import pandas as pd

from sklearn.model\_selection import train\_test\_split

from sklearn import tree

from sklearn.metrics import accuracy\_score, confusion\_matrix

data = pd.read\_csv('customerdata.csv')

pd.options.display.max\_columns = None

print(data)

print(data.info())

data.drop(columns = ['User ID'], inplace=True)

# RECODE GENDER -- male 0 and female 1

gender\_dict = {'Male':0,'Female':1}

data.Gender =[gender\_dict[item] for item in data.Gender]

# \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

# \* CATEGORIZE THE SALARY VALUES

# \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

def categorizeSalary(sal):

cat = sal/10000

if cat < 4:

return 0

elif cat > 8:

return 2

else:

return 1

data['EstimatedSalary'] = data['EstimatedSalary'].apply(categorizeSalary)

# \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

# \* CATEGORIZE THE Generation VALUES

# \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

def categorizeGeneration(gen):

if gen > 59:

return 0

elif gen > 43:

return 1

elif gen > 26:

return 2

else:

return 3

data['Age'] = data['Age'].apply(categorizeGeneration)

# SEPARATE X and y

X = data.drop(columns=['Purchased'])

y = data.Purchased

# split 70-30

X\_train, X\_test, y\_train, y\_test = train\_test\_split(X,y,test\_size=.3)

# CREATE THE MODEL

dectree = tree.DecisionTreeClassifier()

#dectree = tree.DecisionTreeClassifier(criterion= 'entropy')

dectree.fit(X\_train, y\_train)

# HOW GOOD IS OUR MODEL ??

predictions = dectree.predict(X\_test)

print('\n\nAccuracy score:')

print(accuracy\_score(y\_test,predictions))

print('\n\nConfusion matrix:')

print(confusion\_matrix(y\_test,predictions))

import matplotlib.pyplot as plt

fig = plt.figure(figsize=(20,20))

tree.plot\_tree(dectree,feature\_names=X.columns.tolist())

plt.show()

# Predict yes will purchase or no probably not

# for a female millenial 75k salary

sal = categorizeSalary(75000)

xpredict = [[1,2,sal]]

X\_predicts = pd.DataFrame(xpredict,columns=X.columns)

print('\n\nPredicting data: ', X\_predicts)

yForX\_predicts = dectree.predict(X\_predicts)

print(yForX\_predicts)