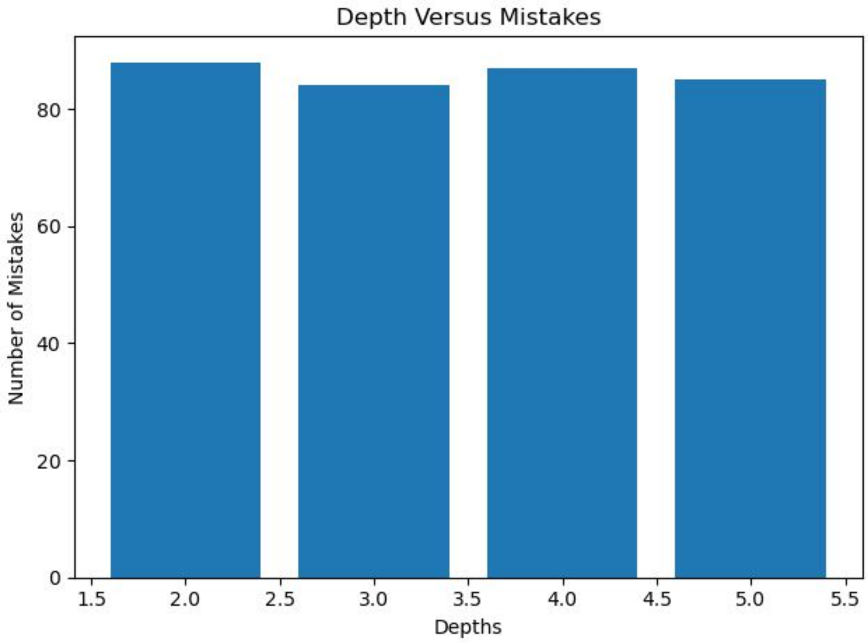
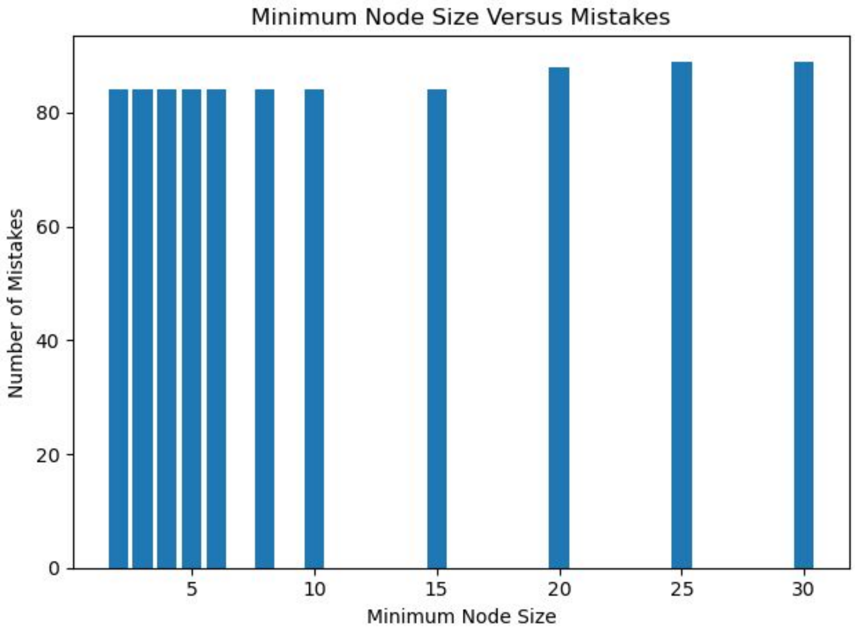
Bahdah Shin

Samuel Adams

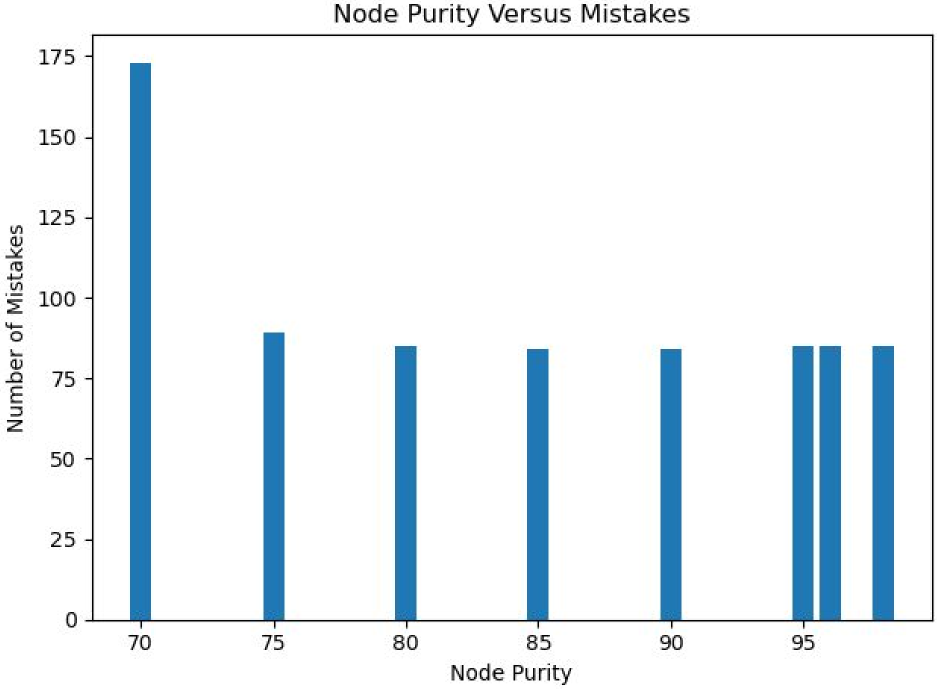
a. **Best minimum Node Depth:** 3



b. **Best Minimum Node Size**: 15



c. **Best Purity**: 85%



d. The csv file name is HW05_Adams_MyClassifications.csv

e. The homework 7 explores the n fold cross validation. This means we will split the training data n times into batches and select one of the batches as the test data while the rest becomes the training data. We will test each variation where we try out each batch as a test data. The decision tree is trained from the training data. Then the number of mistakes is calculated by running the test data on the trained decision tree. The total mistakes are added from n variations. The number of mistakes can be compared with the parameter change.

   There are three parameters that can be changed: the depth, minimum number of data records, and node purity. The depth range was [2, 3, 4, 5], minimum number of data records was [30, 25, 20, 15, 10, 8, 6, 5, 4, 3, 2], and the node purity was [70, 75, 80, 85, 90, 95, 96, 98]. For each range, in each element, the total mistakes are calculated, where the lowest number of mistakes indicate the best parameter.

   The best three parameters are derived from the 400 data points. With these parameters, we trained the tree with the 16000 data points. Finally, we checked the accuracy of the tree by providing the same 16000 data points. The accuracy resulted in 55%. I assume that we got a low accuracy because the 400 data points does not reflect the entire data set. In addition, the split was the first 400 data points rather than a balanced split. We didn't have enough time to get to balancing the split. If there was only one column, we would have sorted the list then split. However, the data has multi-channels, which makes the homogenized split more tricky.

The output for the graphs were underwhelming. In graph one, the error values are within the 4 error values. The error values go up and down but it seems that the values are too close together to show meaningful values. In graph two, the values were within 5 error values, and most values have the same error values. In graph three, we have a really high error value for 70% node purity while the rest sits within 5 error values. Four of the node purity has the same error values.