EECS 349 Problem Set 2 (David Wallach — daw647)

1. **Group members**: David Wallach (me), Nicholas Kotsiantos, Heath Reineke

2. We altered the node struct by adding:

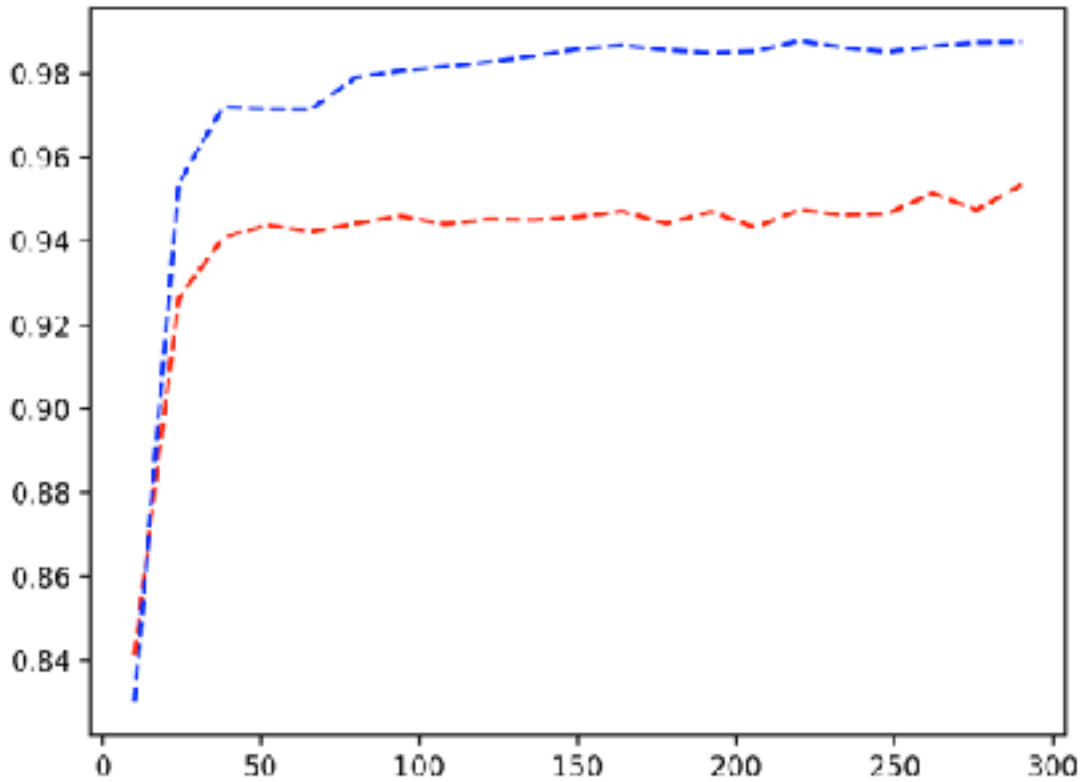
   - **branches** which is a list that contains lists of examples that correspond to values that have been sorted by the value of the attribute the node is splitting on. This was to recursively make the children nodes have the right subset of data

   - **classifiers** is used to determine if the node is a leaf by calling the method we added to the struct **is_leaf** which is then called pruning to make sure we do not try and prune a leaf node.

3. We handled missing attributes by finding the most common attribute values that correlate to the missing attributes class. We chose this strategy because with the limited information we had to make an educated guess. By using this method, we were able to take into account the entire dataset to develop our educated guess.

4. We performed pruning by looking at each node in the original tree. If it was a leaf node we would skip over it. Otherwise, we kept a dictionary of node labels with associated accuracy scores of the tree with that node pruned. If the maximum value of the dictionary is not greater than the base accuracy of the original node then stop pruning. Otherwise, we prune the node with the greatest accuracy on the test data. When we prune a node, we get the mode of all the examples under that node and make that the nodes value.

5.



[RED LINE = NO PRUNING I BLUE LINE = PRUNING]

a. As the size of the training data increases the accuracy percentage increases. This makes sense because as the training data increases, the ID3 tree can more accurately split on attributes by using more examples to train the data.
b. As the training set size increases, it gets to a point where it overfits the data as it tailors the tree too precisely to the data. However, the pruning gets more accurate as the training set increases because it is able to prune more nodes until it gets the optimal tree in which no over fitting occurs.