**CA03 – Decision Tree Algorithm**

1. Data Source and Contents

* Data Source: <https://drive.google.com/file/d/13335ZpTuNi7bE25go_4nNpRXpVm90zxO/view?usp=sharing>
* Provide all packages that are needed for the codes.

1. Data Quality Analysis (DQA)

* Access the dataset.
* Clean the dataset by checking missing values and datatypes.

1. Exploratory Data Analysis (EDA)

* Display the graphs for dependent variables versus independent variables.
* Age vs Income segments.
* Capital Gain/Loss vs Income segments.
* Education vs Income segments.
* Hours per Week vs Income segments.
* Marriage Status and Relationship vs Income segments.
* Occupation vs Income segments.
* Race and Sex vs Income segments.

1. Build Decision Tree Classifier Models

* Slice and label the variables.
* Set up the conditions for the decision tree model.
* Run the model based on the given condition.

1. Visualize Your Decision Tree using GraphViz

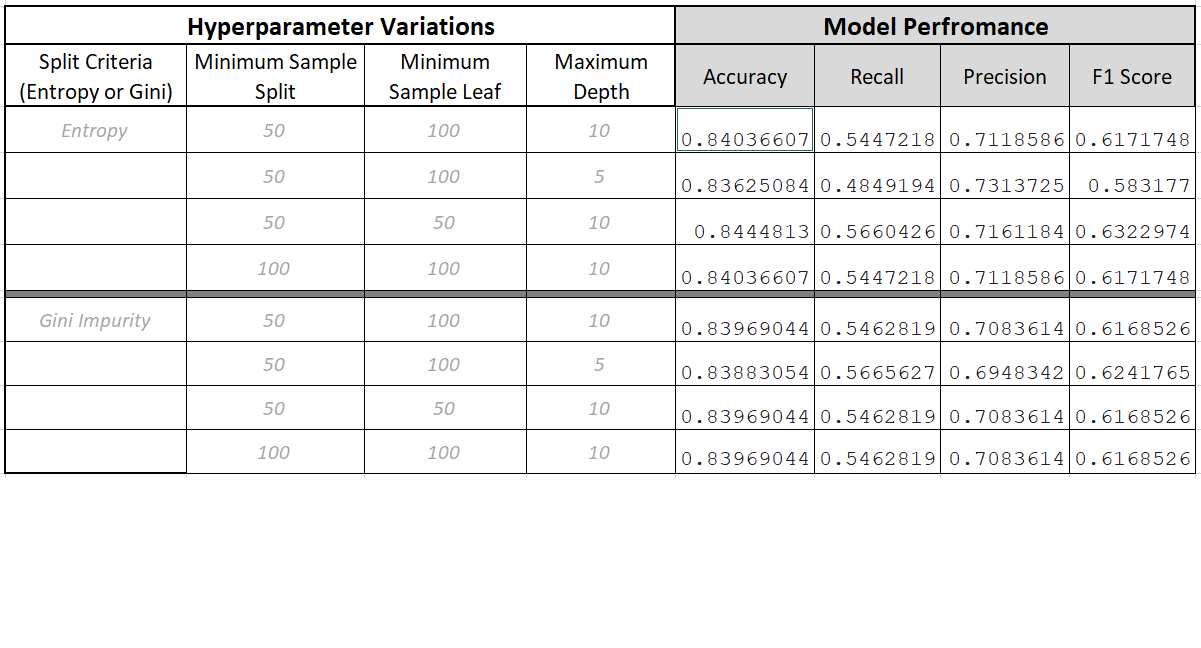
* Display the model.
* 

1. Evaluate Decision Tree Performance

* By using confusion matrix including TP, TN, FP, and FN.
* Also, by using accuracy score, precision, recall, f1 score, AUC value, and ROC curve.

1. Tune Decision Tree Performance

* Set up 8 different conditions to test out the best models.
* Hyperparameter Variations and results are recorded at the table below.



1. Conclusion

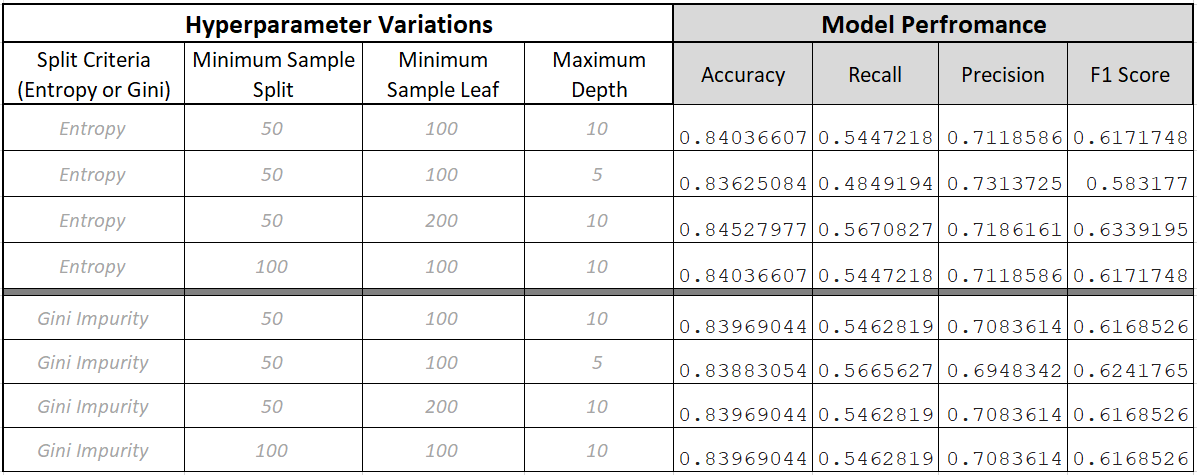
Q.1.1 Why does it makes sense to discretize columns for this problem?

- Because it will be easier to create the decision tree model

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

- There will be so many branches and split which causes the difficulty to predict.

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



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

- 7.87 ms

Q.8.2 Did you find the BEST TREE?

- Here is the condition for the best tree.

criterion= 'entropy', min\_samples\_split =50, min\_samples\_leaf = 50, max\_depth=10

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



Q.8.4 What makes it the best tree?

- The model returns highest accuracy score, recall, and f1 score, and the precision is also high.

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

- 84.5%

1. Automation of Performance Tuning

- Set up the codes to test the hyperparameters.

- Return the table of hyperparameters and evaluating values.

10. Deliverables

Using the given data to predict the income segments.

- Hours Worked per Week = 48

* Occupation Category = Mid – Low
* Marriage Status & Relationships = High
* Capital Gain = Yes
* ace-Sex Group = Mid
* Number of Years of Education = 12
* Education Category = High
* Work Class = Income
* Age = 58

The results show the person has income equal or higher than 50K.