BSAN 6070 Machine Learning

Rongxing (Vincent) Chen

CA03 Decision Tree Classification

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

Dividing the data into groups could help identify the groups of persons since the same group of persons tend to have similar attributes in the dataset. Discretizing data also helps in grouping outliers. It is more convenient to draw general conclusions from the analysis when we group out the values.

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

Continuous data might not work well in some of the decision tree models.

Extreme values could affect the model created later.

Q7.1 Decision Tree Hyper-parameter variation vs. performance



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

I have run 16 times, 8 for entropy and 8 for gini.

Q.8.2 Did you find the BEST TREE?

Yes, based on the combination of hyperparameters that I have tried.

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



Q.8.4 What makes it the best tree?

It should have the highest F1 score and accuracy score compared to other trees.

Based on the hyperparameters tried, the tree with a criterion of gini, min\_samples\_split of 7, min\_samples\_leaf of 20, max\_depth of 8, could generate the best prediction results.

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

It is predicted that this person would have an annual income of over $50,000, with a probability of 0.78947368.