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

If the decision tree was to use continuous variables instead, there would be too many variables to take into account. Rather than that, it is better to discretize the result and narrow down the factors.

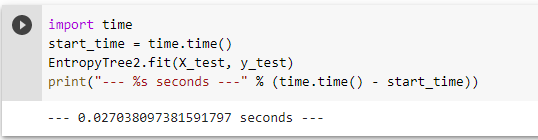
Q.1.2 What might be the issues (if any) if we DID NOT discretize the columns.

If the column was not discretized, there would be issues with outliers and skewed variables. Putting the continuous values into discrete bins helps normalize the distribution of the data.

Q.7.1 Decision Tree Hyper-parameter variation vs. performance

I tested around with the hyper parameters for both entropy and Gini and found out that the performance dips if your tree splits too many splits, leaves, and high max depths. The same result applies to parameters the little splits, leaves, and small max depths. I was able to get the best performing trees when these parameters were set in moderation, around 25 leaf, split, and depth.

Q.8.1 How long was your total run time to train the model?



Q.8.2 Did you find the BEST TREE?

My best tree is Entropy number 2

Q.8.3 Draw the Graph of the BEST TREE Using GraphViz

**Check Colab**

Q.8.4 What makes it the best tree?

Entropy 2 is the best tree because it has the highest Accuracy score

And AUC value is above 0.5 which means that true positive rate is higher than false positive rate.

Q.10.1 What is the probability that your prediction for this person is accurate?

The probability of my prediction ( 0 - income below 50k) being correct is 56.25%

