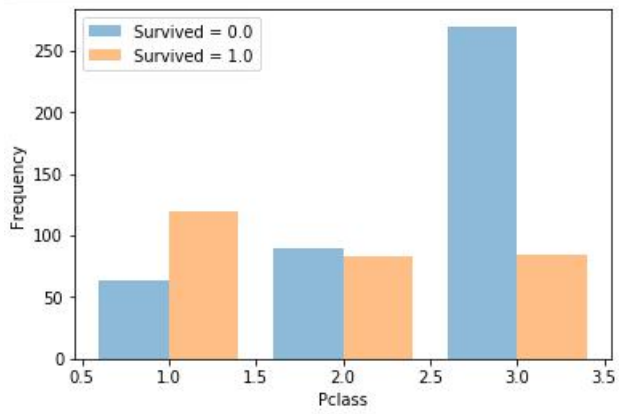
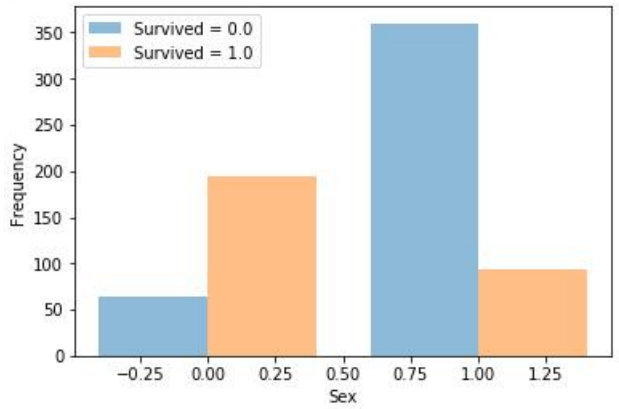
(a)



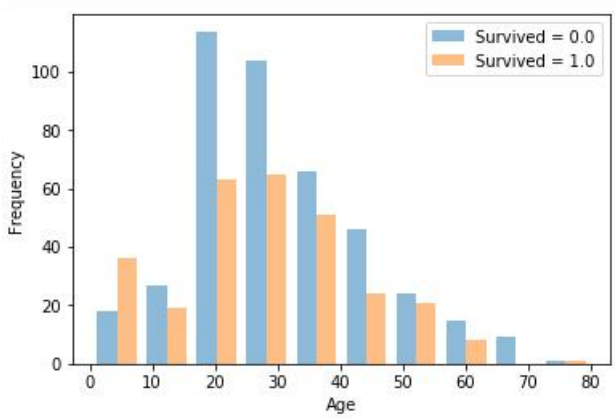
Pclass:

The first class has the highest survival rate. Then the second class. And the third class has the lowest survival rate.



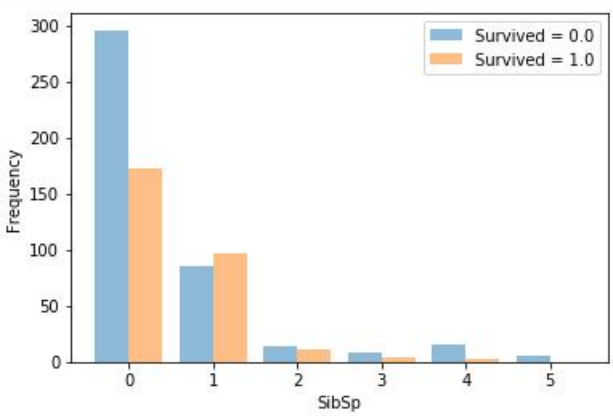
Sex:

Females' survival rate is higher than males.



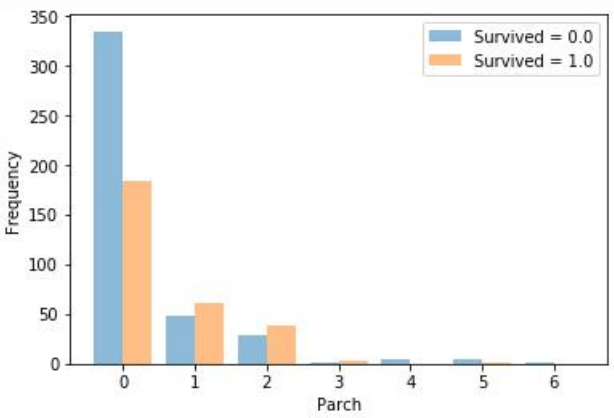
Age:

People whose age is around 20-50 has lowest survival rate. People under 10 are more likely to survive.



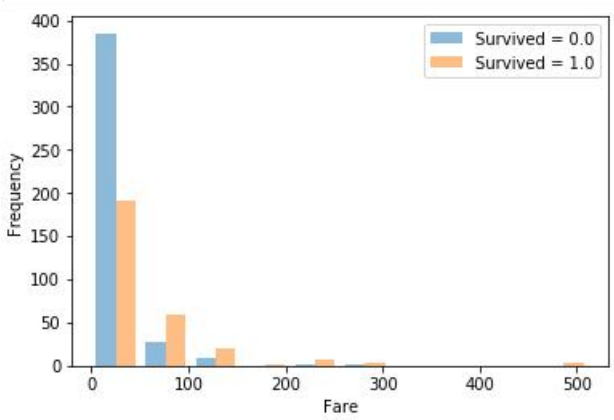
SibSip:

People who traveled with at least 1 sibling or spouse have higher survival rate than those who traveled alone.



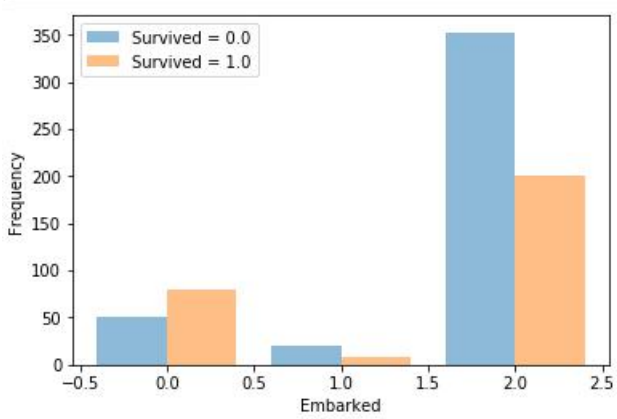
Parch:

People who traveled with at least one parent or child are more likely to survive than those who traveled alone.



Fare:

People who paid more for fare have a higher survival rate.



Embarked:

People embarking from Cherbourg have higher survival rate than embarking from other places.

(c)

The training error of this Decision Tree Classifier is 0.014

(d)

training error (n_neighbors=3): 0.167

training error (n_neighbors=5): 0.201

training error (n_neighbors=7): 0.240

(e)

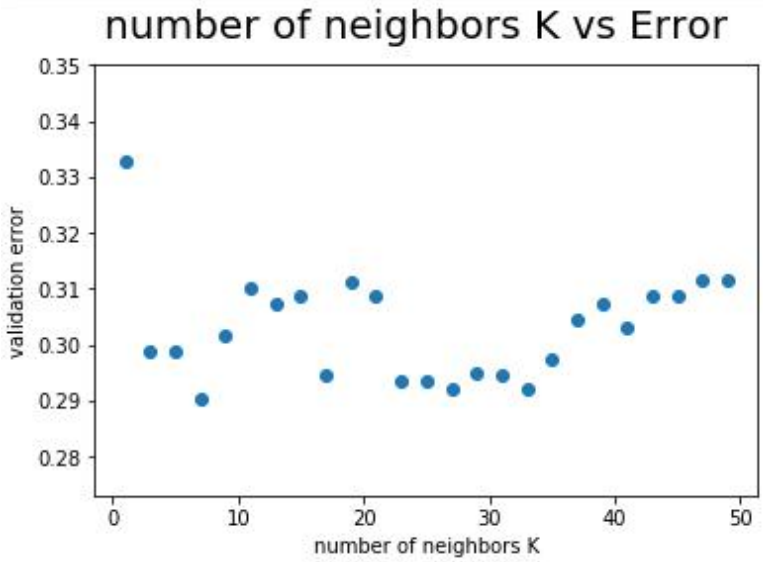
Training error and testing error for MajorityVoteClassifier are: 0.404, 0.407

Training error and testing error for RandomClassifier is: 0.489, 0.487

Training error and testing error for DecisionTreeClassifier is: 0.012, 0.241

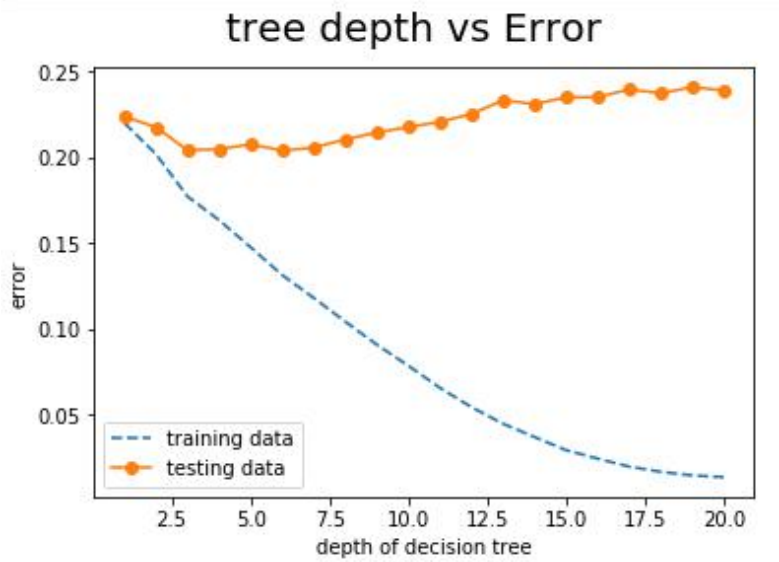
Training error and testing error for KNeighborsClassifier is: 0.212, 0.315

(f)



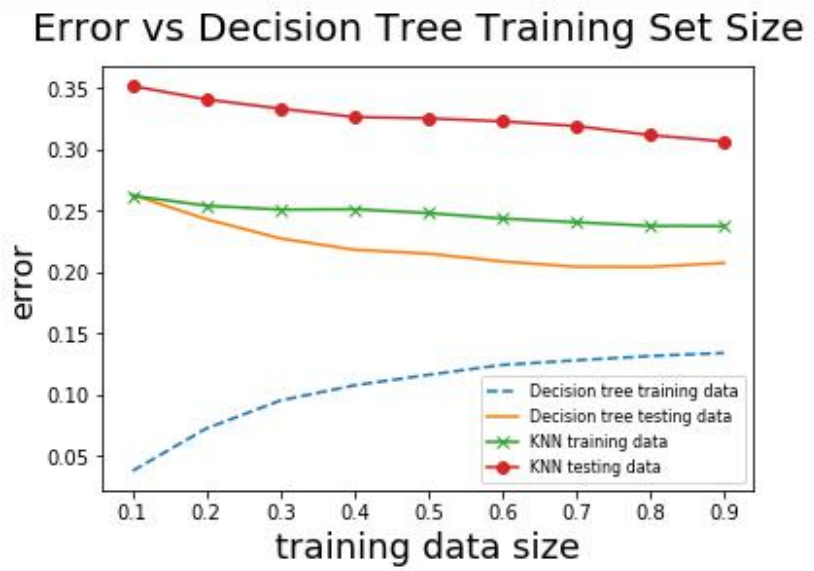
At first when k is bigger, validation error decreases. But later on we can observe overfitting, that is, when k is still getting bigger, validation error will increase. The best k is 7. The validation error for K=7 is 0.290

(g)



The best depth limit is 6. When the tree gets deeper, training error always decreases, but testing error will first decrease and then increase due to overfitting.

(h)



Decision tree training error increases as training data size increases, but still stays in a low level.
Decision tree testing error is high at first and decreases when training data size increases, but it doesn't decrease so much.
There is still a big gap between decision tree training error and testing error, and since training error is low but testing error is high, the model has a high variance. Maybe using more data or delete some features can solve this problem.

KNN training data starts at a high error level and as training data increases, it doesn't decrease much. KNN testing data error starts are a fairly high level and decreases as training data increases, but still in a high level.
Since both training error and testing error are in a high level, we may conclude that the model has a high bias now. Choosing some other features may help.