Munir Vafai, Collins Kariuki

**1. What is the accuracy of the random classifier on the Titanic data set from assignment 1. To calculate this, generate a random 80/20 split (using dataset.split(0.8)) train the model on the 80% fraction and then evaluate the accuracy on the 20% fraction. Repeat this 100 times and average the result (hint: do the repetition in code :).**

We got 0.4967832167832166.

**2. What is the accuracy of your decision tree classifier on the Titanic data set with unlimited depth. As above, average the results over 100 random 80/20 splits.**

We got 0.737482517482517.

**3. What is the best depth limit to use for this data? To answer this, do the same calculations as above (average 100 experiments), but do it for increasing depth limits, specifically 0, 1, 2, ..., 10. Show all of your results.**

Depth: 0 Accuracy: 0.5923776223776219

Depth: 1 Accuracy: 0.5907692307692306

Depth: 2 Accuracy: 0.59020979020979

Depth: 3 Accuracy: 0.5804895104895104

Depth: 4 Accuracy: 0.6151048951048947

Depth: 5 Accuracy: 0.6878321678321678

Depth: 6 Accuracy: 0.7422377622377618

Depth: 7 Accuracy: 0.7334265734265734

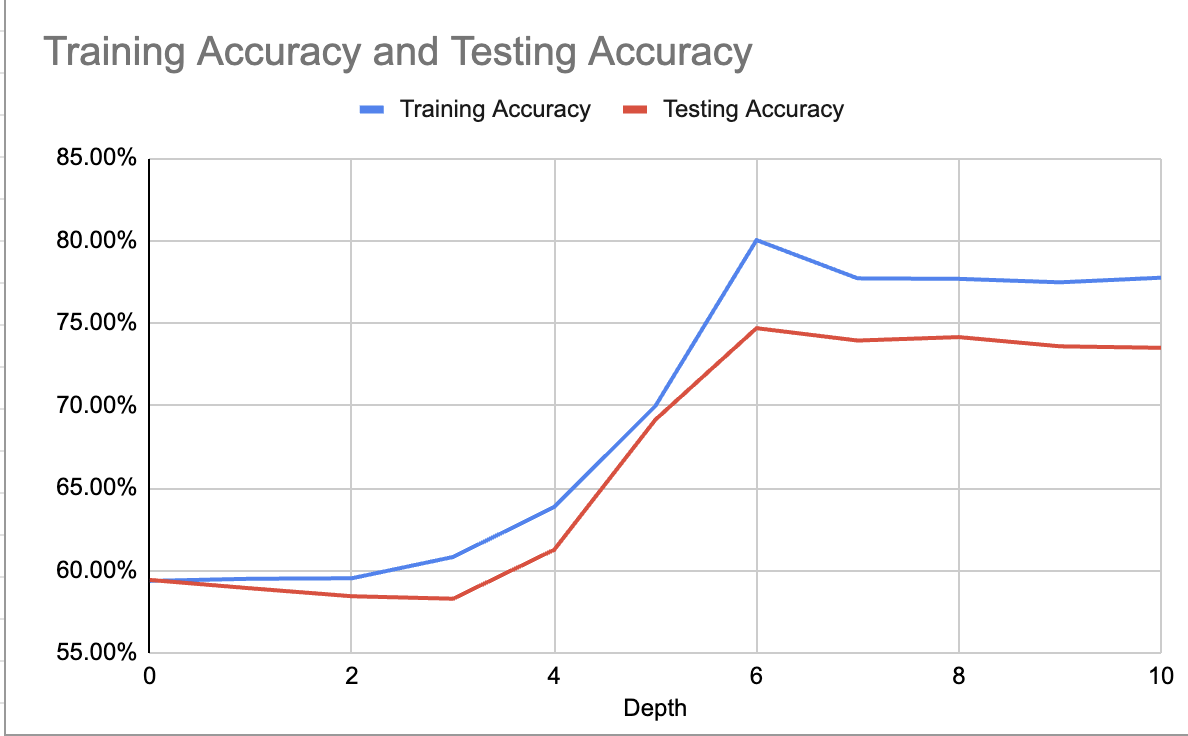
Depth: 8 Accuracy: 0.7425174825174821

Depth: 9 Accuracy: 0.7355244755244754

Depth: 10 Accuracy: 0.7410489510489507

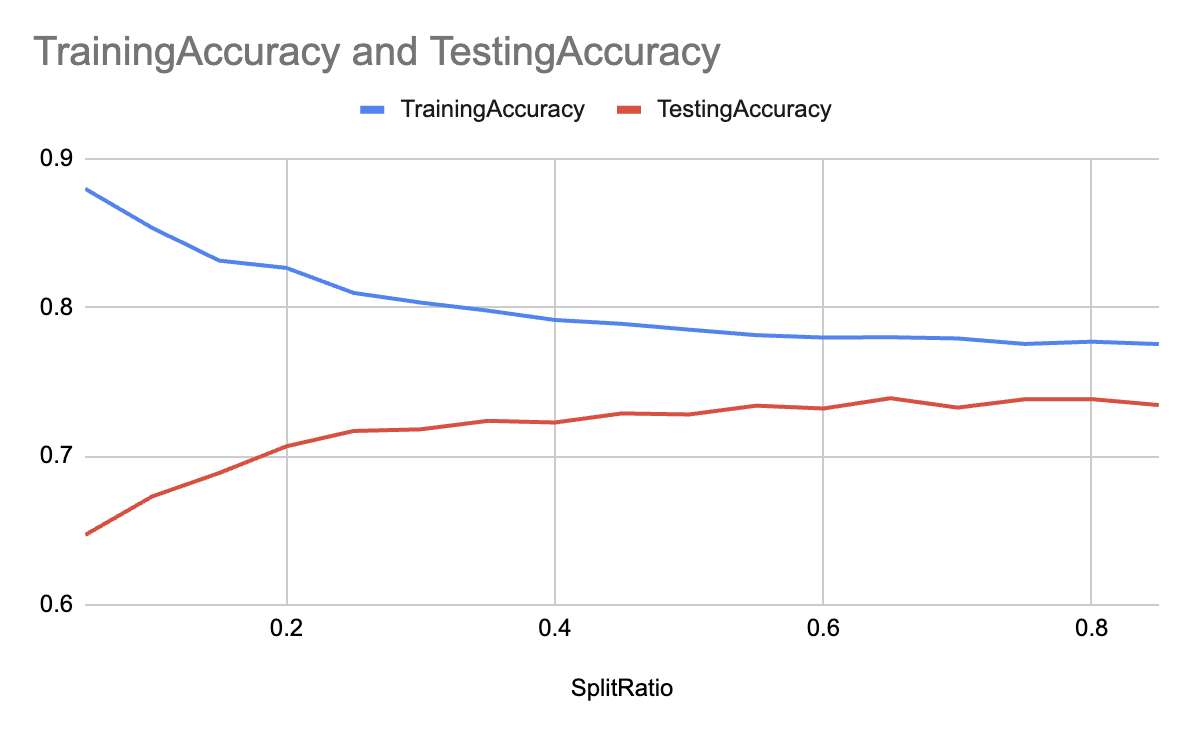
Ideal depth limit: 8 with accuracy: 0.7425174825174821

**4. Do we see overfitting with this data set? Repeat the experiment from question 3 with increasing depth (0, 1, ..., 10) and calculate the accuracy this time on both the testing data (like before) and the training data. Create a graph with these results and then provide a 1-2 sentence answer describing the graph.**



As the decision tree depth increases, both training and testing accuracies rise until reaching a depth of 6. Beyond this depth, accuracies decline and level off. However, this contradicts our expectations. While we anticipate initial increases in both training and testing accuracies, we do not anticipate the training accuracy to plateau or decrease. Testing accuracy aligns with expectations as it decreases due to overfitting over time.

**5. How does the amount of training data affect performance? To answer this, do the same calculations as above (average 100 experiments), but start with splits of 0.05 (5% of the data used for training) and work up to splits of size 0.9 (90% of the data used for training) in increments of 0.05. For these experiments use full depth trees, i.e. trees without any depth limit. Create a graph with these results and then provide a 1-2 sentence answer describing the graph.**



As the split ratio increases, indicating a larger proportion of training data compared to testing data, we observe a decrease followed by stabilization in training accuracy, which is unexpected. We would typically expect training accuracy to improve as we have more data for training. In contrast, testing accuracy increases and stabilizes as anticipated. This aligns with expectations because, with a higher split ratio and less testing data, the model becomes more capable of accurately classifying labels in the smaller testing dataset as it has more training data to train on.

**6. What does the training data size experiment tell us?**

The training data experiment underscores the importance of not repeatedly testing your model on the test data, as it can lead to overfitting. This is evident in question 3, where accuracy initially increases but then decreases beyond a depth limit of 8 due to overfitting. While training accuracy may improve with more nodes or greater depth, the same may not hold true for testing data. This experiment highlights the necessity of using development data during the decision tree model creation process to prevent overfitting to the testing data.

Simultaneously, it's essential to grasp the lower threshold of your split ratio. Ensuring a sufficiently large sample size for your training data is crucial, as it enables your model to acquire adequate knowledge. This, in turn, prevents negative impacts on testing accuracy resulting from a small training dataset, which could lead to underfitting.