CA03 Decision Tree Algorithm

Google Drive Link: https://colab.research.google.com/drive/1ADOom8c-ZqJYATrxias_8Scfa8-NRqy3

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

It makes sense to discretize columns for this problem because there are many features that are categorical per se (Occupation, Education, etc.), and thus, we need to handle all variables with the same type of data. In addition, if we think about the decision tree algorithm and we try to do it by hand, we will realize that if we tried using numerical variables, splitting decisions would be difficult to handle because there would be many splits for each number.

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

The main issues that we would encounter if we would not discretize columns are, as I mentioned before, handling categorical variables in decision trees would be much more convenient than numerical. In addition, discretization allows to handle outliers, placing the values in the corresponding intervals (together with inliers).

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

CA03 - Deicison Tree

Name:	Carla Mariana Fera

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*	*15*	*10*	0.8425	0.75	0.79	0.765
Entropy	*5*	*20*	*5*	0.8278	0.73	0.775	0.73
Entropy	*2*	*5*	*2*	0.8165	0.675	0.77	0.7
Entropy	*2*	*5*	*10*	0.8401	0.745	0.785	0.765
Gini Impurity	*10*	*15*	*10*	0.8423	0.75	0.79	0.765
Gini Impurity	*3*	*5*	*5*	0.8282	0.71	0.775	0.73
Gini Impurity	*2*	*5*	*2*	0.8165	0.675	0.77	0.7
Gini Impurity	*2*	*5*	*10*	0.8404	0.75	0.785	0.765

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

The total runtime to train the model is 1/100 of a second.

Q.8.2 Did you find the BEST TREE?

Yes, I find the best tree which is the first tree on the list of the table above or TREE 5, from the python notebook.

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

Refer to the code to see the tree complete tree.

Q.8.4 What makes it the best tree?

This tree is better than the other trees because all the scores are ranked higher than the rest of the trees. For instance, accuracy is 0.8425, which is the highest. Precision is 0.79, which is also the highest one, and so on.

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

The accuracy for this model is the same as the original model which is 0.8425.