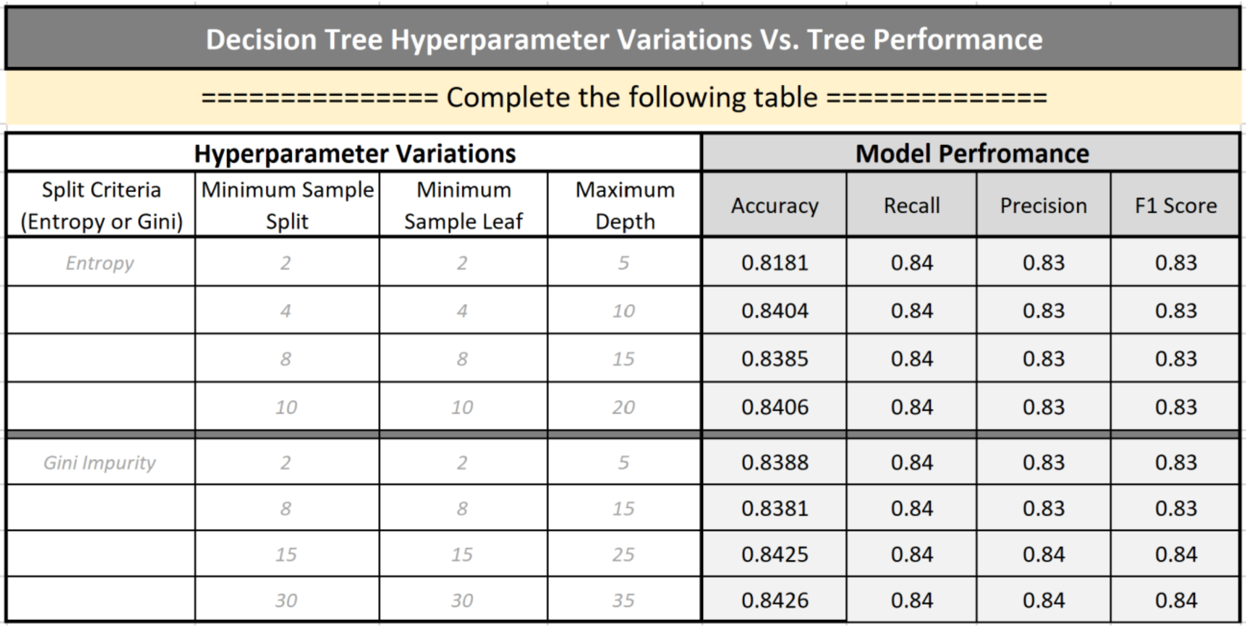
CA03 Assignment Report
Kassie - Xinyu Xie

Q.1.1 Why does it makes sense to discretize columns for this problem?
Q.1.2 What might be the issues (if any) if we DID NOT discretize the columns.

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



Decision Tree Hyperparameter Variations Vs. Tree Performance

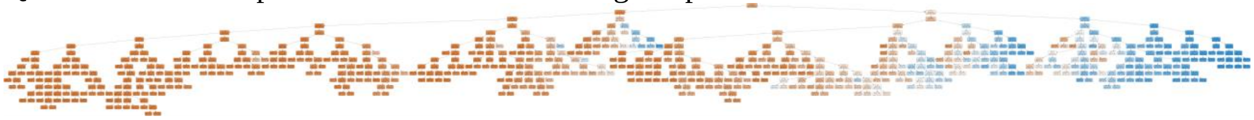
=============== Complete the following table ==============

Hyperparameter Variations				Model Perfromance			
Split Criteria (Entropy or Gini)	Minimum Sample Split	Minimum Sample Leaf	Maximum Depth	Accuracy	Recall	Precision	F1 Score
Entropy	2	2	5	0.8181	0.84	0.83	0.83
	4	4	10	0.8404	0.84	0.83	0.83
	8	8	15	0.8385	0.84	0.83	0.83
	10	10	20	0.8406	0.84	0.83	0.83
Gini Impurity	2	2	5	0.8388	0.84	0.83	0.83
	8	8	15	0.8381	0.84	0.83	0.83
	15	15	25	0.8425	0.84	0.84	0.84
	30	30	35	0.8426	0.84	0.84	0.84

Q.8.1 How long was your total run time to train the model?
I'm not sure about the total run time to train the model. If it's assuming the running time to train t he model in the background, I think it would be about 5 seconds in total. If it's assuming the total running time to train the model in the respect of person who impleted the training, I think it wou ld be around three to four hours to train for me.

Q.8.2 Did you find the BEST TREE?
As far as I'm considering, I think the tree that I got at last in the group of using Gini Impurity as criteria is the best tree so far.

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



Q.8.4 What makes it the best tree?

For this dataset, I am not very sure about between recall and precision, which one is more important in this case. So I decide to use all of the three. Having a higher Recall means there are less FALSE NEGATIVES, and having higher Precision means there are less FALSE

POSITIVES. Furthermore, harmonic mean of precision and recall. With all these being considered, I assume that my 4th Gini tree is the best tree so far.

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