1. Consider the two plots for the data sets -400.csv. For each plot, what is the tree depth corresponding to highest test accuracy. Is the depth higher for entropy or gini impurity index? What is a potential reason for this?

Entropy: the depth=3
Gini: the depth=6
The depth for gini impurity is higher.
Potential reason: When using entropy (information gain) to split features, at each step the particular feature is selected which will give the purest children nodes. As a result, it takes fewer steps (tree depth) to complete splitting.

2. Consider the two plots for gini impurity index over the two datasets. For each plot what isthe tree depth corresponding to the highest test accuracy. Which of the two datasets has the higher tree depth? What is a potential reason for this?

400.csv dataset: the depth=6
200.csv dataset: the depth=3
The “400” dataset has the higher tree depth.
Potential reason: There are 200 more training examples in “400” dataset than “200” dataset. So it takes more steps to split features as purely as possible.

3. Which of two datasets -200.csv and -400.csv has a higher test accuracy? You can just take one depth, say 4 and look at the accuracy on the plots. What is the potential reason for this?

Entropy:
-400.csv dataset has a higher test accuracy.
Potential reason:
More training data makes the decision tree fit better.

Gini index:
-200.csv dataset has a higher test accuracy.
Potential reason:
Too much training data causes overfitting.