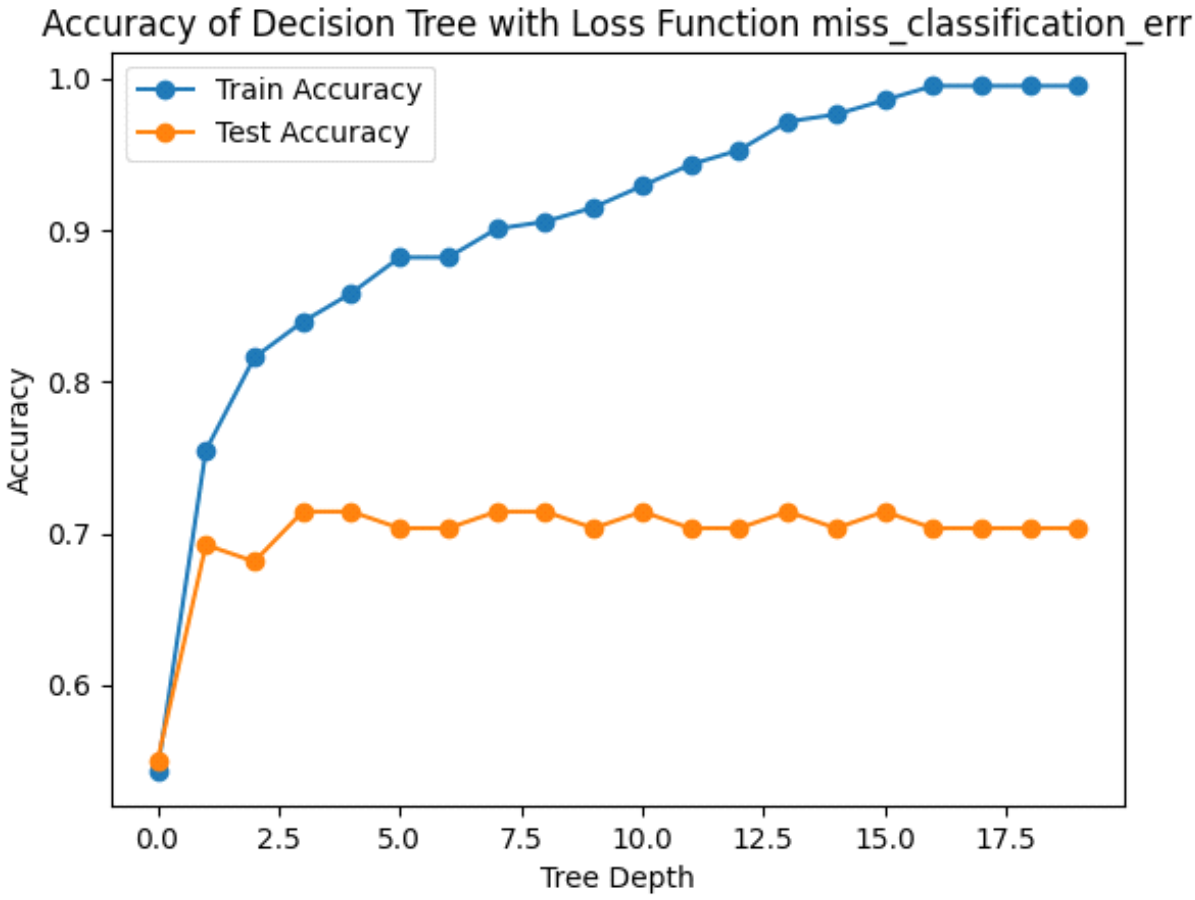
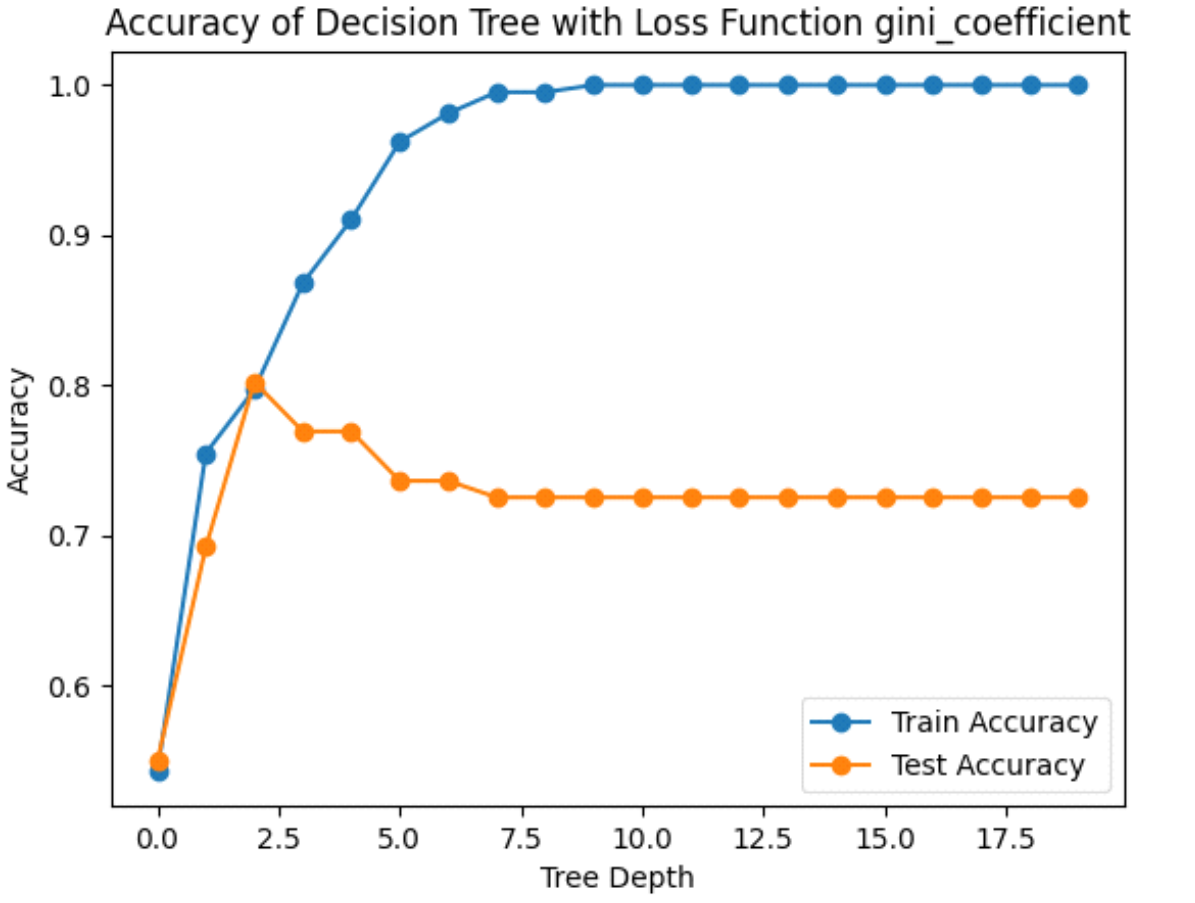
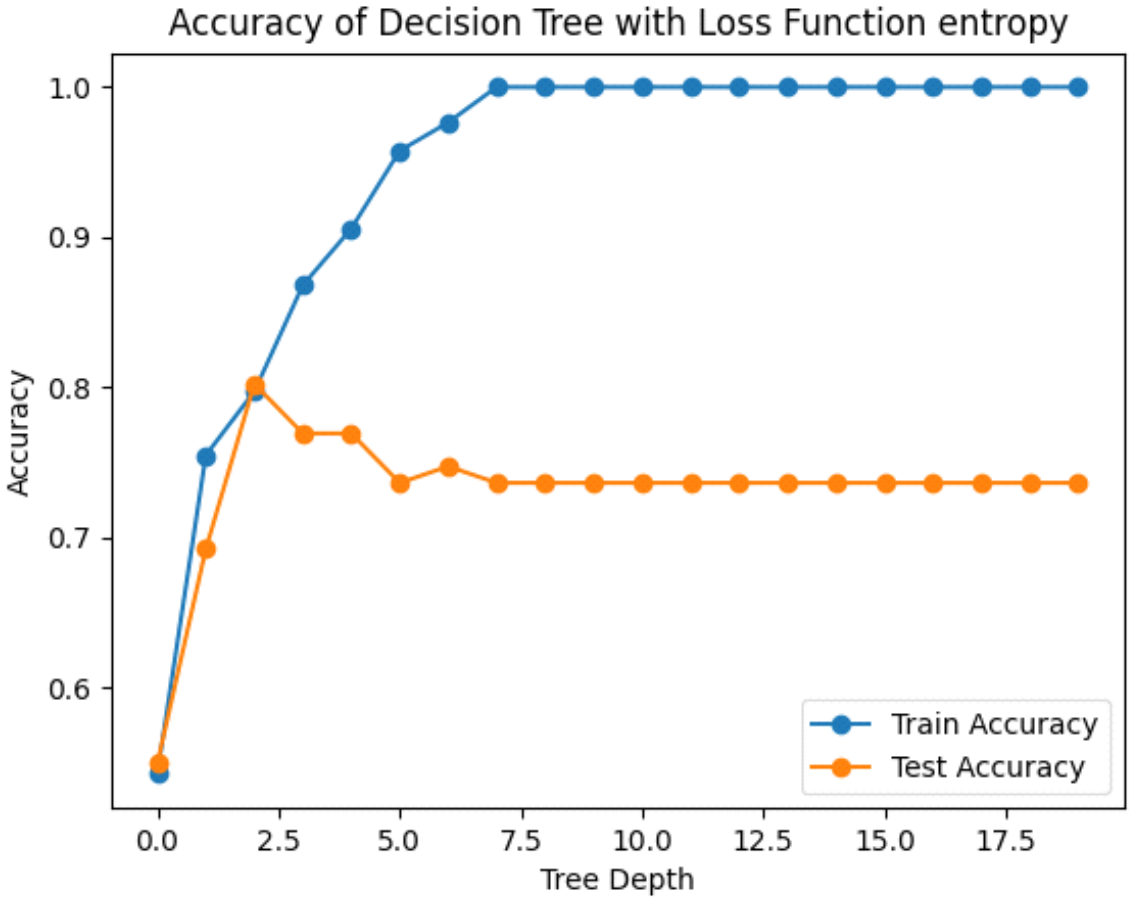
A3Q1

June 29, 2022 7:42 PM

The graphs are as follows:







From those graphs, we can tell that the misclassification error loss function gives the worst performance. It will give a very deep decision tree (as the train accuracy did not reach 100% before depth 15), and relatively lower test accuracy than the other two loss functions. The entropy loss function gives a shallower tree (less depth) than the gini index loss function does, but the test accuracy of those two follows a similar trend and do not vary that much.

For entropy and gini index loss functions, the test accuracy increases rapidly at the beginning, reaching 0.8-is, but then drop to around 0.75 as the maximum depth grows, and become stable after the maximum depth reaches 8. The increase at the beginning is because that our model starts to work, and the drop later indicates an overfit. The fact that the train accuracy keeps increasing as the maximum depth grows until it reaches its limit, which also indicates that an overfit may happen.

For the misclassification error loss function, the train accuracy keeps growing up to 100%, while the test accuracy increase at the first, then keeps validating around the line of 70%. This is also because that our model starts working at the beginning and then we encountered some overfitting as the maximum depth grows. And the performance of misclassification is also worse than the other two.

The train accuracy of all three curves will reach 100% eventually, which indicates an overfitting. In fact, as long as the train accuracy reaches 100%, the tree has reached the maximum depth. Although I still train the model for more times which means I treat it as it will still grow, the actual tree will stop growing, and that is why the curves will become stable at the end.

B)

Bagging accuracy data: median: 81.32%, min: 78.02%, max: 82.42%
Random Forest accuracy data: median: 81.32%, min: 79.12%, max: 83.52%

The difference between these two methods are not significant, but the random forest has a slightly higher min and max accuracy than the bagging method. And both methods have a better accuracy than the previous three methods when the maximum depth is 3.

Both methods are better than the previous three because that we used bagging in both methods given that we have a relatively small dataset, so we can decrease the variance and increase the accuracy.
Random forest accuracy will behave slightly better than the bagging strategy because the RF method will further decrease the variance. The RF method randomly chooses features so that the result will not highly rely on one single feature, adding more randomness so adding more independency to the data. This can reduce the variance and thus increase the test accuracy / performance.