Project 1: Build Decision Tree (DV-"Survived", IDV-"Age, Gender and Fare") and Prediction

Step 1: Load the Dataset

import pandas as pd

train\_dataset = pd.read\_csv("D:/AI\_ML\_Course/Day24/train.csv")

train\_dataset.columns

Out[25]:

Index(['PassengerId', 'Survived', 'Pclass', 'Name', 'Sex', 'Age', 'SibSp',

'Parch', 'Ticket', 'Fare', 'Cabin', 'Embarked'],

dtype='object')

train\_dataset["Age"].isnull().sum()

Out[33]: 0

Step 2: Build the Decision Tree with important features (Independent variable as Age, Sex and Fare) and (Dependent variable as Survived)

from sklearn import tree

from sklearn import preprocessing

label\_encoder = preprocessing.LabelEncoder()

#Encoding gender from categorical to continuous

encoded\_sex\_test = label\_encoder.fit\_transform(train\_dataset["Sex"])

encoded\_survived\_test = label\_encoder.fit\_transform(train\_dataset["Survived"])

test\_features = pd.DataFrame([train\_dataset["Age"],encoded\_sex\_test,train\_dataset["Fare"]]).T

tree\_model = tree.DecisionTreeClassifier(max\_depth = 8)

tree\_model.fit(X=test\_features,y=train\_dataset['Survived'])

with open("D:/AI\_ML\_Course/Day24/Dtree2.dot",'w') as f:

f = tree.export\_graphviz(tree\_model,feature\_names=["Age","Sex","Fare"],out\_file=f);



Step 3: To calculate the accuracy of the model

tree\_model.score(X=test\_features, y=train\_dataset['Survived'])

Out[52]: 0.8706411698537683

Inference: The model accuracy is 87.06%

Step 4: Based on the training Dataset predicted the test dataset output

test\_preds = tree\_model.predict(X=test\_features)

predicted\_output =pd.DataFrame({'PassengerId':train\_dataset['PassengerId'],'Survived':test\_preds})

predicted\_output.to\_csv("D:/AI\_ML\_Course/Day24/OutputSP.csv",index=False)

