Sarah Olsen
BSAN 6070 - CA03

Question 1.1

Why does it makes sense to discretize columns for this prediction problem?

In this instance, discretisation helps the model handle outliers in each demographic column. By creating a set of contiguous intervals that span the range of variable values, it places outlier values into the lowest or highest interval, with the remaining values of the distribution in their respective 'inlier' bins.

Question 1.2

What might be the issues (if any) if we DID NOT discretize the columns?

If we used continuous data in the model, we might disproportionately value attributes that have outliers. In this case, the program might try to reduce the continuous variables into a more manageable range. It would make the program more efficient, but the outcomes less accurate.

Question 7.1

How long was your total run time to train the model?

I can't figure out how to record the time it takes to train the model.

Question 7.2

Did you find the BEST TREE?

Yes! My best tree was Tree8, with the highest accuracy (84.4%), recall, precision, and F1 score.

Question 7.4

What makes it the best tree?

This is the best tree because it is the most accurate. It predicts the correct category more often than the other models, while maintaining high recall, precision, and F1.
Question 10.1

What is the probability of the outcome of the prediction for this? What is your decision probability threshold and what is your predicted decision based on that?

The probability of the outcome is 1. The decision probability threshold is 0.5, meaning if a record gets a score above 0.5, the are classified as 1 (>50k income), and if below, classified as 0 (<50k income).

Question 10.1

What is the probability that your outcome prediction is accurate?

The probability of the outcome being correct is 78%.