Megan Rice
BSAN 6070

CA03 – Decision Tree Algorithm

Q.1.1 Why does it makes sense to discretize columns for this problem?

For this dataset many of the variables would be continuous if they were not discretized. Because we are using this dataset for a decision tree classifier it makes sense to make them discrete.

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

If we did not discretize the columns there could possibly be issues with outliers and highly skewed variables. Discretization helps handle outliers by placing these values into the lower or higher intervals together with the remaining inlier values of the distribution. Thus, these outlier observations no longer differ from the rest of the values at the tails of the distribution, as they are now all together in the same interval/bucket. In addition, by creating appropriate bins or intervals, discretization can help spread the values of a skewed variable across a set of bins with an equal number of observations.

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

Hyperparameter Variations				**Model Perfromance**			
Split Criteria (Entropy or Gini)	Minimum Sample Split	Minimum Sample Leaf	Maximum Depth	Accuracy	Recall	Precision	F1 Score
Entropy	*2*	*5*	*3*	0.83	0.83	0.82	0.83
	3	*2*	*2*	0.82	0.82	0.81	0.80
	4	*8*	*5*	0.84	0.84	0.83	0.83
	3	*6*	*8*	0.85	0.85	0.84	0.84
Gini Impurity	*2*	*5*	*3*	0.83	0.83	0.82	0.83
	3	*2*	*2*	0.82	0.82	0.81	0.8
	4	*8*	*5*	0.84	0.84	0.83	0.83
	3	*6*	*8*	0.85	0.85	0.84	0.84

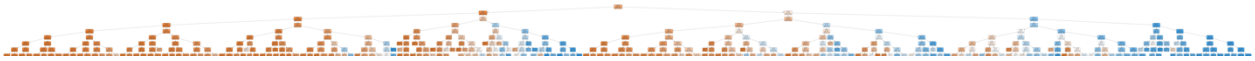
Q.8.1 How long was your total run time to train the model?

100 loops, best of 3: 17.9 ms per loop

Q.8.2 Did you find the BEST TREE?

I found one tree that worked better than others I tried but I do not know for sure that it is the best tree possible.

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



Q.8.4 What makes it the best tree?

This tree has the best scored across the board for accuracy, recall, precision, and f1 score. It also had the best AUC.

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

Based on the AUC value for my best model... the probability that this prediction is accurate is around 76%.