Homework 4
Problem 1
Ayad Masud
733009045

I certify that I have personally done the coding, generated the figures and written the report without aid from anybody else, and that I have not plagiarized, self-plagiarized, or used AI-generated text. I certify that I have acknowledged any sources I used to complete this assignment. ARM

1 Part 1: Classification Tree with Pruning

Figure 1 through figure 5 shows a decision tree with increasing tree depth. We can see for figure 1 with a depth of one the biggest factor to tell whether a passengers survived is whether the passenger is male or not. In figure 2 the next biggest factor that is used to tell whether a passenger survives or not is SibSp and age. As the depth of the tree increases and there are more nodes the tree uses more and more features to determine which way to split. This results in increased complexity which we can see in figure 4 and figure 5 with depth of 5 and a full tree respectively. This is a time where we would want to figure out a way to make things simpler, for example, pruning the tree. Pruning the tree would results in an easier to read tree while conserving the accuracy we gained by having lots and lots of nodes.

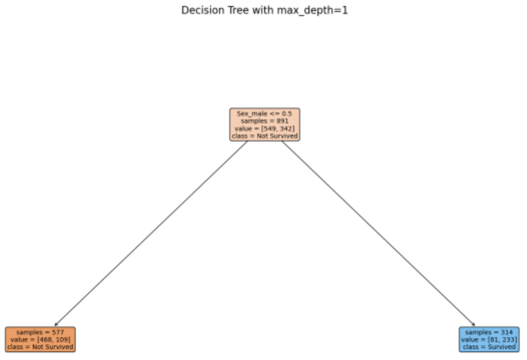

Figure 1. Decision tree of depth 1.

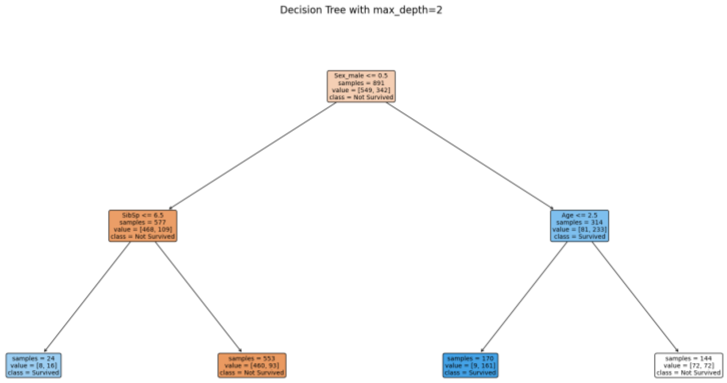

Figure 2. Decision tree of depth 2.

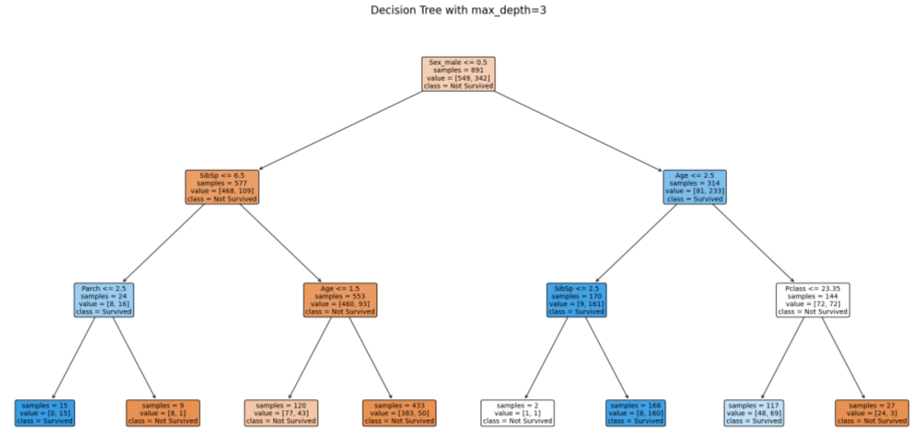

Figure 3. Decision tree of depth 3.

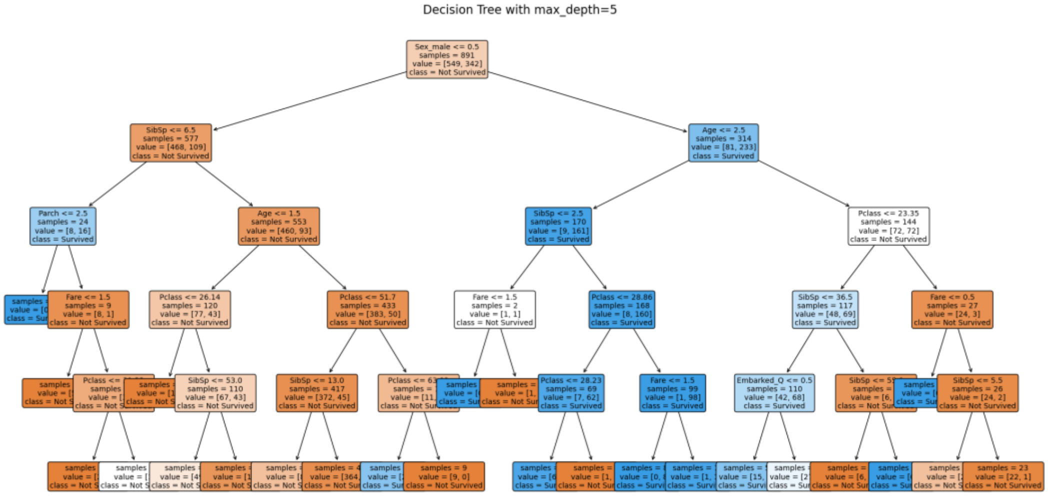

Figure 4. Decision tree of depth 5.

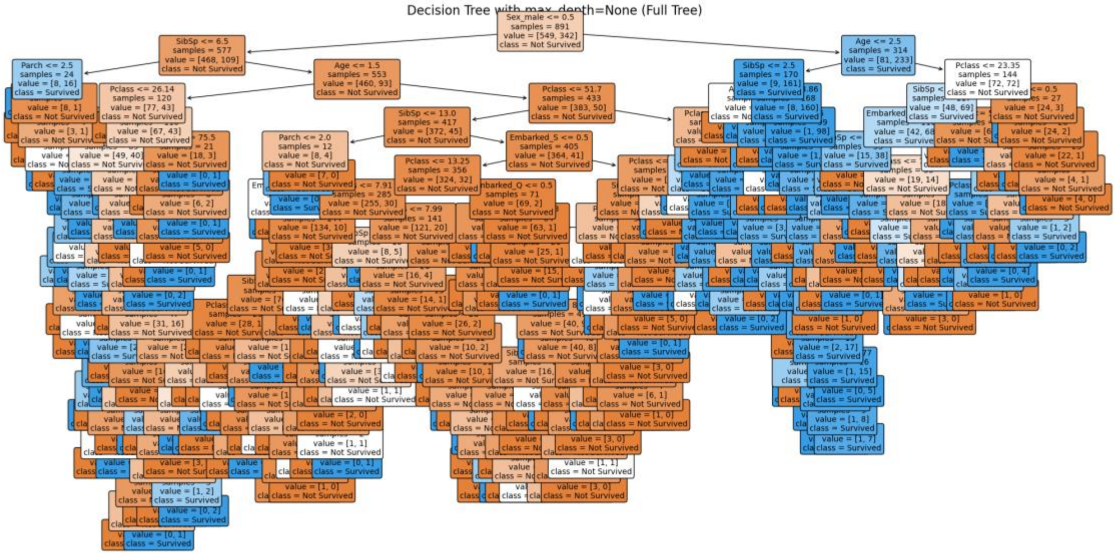

Figure 5. Decision tree of depth None.

Figure 5. Decision tree of depth full tree.

2 Part 2: Classification Tree without Pruning

Figure 6 shows the classification accuracy by using a training and validation split starting from 10/90 to 90/10. The graph shows that gap between training and validation accuracy as the training set ratio increases. As we can see the gap between the two sets decreases as we increase the training set ratio indicating less overfitting. This is expected as more training data usually leads to better generalization. We can also see an interesting trend in the training data as the training set ratio increases. We can see that the outliers in the box plots become smaller and smaller. Overall, we can say that as training set ratio increases the accuracy of the classification tree also increases and therefore results in a better performing model.

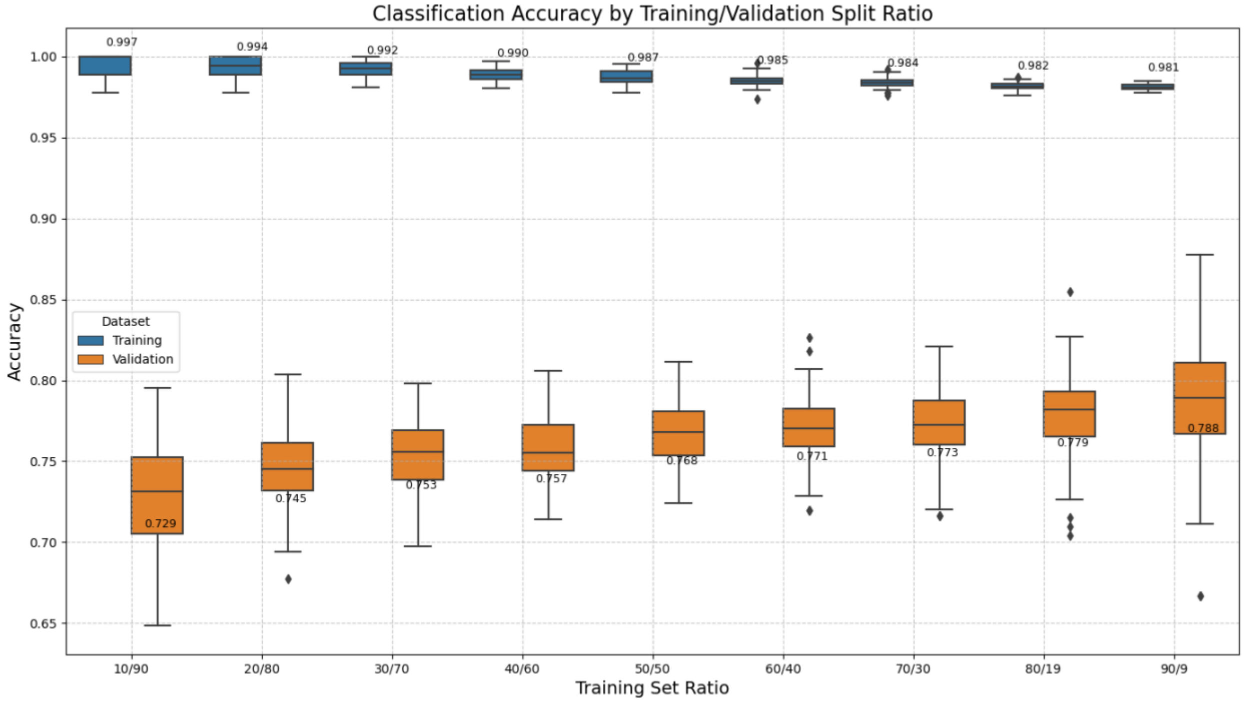

Figure 6. Classification accuracy by training/validation split ratio

3 Part 3: Classification Tree with Train/Validation Ratios

Figure 7 shows the classification accuracy vs tree size. The relationship between the number of nodes in the pruned decision tree and the classification accuracy is clearly seen and indicates the model's performance as complexity increases or more nodes are added to the tree. We can see that as the number of nodes increases, the training accuracy tends to increase, while the validation accuracy stays more or less the same and maybe shows more variability indicated by the spread of the outliers in each of the box plots. We can also see that after a certain point adding mores nodes is not really going to help accuracy, it levels out.

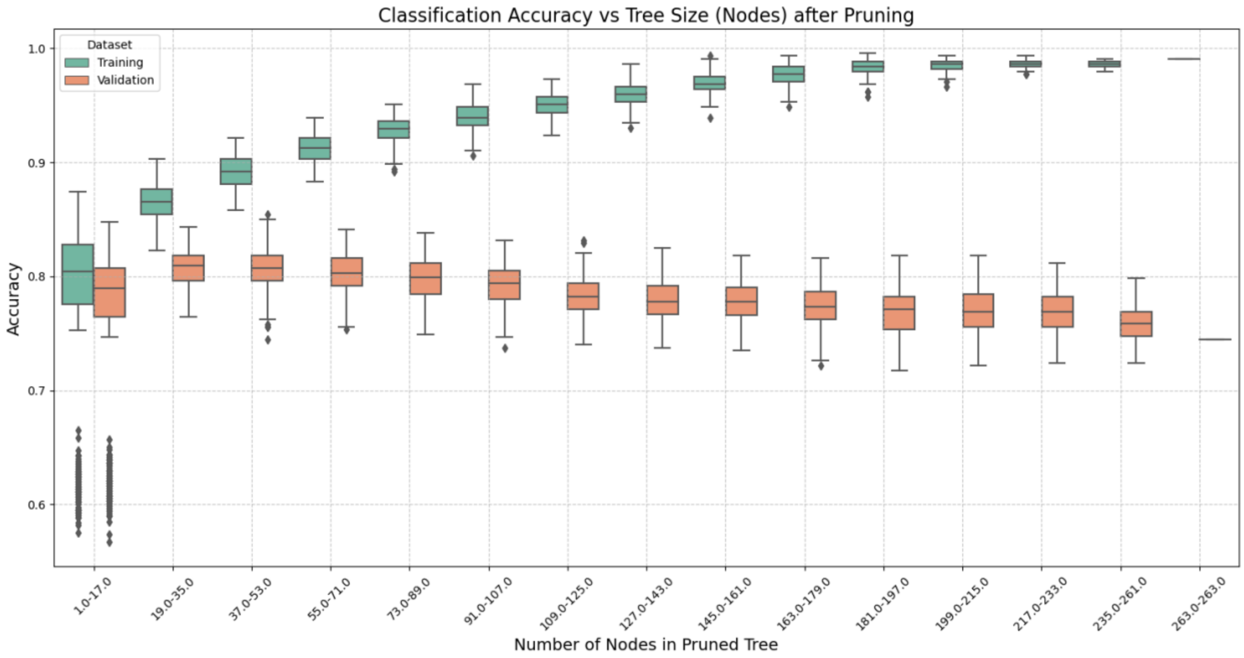

Figure 7. Classification accuracy vs tree size.

4 Resources used to achieve this goal

Canvas: Homework template

Python Libraries: NumPy, pandas, seaborn, matplotlib, sci-kit learn

5 References

scikit-learn. “1.10. Decision Trees — Scikit-Learn 0.22 Documentation.” *Scikit-Learn.org*, scikit-learn.org/stable/modules/tree.html.