Vania Revelina

Professor Arin Brahma

CA03 – Questions

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

Since we are using Decision Tree Classifier and not Regressor, our model requires us to classify our attributes into different categories/bins. There are also too many integers/values for age and hours worked per week attributes. It doesn’t make sense to differentiate each unique integer values.

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

As we can see from our analysis, we needed to create dummy variables for each column to fit the model into our Decision Tree Classifier. The resulting amount of attributes including the dummies are the amount of distinct categories in the dataset (in this case 31 without the education\_num\_bin column). If we did not discretize the columns, let’s say for hours worked per week attribute, we would have 100 unique values (1 through 100). Our Decision Tree Classifier would try to create dummies for each of those unique values because they would interpret them as categories. As a result, we get these overly specific model where our decision trees break out into unique hours worker per week integers from 1 through 100. Having these would higher the chance of overfitting.

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

******

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

It took roughly 0.05 seconds to train the best decision tree model.

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

Yes, the best tree is the one with the following hyperparameters:

* Criterion: Gini Impurity
* Minimum Samples Split: 10
* Minimum Samples Leaf: 30
* Maximum Depth: 9

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

***A picture containing screenshot

Description automatically generated***

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

It has the highest F1-Score (0.64) which means it has the best weighted average of the precision and recall values. There are 2 sets of hyperparameters with the highest F1-Score, but I chose this particular one because it has a slightly better precision score.

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

The accuracy score for the best tree is 0.84, so the probability of accurately predicting this person using the best decision tree model is 84%.