M146 Homework Set #1

Name: Yangyang Mao UID:504945234

M146 Homework Set #1

Name: Yangyang Mao UID:504945234

1. (a) $H(\text{Is poisonous}) = \frac{3}{8}\log_2\frac{8}{3} + \frac{5}{8}\log_2\frac{8}{5} = 0.954.$

(b) Is Smooth

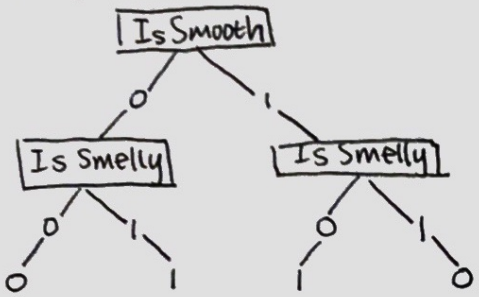
(c) ~~Information gain~~ $= 1 + \frac{1}{4}\log_2 4 + \frac{3}{4}\log_2\frac{4}{3} = 0.905$

$H(\text{Is Poisonous} \mid \text{Is Smooth})$

(d) $H(\text{Is Poisonous}) = 0.954$

Information gain $= H(\text{Is Poisonous}) - H(\text{Is Poisonous} \mid \text{Is Smooth}) = 0.049$

(d)



(e) U : not poisonous

V : not poisonous

W : poisonous

2. $p = \sum_k p_k \quad n = \sum_k n_k$

$\frac{p_k}{p_k+n_k}$ is the same for all k, thus $\frac{p_k}{p_k+n_k} = \frac{p}{p+n} = q$

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

$\therefore$ Information gain $= B\left(\frac{p}{p+n}\right) - B\left(\frac{p}{p+n}\right) = 0$

3. (a) $k=1$. Since when $k=1$, the point is its own neighbour.

The resulting training error $= 0$

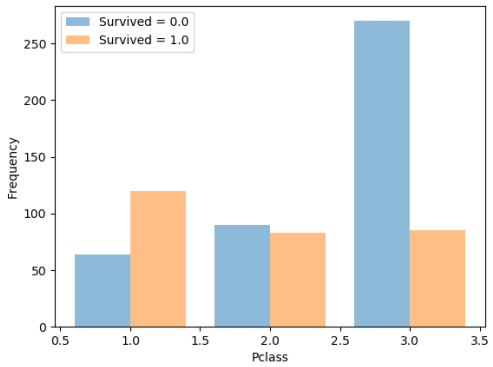
(b) Using too large k will lead to misclassification.

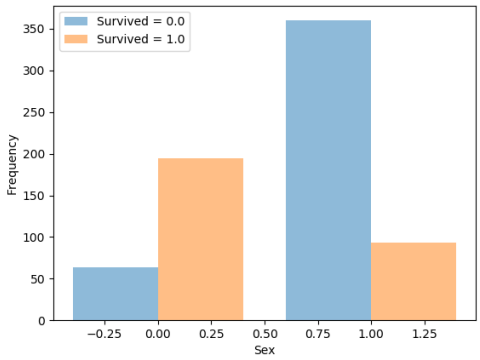
Using too small k will lead to overfitting.

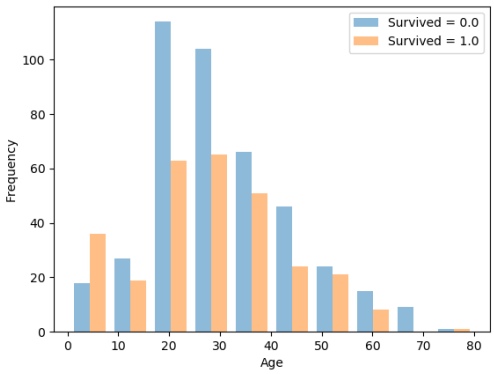
(c) $k=5$ or $k=7$

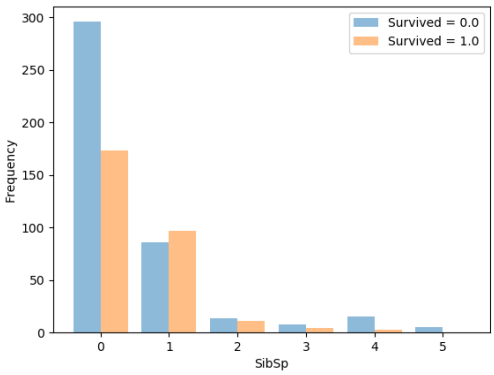
The resulting error is $\frac{4}{14}$

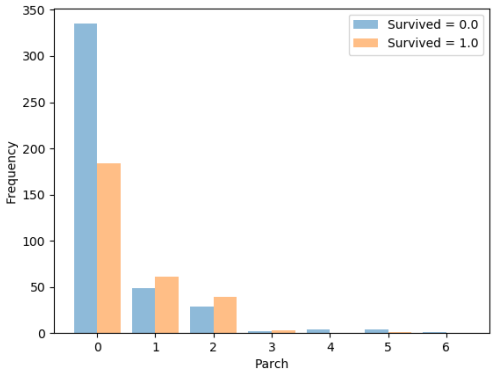
4.1(a)

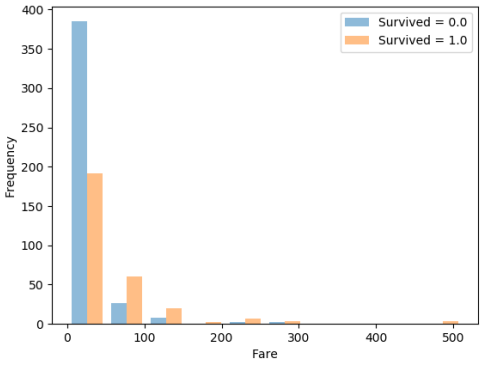

Survived = 0.0
Survived = 1.0
Frequency
Pclass

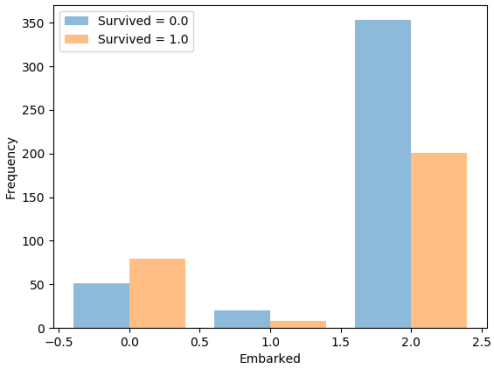

Survived = 0.0
Survived = 1.0
Frequency
Sex


Survived = 0.0
Survived = 1.0
Frequency
Age


Survived = 0.0
Survived = 1.0
Frequency
SibSp

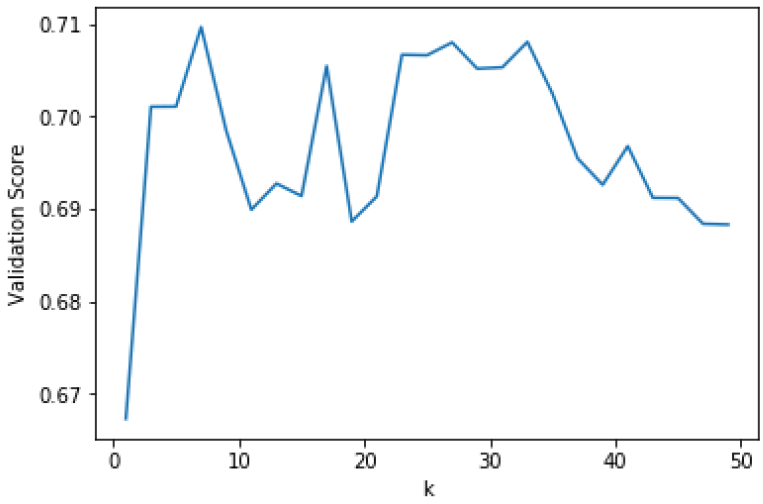

Survived = 0.0
Survived = 1.0
Frequency
Parch


Survived = 0.0
Survived = 1.0
Frequency
Fare


Survived = 0.0
Survived = 1.0
Frequency
Embarked

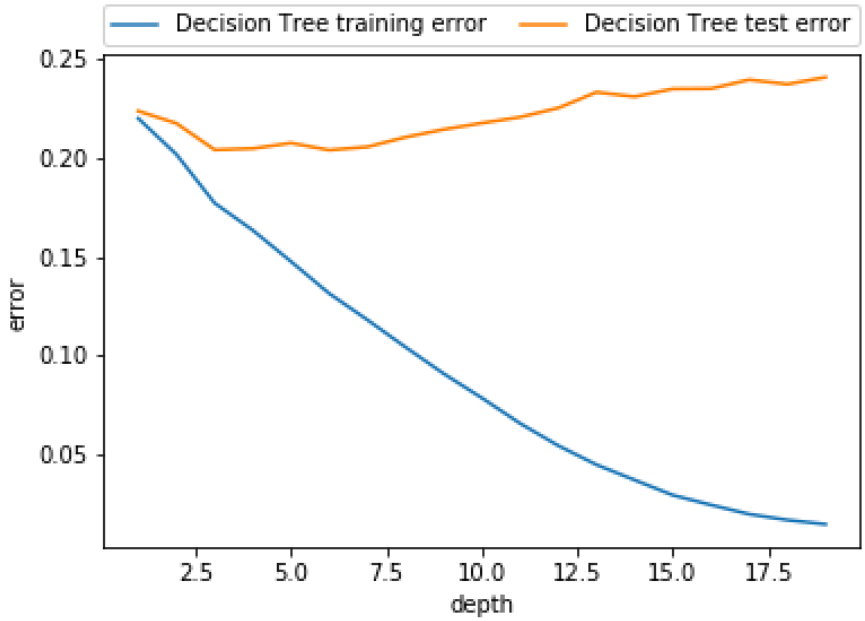
For feature Pclass, first class had the highest survival rate, and third class had the lowest survival rate.
For feature Sex, females had a higher survival rate than males.
For feature Age, people whose age below had the highest survival rate, people who are older than 65 had the lowest survival rate.
For feature Sibsp, passengers with at least one travelling sibling or spouse had higher survival rates than a passengers travelling alone.
For feature Parch, passengers with at least one travelling parent or child had higher survival rates than passengers travelling alone
For feature Fare, passengers who paid more for their fare had a higher survival rate.
For feature Embarked, passengers with embarked smaller than 0.5 has a higher survival rate.

4.2(b) Classifying using Random -- training error: 0.485
(c) Classifying using Decision Tree -- training error: 0.014
(d) Classifying using k-Nearest Neighbors...k=3-- training error: 0.167
Classifying using k-Nearest Neighbors...k=5-- training error: 0.201
Classifying using k-Nearest Neighbors...k=7-- training error: 0.240
(e) Classifying using Majority Vote...
-- training error: 0.404
-- test error: 0.407
Classifying using Random...
-- training error: 0.489
-- test error: 0.487
Classifying using Decision Tree...
-- training error: 0.012
-- test error: 0.241
Classifying using k-Nearest Neighbors...k=5
-- training error: 0.212
-- test error: 0.315
(f)



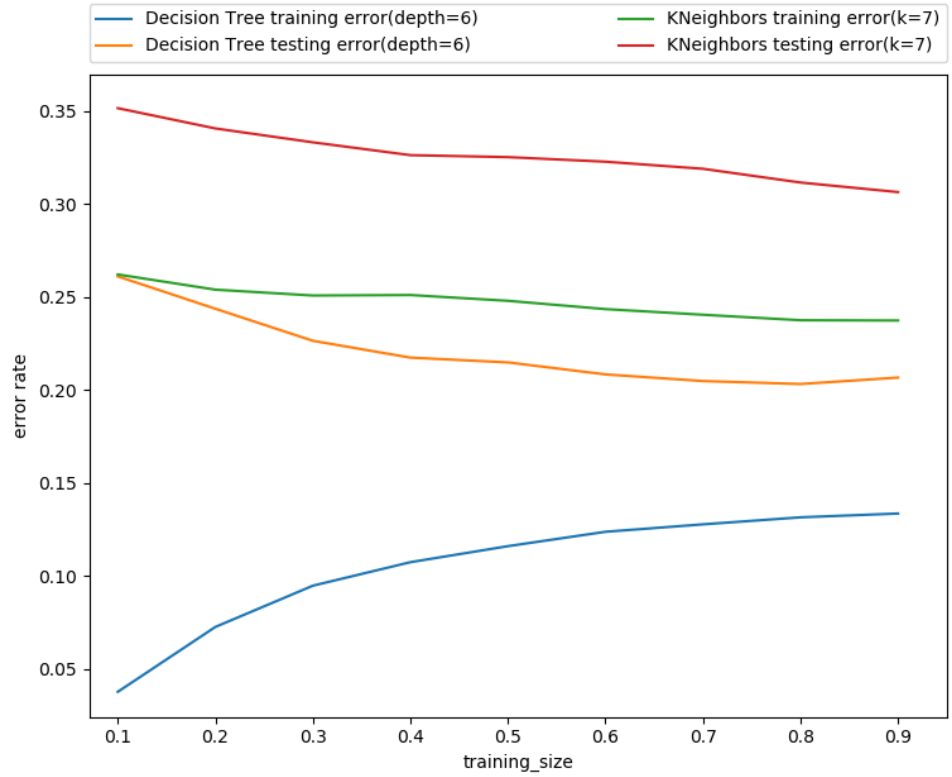
Best value for k is 7. Validation score is fluctuating with increasing of k.

(g)



The figure shows the trend that with increasing of depth, the train error is decreasing and the test error is decreasing then increasing. Best depth limit to use for this data is 6. We can see overfitting since the test error is increasing when depth is greater than 6.

(h)



With increasing of training size, the difference betwwen test error and training error is

becoming smaller. In decision tree, the training error is incrasing and test error is decreasing then increasing with the increasing of k; In KNN, both of the test error and the train error is decreasing.