**CA03 – Decision Tree Algorithms**

Q1.1: Why does it makes sense to discretize columns for this prediction problem?

The reason why it makes sense to discretize columns for this prediction problems is because in order to measure frequency for a large number of numeric values, unique records will be required for each value. For decision tree analysis, these records would all have to be accounted for, resulting in an extremely complicated and large analysis. Therefore, using this method would reduce the volume of the data and make it easier to work with it.

Q1.2: What might be the issues (if any) if we DID NOT discretize the columns?

If discretizing was not performed then there would be an excessive number of leaves. With each unique value having to be a leaf, it could lead to an enormous number of leaves for some columns. As a result of discretizing the columns, the number of leaves can be constrained by grouping them into bins in order to simplify the analysis and prevents overfitting.

Q8.1: How long was your total run time to train the model?

Running the model and manually entering the values into an Excel spreadsheet took approximately 45 minutes to an hour. I am still learning how to build models the fastest way. I still got some errors when building the model, which took me more time to solve.

Q8.2: Did you find the BEST TREE?

By manually running each of my hyperparameter combinations, I was able to find a tree with the best accuracy, precision, and f1 score. With these results, I would rate this tree as the best tree I can possibly obtain.

Q8.3: Draw the Graph of the BEST TREE Using GraphViz

Q8.4: What makes it the best tree?

This Tree had overall the best tree performance with the values given. It had an accuracy of 84.03% and overall higher performance than the rest.

Q10.1: 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 [0.71428571 0.28571429]. The decision probability threshold is of 0.5. Based on that, my predicted decision is that the predicted decision is [0], and the predicted income level will equal <=50k.

Q10.2: What is the probability that your outcome prediction is accurate?

The probability that the outcome prediction of 0, meaning income level is less than or equal to 50k, is accurate is .714 and the probability of the outcome being 1, income level being greater than 50k, is .286.