PROB #1 : completed on CCLE

PROB #2 .

$$\text{Gain}(S, A) = \text{Entropy}(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} \text{Entropy}(S_v).$$

$A = X_j$

$\text{Values}(X_j) = \{1, \ldots, k\}$

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

$$\sum_{v \in values(A)} \frac{|S_v|}{|S|} \text{Entropy}(S_v) = \sum_{i=1}^{k} \frac{|S_i|}{|S|} \text{Entropy}(S_i) = \sum_{i=1}^{k} \frac{p_i + n_i}{p+n} H(S_i)$$

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

→ since $\frac{p_i}{p_i+n_i}$ is same for all values of i

$$= B\left(\frac{p_1}{p_1+n_1}\right) \frac{1}{p+n} \sum_{i=1}^{k} p_i + n_i$$

→ This summation is just sum of of all points in $S = |S|$.

$$= B\left(\frac{p_1}{p_1+n_1}\right) \frac{1}{p+n} (|S|) = B\left(\frac{p_1}{p_1+n_1}\right) \frac{\cancel{p+n}}{\cancel{p+n}}$$

$$= B\left(\frac{p_1}{p_1+n_1}\right)$$

→ since the ratio is the same for all i this means it is the same ratio for the whole data set.

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

$$\therefore \text{Gain}(S, X_j) = \text{Entropy}(S) - \sum_{v \in Values(X_j)} \frac{|S_v|}{|S|} \text{Entropy}(S_v)$$

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

$$= 0.$$

$\therefore$ Information gain $= 0$. ✓

PROB # 3

a.) Since a point can be its own neighbor, $k=0$ minimizes training set error for this dataset. The resulting training error is 0.

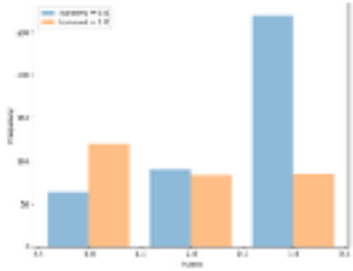
b.) Large values of k would be bad for this data set because it would cause most or all data points to be labeled incorrectly. A k that is too small would lead to overfitting.

c.) $k=5$ minimizes error. The resulting error is $4/14$.

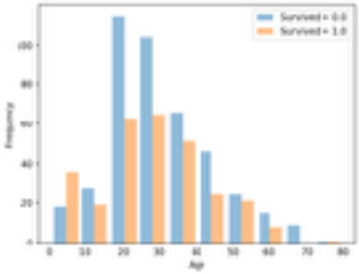
Problem 4:

a.)

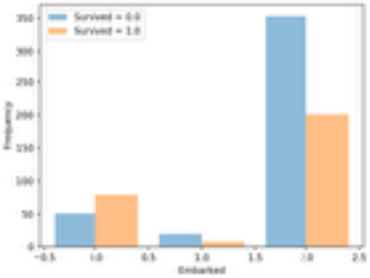
- Ticket Class (Pclass)
  - The ratio of those that survive to those that did not decreases dramatically the lower the class. The number of people who survived from each class is about the same, but the total number of people in each class increases from 1-3.
    - This implies that people in 3rd class were more likely to have died that those in 2nd class and those in second class are more likely to have died than those in first class.



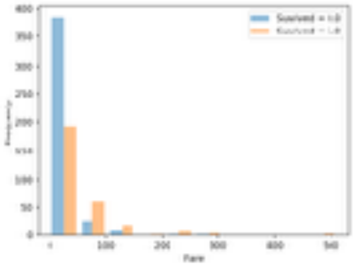
- Age
  - People under the age of 20 were more likely to survive than those between the ages 20-40. The ratio of survivors to non-survivors increases after the 40 year mark



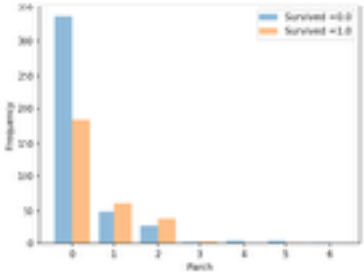
- Port of Embarkation
  - Assuming 0 = Cherbourg, 1 = Queenstown, 2 = Southampton
    - People who got on at Cherbourg were more likely to survive than those who got on at the other 2 ports



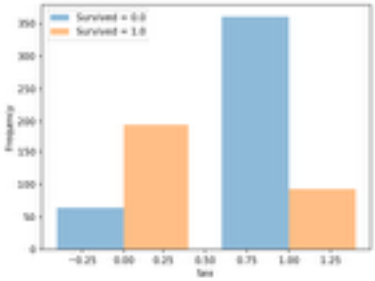
- Fare
  - Those who paid a higher fair were much more likely to survive



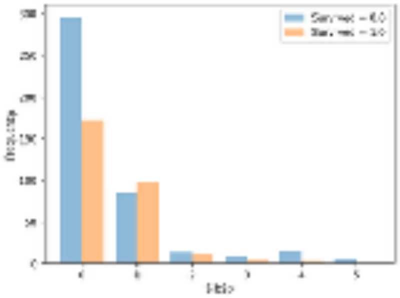
- # parents/children aboard
  - Those who did not have parents or children aboard were more likely to die.



- Sex
  - Females were much more likely to survive than males



- # siblings and spouses aboard
  - Those who had no siblings or spouses aboard were much more likely to die than those who did



b.) I implemented RandomClassifier, when trained on the entire training data set its training error was 0.485.

c.)

-- training error for decision tree: 0.014

d.)

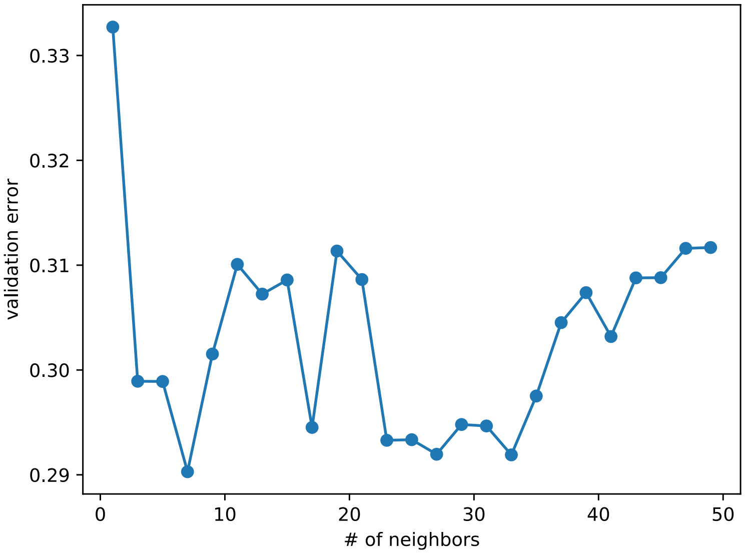
-- training error for 3 nearest neighbors: 0.167
-- training error for 5 nearest neighbors: 0.201
-- training error for 7 nearest neighbors: 0.240

e.)

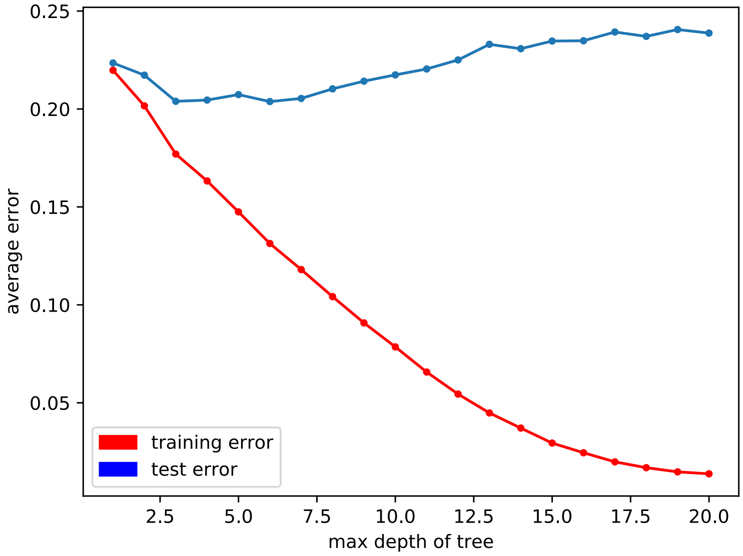
-- Average training error for majority: 0.404
-- Average test error for majority: 0.407
-- Average training error for random: 0.489
-- Average test error for random: 0.487
-- Average training error for decision tree: 0.012
-- Average test error for decision tree: 0.241
-- Average training error for 5 nearest neighbors: 0.212
-- Average test error for 5 nearest neighbors: 0.315

f.) The best value for k is 7.

The values of k can dramatically change the validation error. As we can see from the figure below all values of k under 7 show a high error, most likely due to not taking enough points into consideration after that the model fluctuates between high and low data based on the # of neighbors, but after 30 it continually increases again, probably because the diversity of the 30+ closest neighbors is too high.

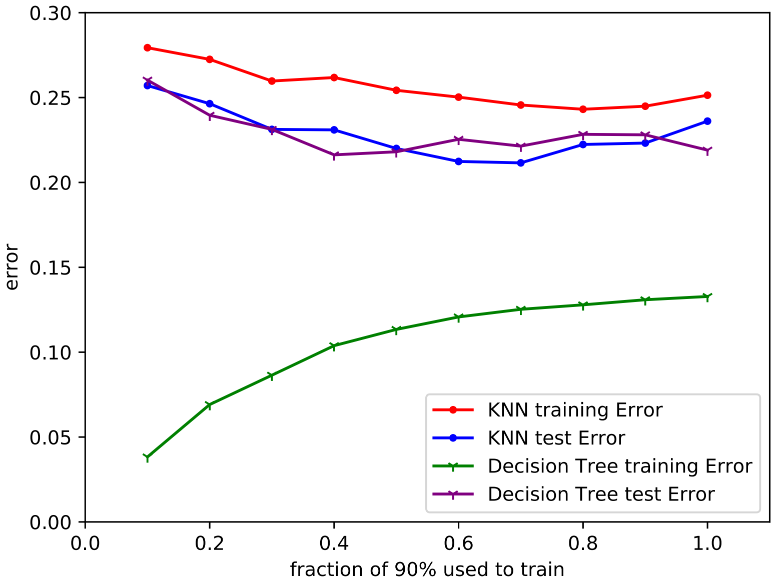


g.) The best depth limit to use is 6.
I do see overfitting, after a depth of 6 the average test error starts increasing even though the training error continues to decrease. Therefore, it is best to stop at 6, because it gives the minimum test error.



h.) For KNN, as the fraction of the training data used to train was increased, the training error decreased; however the test error decreased and then started increasing again at around 70% of the training data.

For decision tree, the more training data was used the higher the training error rate became, probably because there were more points to misclassify. The test error fluctuated slightly, it started off high then decreased and increased slightly again.


error
fraction of 90% used to train
KNN training Error
KNN test Error
Decision Tree training Error
Decision Tree test Error