CA03 – Decision Tree Models

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

The data used in this problem is continuous; discretizing the columns to discrete data can significantly improve the classification performance of certain machine learning algorithms, including decision trees. Moreover, many machine learning algorithms, including Decision Trees and Naïve Bayes, exclusively use discrete data and are unable to process continuous data.

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

Not discretizing the data will not only slow the model down, but also negatively affect its performance. Continuous data tends to be more ambiguous and thus more difficult to interpret by machines, which can result in errors.

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

**Table

Description automatically generated**

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

The total runtime to train the model was 6.44 µs.

**Q.8.2 Did you find the BEST TREE?**

The decision tree with the best performance scores was variation 8. The code is as follows:

best\_dtree = DecisionTreeClassifier(criterion='gini', max\_depth=30, min\_samples\_leaf=15, min\_samples\_split=200)

best\_dtree.fit(X\_train, y\_train)

y\_pred = best\_dtree.predict(X\_test)

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

**Chart, scatter chart

Description automatically generated**

**Q.8.4 What makes it the best tree?**

Tree variation #8 was the best decision tree as its classification report returned the highest scores on average.

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

The probability that this prediction is accurate is about 84.5%, according to the returned accuracy score.