Packages that were used

Sklearn

Numpy

Pandas

Graphviz

Counter

Purpose of this project:

Convert census data from continuous variables into discrete bins then analyze and create decision trees. Furthermore, test the accuracy of the predictions that were made with the sklearn packages.

Flag are the y values / labels for this assignment

0 means income below 50 k and 1 means income above 50k

Task 1:

I explained the why discretization is needed for the project in the word document CA03 Questions.

Task 2:

Perform data analysis and cleaning the dataframe

Make sure there are no null values, remove null values if they exist

Task 3:

Utilize counter, dictionary, list and converting them into a dataframe to format the graph for the

7 variables it wants us to show visually –

“age (5), Capital gain/loss (3), education (5), hours per week (5), Marraige Status n relationship (3), Occupation (5), Race and Sex (3)”

Task 4:

Create X train dataframe with the data labelled as train and test dataframe with data labelled as test.

Create the decision tree, then utilize fit to do the predictions.

Task 5:

Visualize the tree I created in task 4

Task 6:

Create confusion Matrix.

Find out performance metrics – accuracy, precision, recall, f1\_score, AUC value, and ROC plot.

Task 7:

Create 4 entropy and gini trees testing out different parameters to find out which of them generates the best tree.

Task 8:

Questions answered in questions word document

Task 9:

Automate the process of task 7. I utilized iterrows() on the dataframe csv with inputted hyperparameters and empty performance values. The function I created iterates through each row. It passes the values in the first 4 hyperparameters column into my decision tree which then predicts and generate the performance metrics. It checks if split criteria column is equal to entropy or gini before filling out the metrics.

Task 10:

Display unique values from each column with unique function.

Utilize the entropy 2 (best tree) to do the prediction on the data provided.

I categorized the labels manually then ran the prediction and prediction probability.

I found out that I have 56.25% of my prediction being correct