Author: Yatharth Kumar
Entry Number: 2020CS10413

# Decision Trees

## (a)


| MAX DEPTH | TRAIN ACCURACY | TEST ACCURACY | ONLY WIN ACCURACY | ONLY LOSS ACCURACY |
|---|---|---|---|---|
| 5 | 88.59% | 57.70% | 49.64% | 50.36% |
| 10 | 99.63% | 60.29% | 49.64% | 50.36% |
| 15 | 99.69% | 60.29% | 49.64% | 50.36% |
| 20 | 99.69% | 60.29% | 49.64% | 50.36% |
| 25 | 99.69% | 60.29% | 49.64% | 50.36% |

The accuracy levels for 'win only' and 'lose only' categories are close to 50 percent, which aligns with our expectations, as random guessing in a two-class classification problem typically yields a probability of success around 0.5. Examining the graphs, we can observe that the training accuracy improves as the maximum depth of the tree is increased, as anticipated. This behavior is due to the tree's ability to capture more information from the training data with greater depth.

However, when one-hot encoding is not employed, it was noticed that without a depth constraint, the tree grows fully to a depth of approximately 12. As a result, both the training and test accuracy levels plateau beyond a certain depth. Interestingly, the test accuracy also increases as the depth is raised. It's worth noting that the test dataset is relatively small, which may explain why some observations do not align with our initial expectations, such as the occasional decrement in test accuracy with increasing depth.

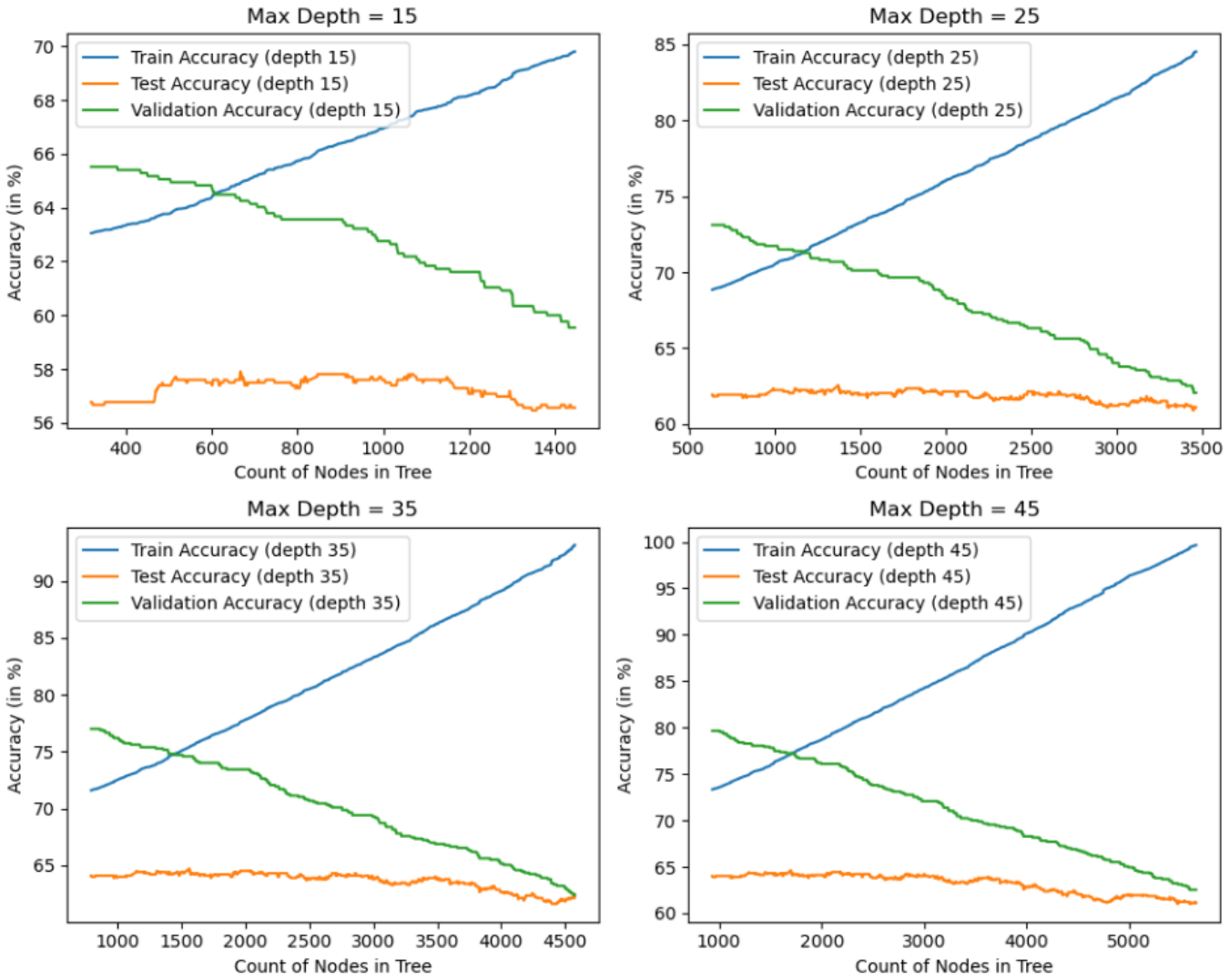
# (b)

The one-hot encoding data exhibits superior performance on the test dataset. Specifically, when the tree's depth is set to 25, 35, and 45, the achieved accuracies surpass those obtained without one-hot encoding. The one-hot encoding version of the data also follows the expected pattern as the maximum tree depth increases.

It demonstrates the best results around a depth of 35, after which the training accuracy notably increases due to the incorporation of more training information. However, this results in overfitting of the model on the training data, causing a corresponding decrease in test accuracy. It's important to note that, unlike in part A, the tree in this case doesn't grow to its maximum depth, primarily because there are a significantly larger number of attributes to consider. The input dimension is considerably larger than in the previous scenario.

The maximum depth is 48 for the fully grown tree.

# (c)

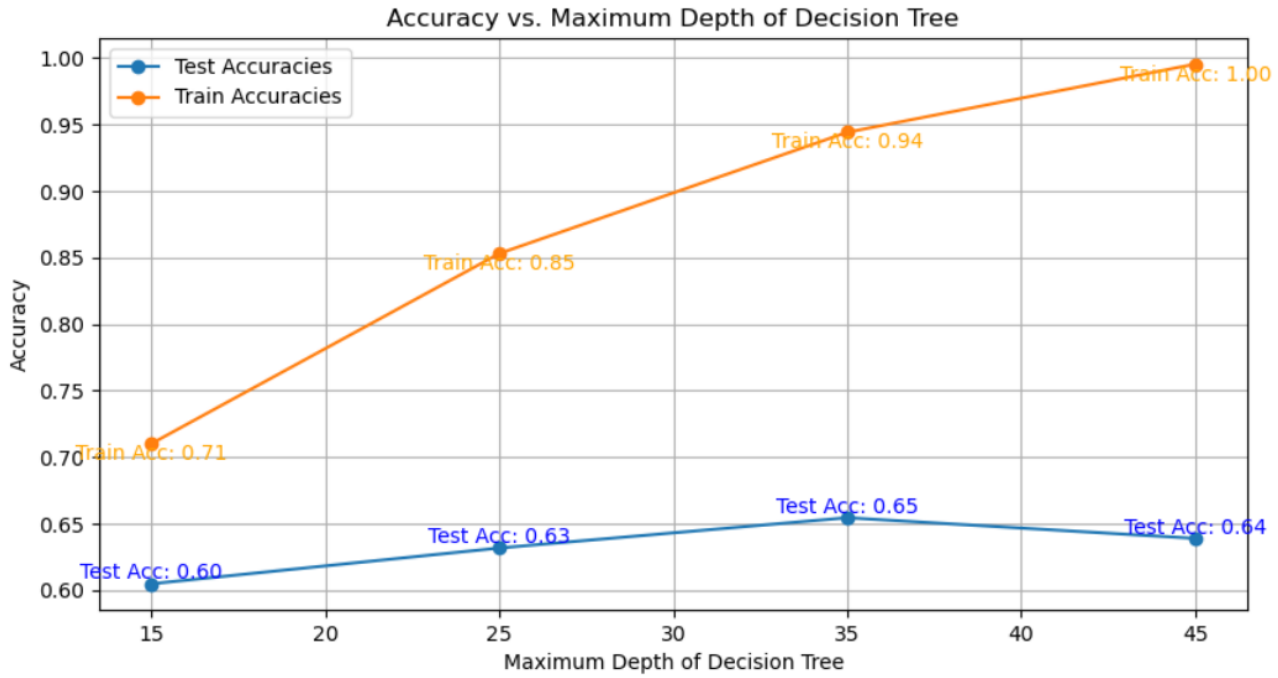


| MAX DEPTH | BEFORE PRUNING ACCURACY | POST PRUNING ACCURACY |
|---|---|---|
| 15 | 55.33% | 56.77% |
| 25 | 61.01% | 61.94% |
| 35 | 62.67% | 64.12% |
| 45 | 61.12% | 64.01% |

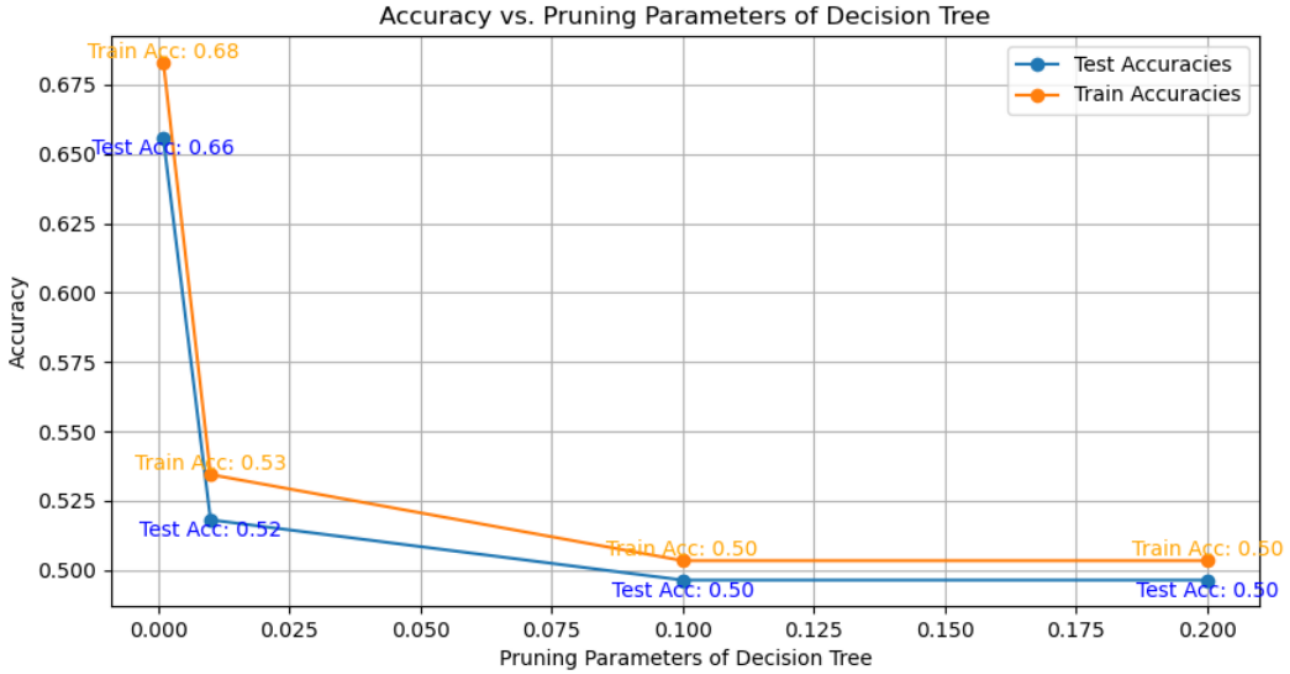
The obtained graphs align closely with our expectations. When we maintain a fixed depth and allow the tree to grow fully, the training accuracy steadily increases. Essentially, the tree progressively incorporates more and more training data until the training accuracy reaches 100 percent, indicating complete overfitting. Subsequently, we employ validation data and commence pruning, removing nodes at each step that contribute the most to an increase in accuracy on the validation data. This process grows the entire tree and then prunes away unnecessary overfitting. Consequently, the curves for training and validation accuracy exhibit a monotonically increasing or decreasing pattern. They do not exhibit fluctuations, maintaining a consistent direction of change.

The behavior of the test data improves after pruning, which is also evident in the graph. Among various tree depths, it becomes apparent that as the depth increases, the training data overfits more severely. Depth values of 25 and 35 appear to yield the best test accuracies after pruning. In contrast, a depth of 15 underperforms due to underfitting.

# (d)




| DEPTH | ACCURACY |
|---|---|
| 15 | 57.70% |
| 25 | 61.72% |
| 35 | 61.61% |
| 45 | 62.30% |



| ALPHA | ACCURACY |
|---|---|
| 0.001 | 62.76% |
| 0.01 | 50.00% |
| 0.1 | 47.36% |
| 0.2 | 47.36% |

**Best Depth**: 45
**Best Alpha**: 0.001
**Accuracy**: 65.5635987590486%

The accuracy remains consistent across the test, training, and validation datasets, and the results obtained by the decision tree that utilises the library are remarkably similar. The trend observed in the depth graph is also reflected in the graph produced in the c portion. A rise in the maximal tree depth is accompanied by a subsequent decline in the accuracy of the test data. The precision of the training data increases in a similar monotonic fashion. CCP alpha is a parameter that signifies the extent of pruning. It is apparent that the level of accuracy is greatest when pruning is minimal to nonexistent. However, as pruning becomes more extensive, there is a substantial decline in accuracy, which eventually approaches 0.5, representing a random chance prediction. Nearly 0.675 is the accuracy at ccp=0.001, which our model is incapable of producing. Consequently, the efficacy of a decision tree classifier is substantially influenced by the degree and depth of pruning.

# (e)

**Best Hyperparameters**: {
'max_features': 0.9,
'min_samples_split': 10,
'n_estimators': 350
}
**Training Accuracy**: 96.93369106937524 %
**Test Accuracy**: 72.49224405377456 %
**Validation Accuracy**: 70.91954022988506 %
**Out-of-Bag Accuracy**: 72.68429794301776 %

The results obtained in this section for testing and substantiation are more precise than the combined accuracy of the results obtained in the previous two sections of this investigation. This showcases the efficacy of random forests.

The collective output of multiple decision trees yields a considerably more accurate prediction in comparison to employing a solitary tree, due to the fact that each tree collects a unique type of data information.

The training accuracy has decreased from nearly 0.99 in the previous result.