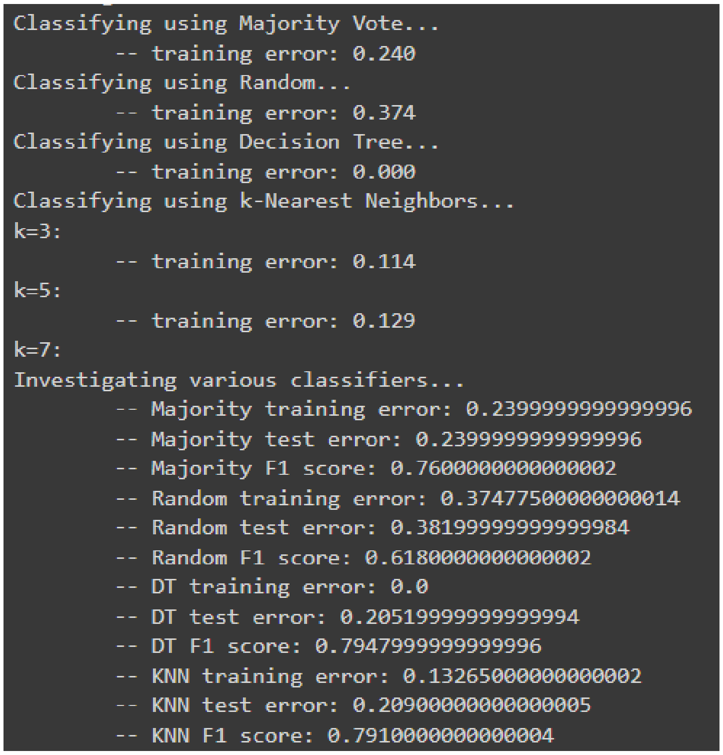
h.



```
Investigating training set sizes...
```

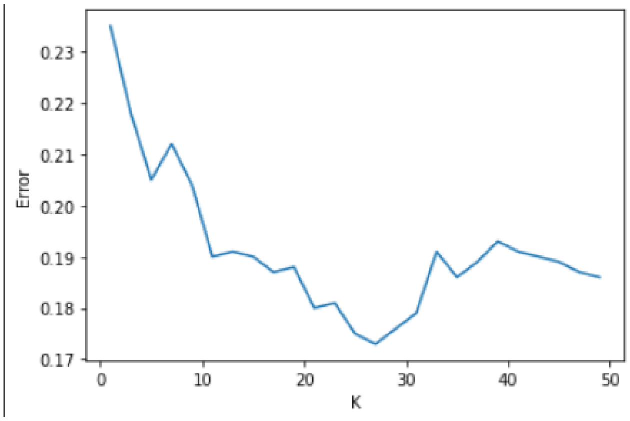
The DT training error (depth 5) decreases as the amount of training data increases, but the test error increases once the training data increases. The KNN training and test data practically stay the same.

i. bcde:

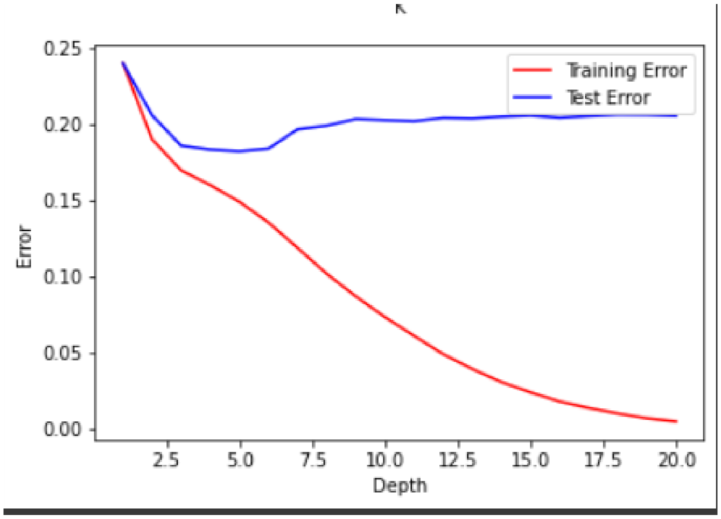


```
Classifying using Majority Vote...
        -- training error: 0.240
Classifying using Random...
        -- training error: 0.374
Classifying using Decision Tree...
        -- training error: 0.000
Classifying using k-Nearest Neighbors...
k=3:
        -- training error: 0.114
k=5:
        -- training error: 0.129
k=7:
Investigating various classifiers...
        -- Majority training error: 0.23999999999999996
        -- Majority test error: 0.23999999999999996
        -- Majority F1 score: 0.7600000000000002
        -- Random training error: 0.37477500000000014
        -- Random test error: 0.38199999999999984
        -- Random F1 score: 0.6180000000000002
        -- DT training error: 0.0
        -- DT test error: 0.20519999999999994
        -- DT F1 score: 0.7947999999999996
        -- KNN training error: 0.13265000000000002
        -- KNN test error: 0.20900000000000005
        -- KNN F1 score: 0.7910000000000004
```

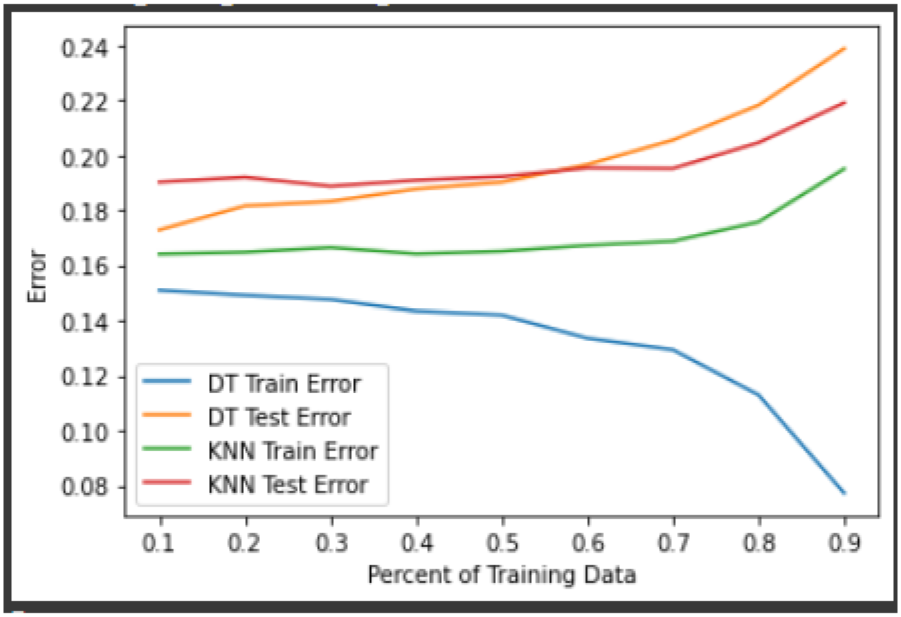
f:


0.23
0.22
0.21
0.20
0.19
0.18
0.17
Error
0
10
20
30
40
50
K

g:


K
0.25
0.20
0.15
0.10
0.05
0.00
Error
Training Error
Test Error
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0
Depth

h:


0.24
0.22
0.20
0.18
0.16
0.14
0.12
0.10
0.08
Error
DT Train Error
DT Test Error
KNN Train Error
KNN Test Error
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
Percent of Training Data

Practically no difference for parts b-e. However, we find that a K value of about 27 seems to be the value with the least error in part f. Moreover, the errors in part h are all lower, even though the KNN k value wasn't optimized. Seems like in general performance was improved.