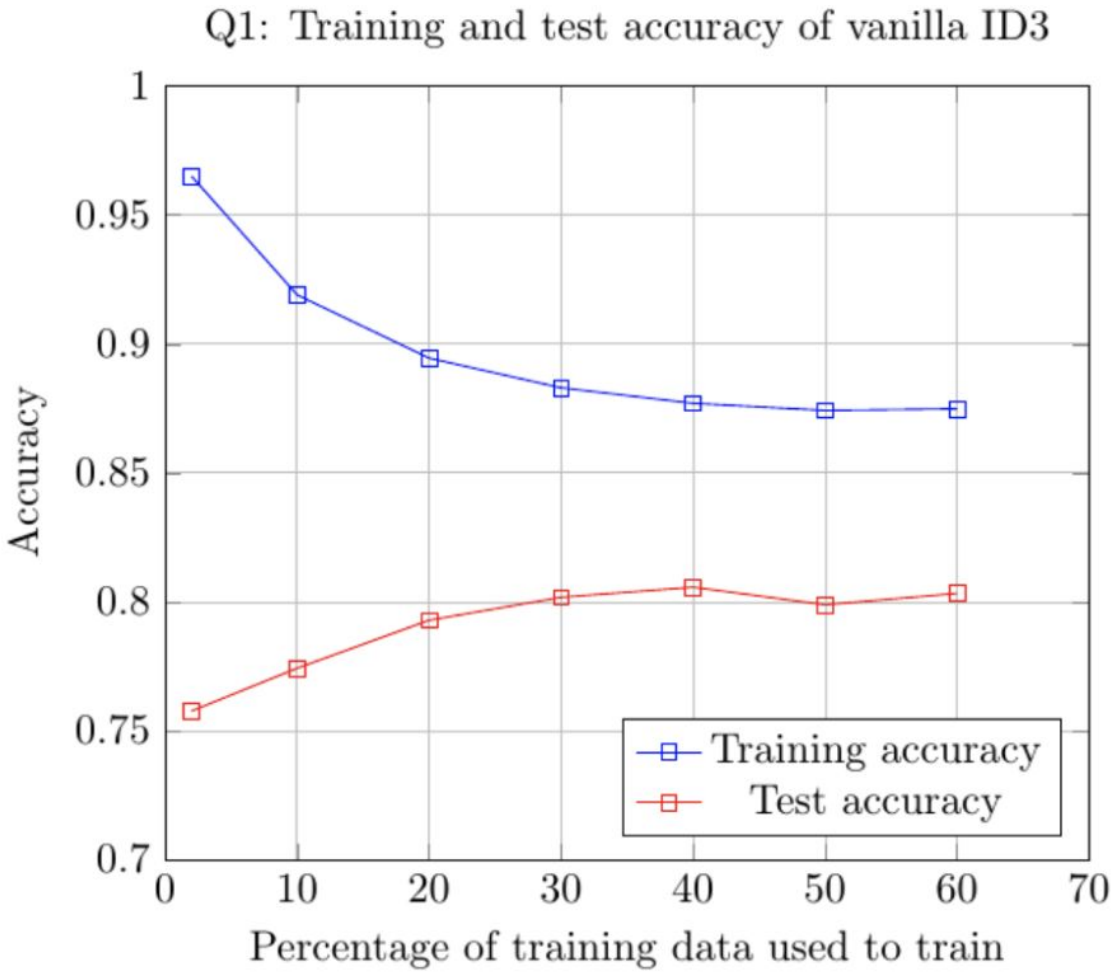
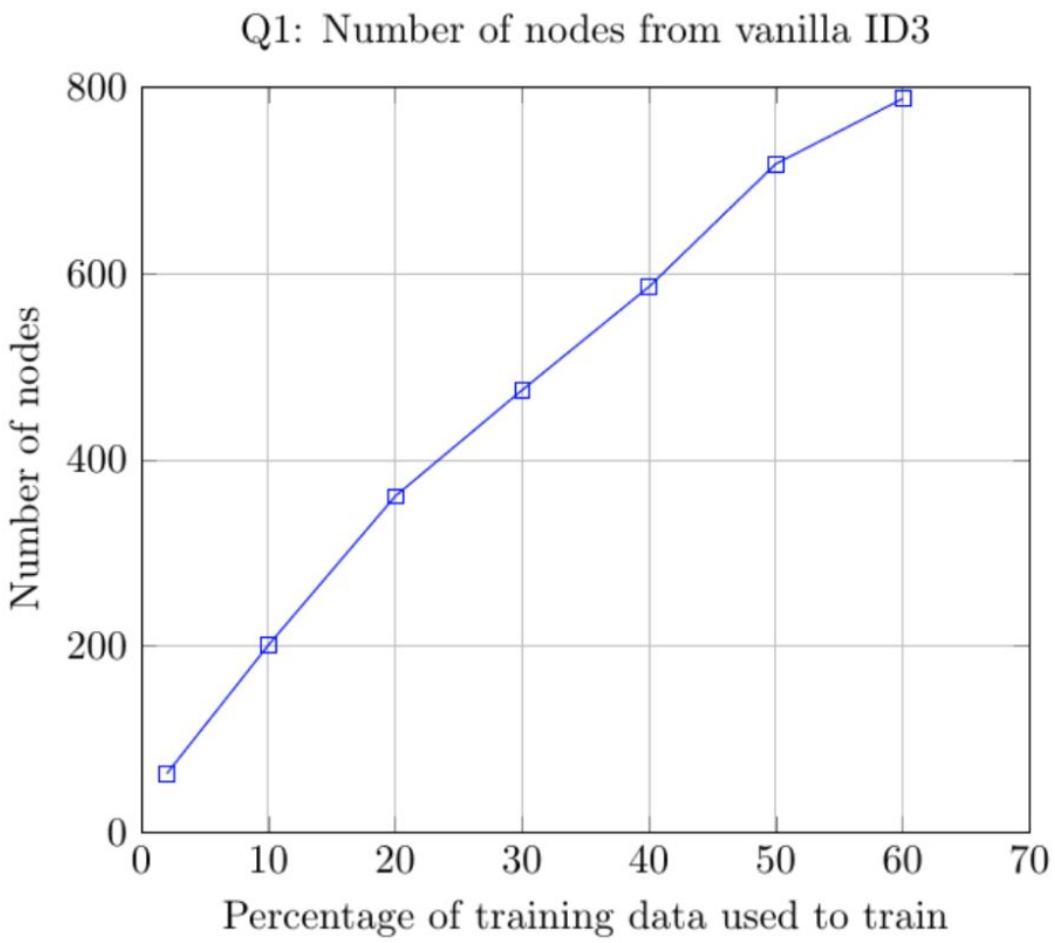
Raman Kahlon
kahlonr@purdue.edu
PUID: 00291-51712

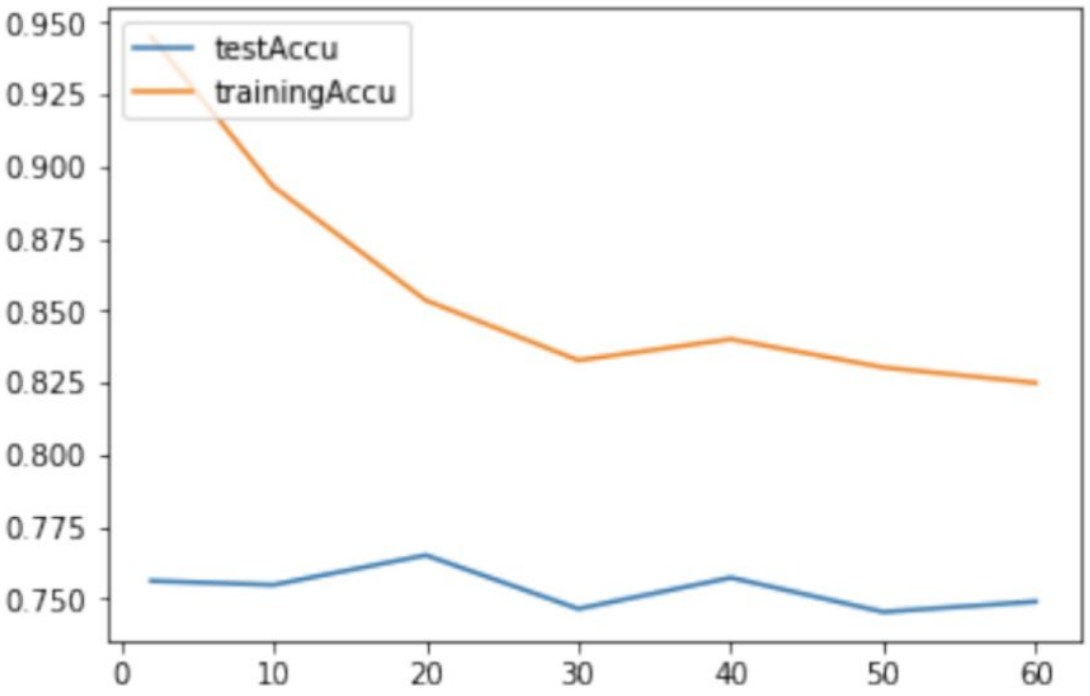
HW 2 (5 late days)

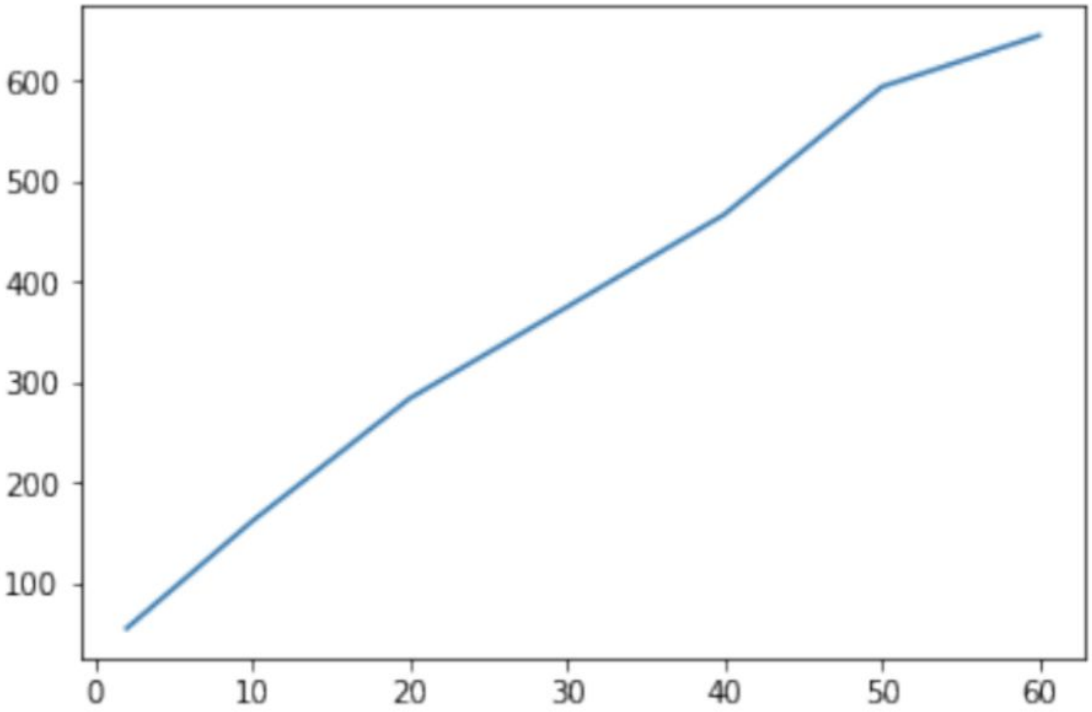
1.

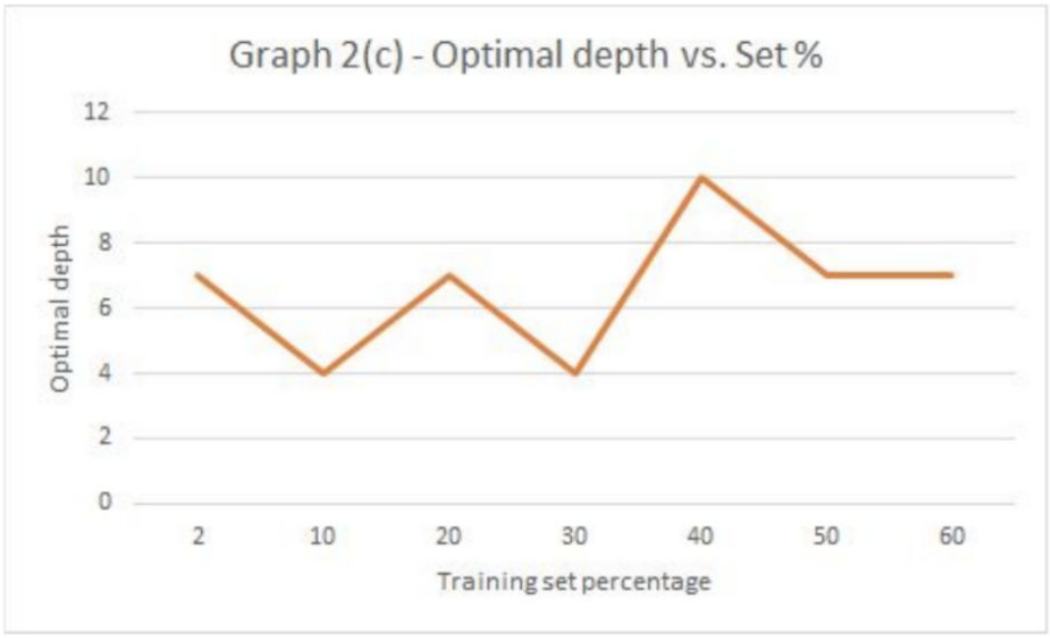



Q1: Number of nodes from vanilla ID3
Number of nodes
Percentage of training data used to train
0
200
400
600
800
0
10
20
30
40
50
60
70

2.

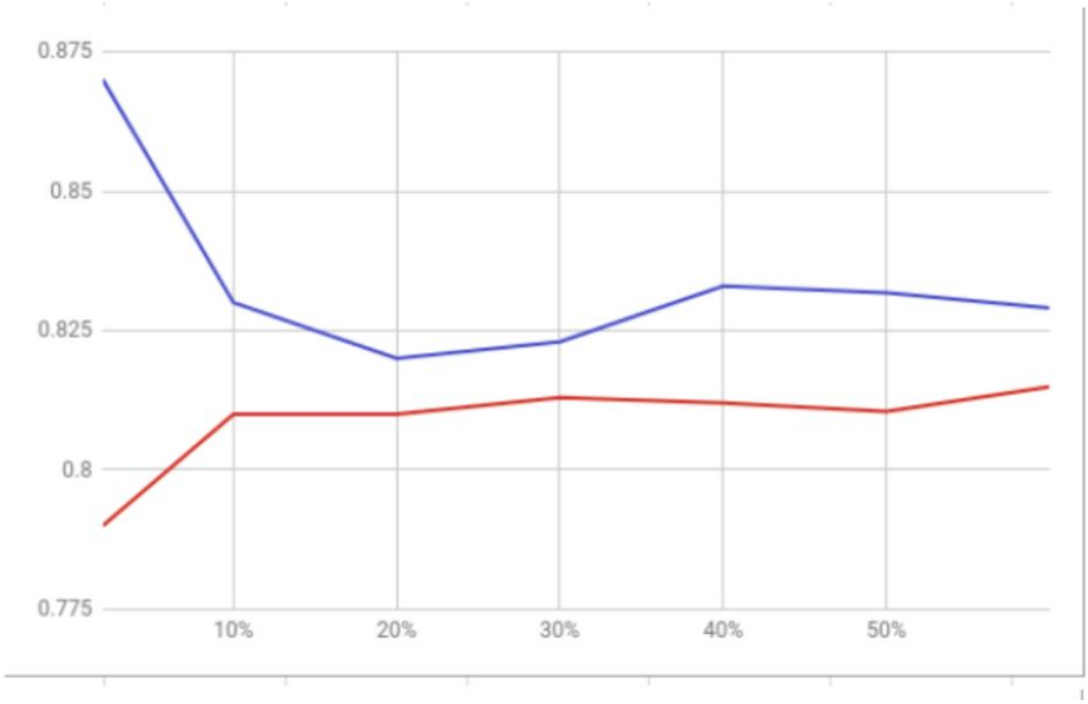

testAccu
trainingAccu
0.950
0.925
0.900
0.875
0.850
0.825
0.800
0.775
0.750
0
10
20
30
40
50
60

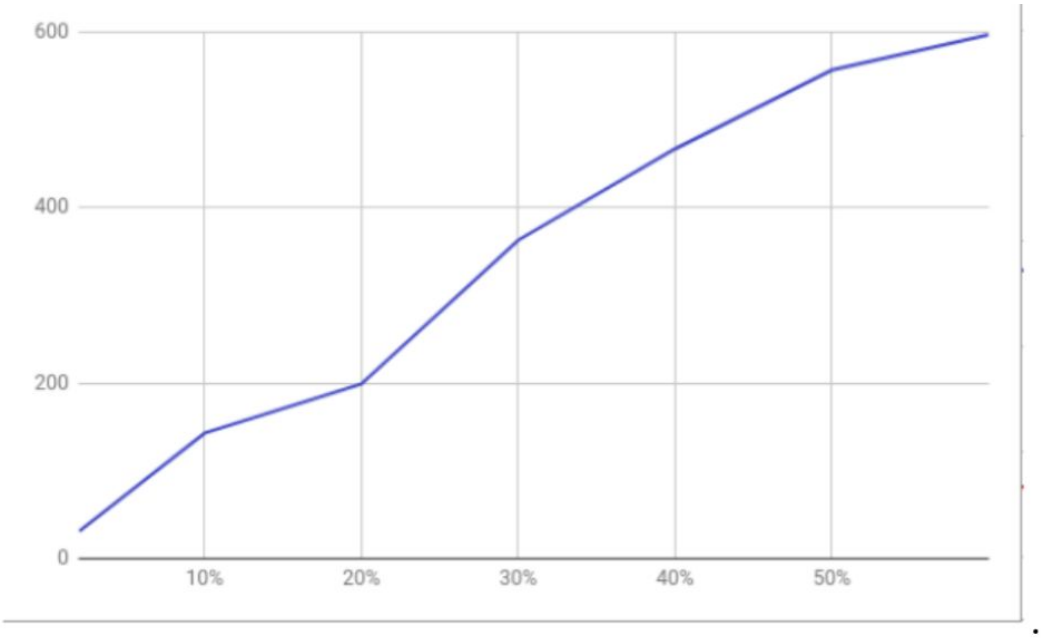

600
500
400
300
200
100
0
10
20
30
40
50
60


Graph 2(c) - Optimal depth vs. Set %
Optimal depth
12
10
8
6
4
2
0
2
10
20
30
40
50
60
Training set percentage

.

3.


0.875
0.85
0.825
0.8
0.775
10%
20%
30%
40%
50%



4. The testing set is only supposed to represent unseen data. If we prune it, then we would fit our model to it inadvertently, eventually getting an inaccurate measure of its performance on unseen data.

5. I would keep track of all possible class labels, instead of a single class label, and then output the class labels in order of decreasing frequency.