In [1]:

```
import pandas as pd
import os
import joblib as jb
import sklearn
import pydotplus
```

In [2]:

```
from sklearn.preprocessing import LabelEncoder
```

In [35]:

```
data=pd.read_excel('Combined.xlsx')
```

In [36]:

```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 48842 entries, 0 to 48841
Data columns (total 15 columns):
 #   Column         Non-Null Count  Dtype 
---  ------         --------------  ----- 
 0   Age            48842 non-null  int64 
 1   Workclass      48842 non-null  object
 2   Fnlwgt         48842 non-null  int64 
 3   Education      48842 non-null  object
 4   EducationNum   48842 non-null  int64 
 5   MaritalStatus  48842 non-null  object
 6   Occupation     48842 non-null  object
 7   Relationship   48842 non-null  object
 8   Race           48842 non-null  object
 9   Sex            48842 non-null  object
 10  CapitalGain    48842 non-null  int64 
 11  CapitalLoss    48842 non-null  int64 
 12  HoursPerWeek   48842 non-null  int64 
 13  NativeCountry  48842 non-null  object
 14  Class          48842 non-null  object
dtypes: int64(6), object(9)
memory usage: 5.6+ MB
```

In [44]:

```
data['Age'].value_counts()
```

Out[44]:

```
36    1348
35    1337
33    1335
23    1329
31    1325
      ...
88       6
85       5
87       3
89       2
86       1
Name: Age, Length: 74, dtype: int64
```

In [45]:

```
data['Workclass'].value_counts()
```

Out[45]:

```
 Private             33906
 Self-emp-not-inc     3862
 Local-gov            3136
 ?                    2799
 State-gov            1981
 Self-emp-inc         1695
 Federal-gov          1432
 Without-pay            21
 Never-worked           10
Name: Workclass, dtype: int64
```

In [46]:

```
data['Workclass'] = data['Workclass'].str.strip().replace('?', 'Private')
```

In [47]:

```
data['Workclass'].value_counts()
```

Out[47]:

```
Private             36705
Self-emp-not-inc     3862
Local-gov            3136
State-gov            1981
Self-emp-inc         1695
Federal-gov          1432
Without-pay            21
Never-worked           10
Name: Workclass, dtype: int64
```

In [48]:

```
data['Fnlwgt'].value_counts()
```

Out[48]:

```
203488    21
120277    19
190290    19
125892    18
126569    18
          ..
286983     1
185942     1
234220     1
214706     1
350977     1
Name: Fnlwgt, Length: 28523, dtype: int64
```

In [49]:

```
data['MaritalStatus'].value_counts()
```

Out[49]:

```
 Married-civ-spouse       22379
 Never-married            16117
 Divorced                  6633
 Separated                 1530
 Widowed                   1518
 Married-spouse-absent      628
 Married-AF-spouse           37
Name: MaritalStatus, dtype: int64
```

In [11]:

```
data['Occupation'].value_counts()
```

Out[11]:

```
 Prof-specialty       6172
 Craft-repair         6112
 Exec-managerial      6086
 Adm-clerical         5611
 Sales                5504
 Other-service        4923
 Machine-op-inspct    3022
 ?                    2809
 Transport-moving     2355
 Handlers-cleaners    2072
 Farming-fishing      1490
 Tech-support         1446
 Protective-serv       983
 Priv-house-serv       242
 Armed-Forces           15
Name: Occupation, dtype: int64
```

In [50]:

```
data['Occupation'] = data['Occupation'].str.strip().replace('?', 'Prof-specialty')
data['Occupation'].value_counts()
```

Out[50]:

```
Prof-specialty       8981
Craft-repair         6112
Exec-managerial      6086
Adm-clerical         5611
Sales                5504
Other-service        4923
Machine-op-inspct    3022
Transport-moving     2355
Handlers-cleaners    2072
Farming-fishing      1490
Tech-support         1446
Protective-serv       983
Priv-house-serv       242
Armed-Forces           15
Name: Occupation, dtype: int64
```

In [51]:

```
data['Relationship'].value_counts()
```

Out[51]:

```
 Husband           19716
 Not-in-family     12583
 Own-child          7581
 Unmarried          5125
 Wife               2331
 Other-relative     1506
Name: Relationship, dtype: int64
```

In [52]:

```
data['Race'].value_counts()
```

Out[52]:

```
 White                 41762
 Black                  4685
 Asian-Pac-Islander     1519
 Amer-Indian-Eskimo      470
 Other                   406
Name: Race, dtype: int64
```

In [53]:

```
data['Sex'].value_counts()
```

Out[53]:

```
 Male      32650
 Female    16192
Name: Sex, dtype: int64
```

In [54]:

```
data['CapitalGain'].value_counts()
```

Out[54]:

```
0        44807
15024      513
7688       410
7298       364
99999      244
         ...
22040        1
2387         1
1639         1
1111         1
6612         1
Name: CapitalGain, Length: 123, dtype: int64
```

In [55]:

```
data['CapitalLoss'].value_counts()
```

Out[55]:

```
0       46560
1902      304
1977      253
1887      233
2415       72
        ...
1539        1
1870        1
1911        1
2465        1
1421        1
Name: CapitalLoss, Length: 99, dtype: int64
```

In [43]:

```
data['NativeCountry'].value_counts()
```

Out[43]:

```
United-States                 44689
Mexico                          951
Philippines                     295
Germany                         206
Puerto-Rico                     184
Canada                          182
El-Salvador                     155
India                           151
Cuba                            138
England                         127
China                           122
South                           115
Jamaica                         106
Italy                           105
Dominican-Republic              103
Japan                            92
Guatemala                        88
Poland                           87
Vietnam                          86
Columbia                         85
Haiti                            75
Portugal                         67
Taiwan                           65
Iran                             59
Greece                           49
Nicaragua                        49
Peru                             46
Ecuador                          45
France                           38
Ireland                          37
Hong                             30
Thailand                         30
Cambodia                         28
Trinadad&Tobago                  27
Laos                             23
Yugoslavia                       23
Outlying-US(Guam-USVI-etc)       23
Scotland                         21
Honduras                         20
Hungary                          19
Holand-Netherlands                1
Name: NativeCountry, dtype: int64
```

In [57]:

```
data['NativeCountry'] = data['NativeCountry'].str.strip().replace('?', 'United-States')
data['NativeCountry'].value_counts()
```

Out[57]:

```
United-States                 44689
Mexico                          951
Philippines                     295
Germany                         206
Puerto-Rico                     184
Canada                          182
El-Salvador                     155
India                           151
Cuba                            138
England                         127
China                           122
South                           115
Jamaica                         106
Italy                           105
Dominican-Republic              103
Japan                            92
Guatemala                        88
Poland                           87
Vietnam                          86
Columbia                         85
Haiti                            75
Portugal                         67
Taiwan                           65
Iran                             59
Greece                           49
Nicaragua                        49
Peru                             46
Ecuador                          45
France                           38
Ireland                          37
Hong                             30
Thailand                         30
Cambodia                         28
Trinadad&Tobago                  27
Laos                             23
Yugoslavia                       23
Outlying-US(Guam-USVI-etc)       23
Scotland                         21
Honduras                         20
Hungary                          19
Holand-Netherlands                1
Name: NativeCountry, dtype: int64
```

In [58]:

```
data['Class'].value_counts()
```

Out[58]:

```
 <=50K    37155
 >50K     11687
Name: Class, dtype: int64
```

In [59]:

```
data["Age-Category"]=pd.cut(data.Age, bins=[0,20,40,60,120],labels=["Child","Young","Adu
data
```

Out[59]:

	Age	Workclass	Fnlwgt	Education	EducationNum	MaritalStatus	Occupation	Relati
0	39	State-gov	77516	Bachelors	13	Never-married	Adm-clerical	Not-in
1	50	Self-emp-not-inc	83311	Bachelors	13	Married-civ-spouse	Exec-managerial	Hu
2	38	Private	215646	HS-grad	9	Divorced	Handlers-cleaners	Not-in
3	53	Private	234721	11th	7	Married-civ-spouse	Handlers-cleaners	Hu
4	28	Private	338409	Bachelors	13	Married-civ-spouse	Prof-specialty	
...	...	...	...	...	...	...	...	
48837	39	Private	215419	Bachelors	13	Divorced	Prof-specialty	Not-in
48838	64	Private	321403	HS-grad	9	Widowed	Prof-specialty	
48839	38	Private	374983	Bachelors	13	Married-civ-spouse	Prof-specialty	Hu
48840	44	Private	83891	Bachelors	13	Divorced	Adm-clerical	Ow
48841	35	Self-emp-inc	182148	Bachelors	13	Married-civ-spouse	Exec-managerial	Hu

48842 rows × 16 columns

In [60]:

```
del data['Age']
```

In [61]:

```
data = data.rename(columns={'Age-Category': 'AgeCategory'})
data
```

Out[61]:

wgt	Education	EducationNum	MaritalStatus	Occupation	Relationship	Race	Sex	Capit
516	Bachelors	13	Never-married	Adm-clerical	Not-in-family	White	Male	
311	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband	White	Male	
646	HS-grad	9	Divorced	Handlers-cleaners	Not-in-family	White	Male	
721	11th	7	Married-civ-spouse	Handlers-cleaners	Husband	Black	Male	
409	Bachelors	13	Married-civ-spouse	Prof-specialty	Wife	Black	Female	
...	...	...	...	...	...	...	...	
419	Bachelors	13	Divorced	Prof-specialty	Not-in-family	White	Female	
403	HS-grad	9	Widowed	Prof-specialty	Other-relative	Black	Male	
983	Bachelors	13	Married-civ-spouse	Prof-specialty	Husband	White	Male	
891	Bachelors	13	Divorced	Adm-clerical	Own-child	Asian-Pac-Islander	Male	
148	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband	White	Male	

ıns

In [62]:

```
data["FnlwgtCategory"]=pd.cut(data.Fnlwgt, bins=[-1,100000,500000,1000000,1500000],label
del data['Fnlwgt']
data
```

3	Private	11th	7	Married-civ-spouse	Handlers-cleaners	Husband	Black	Male
4	Private	Bachelors	13	Married-civ-spouse	Prof-specialty	Wife	Black	Female
...	...	...	...	...	...	...	...	...
48837	Private	Bachelors	13	Divorced	Prof-specialty	Not-in-family	White	Female
48838	Private	HS-grad	9	Widowed	Prof-specialty	Other-relative	Black	Male
48839	Private	Bachelors	13	Married-civ-spouse	Prof-specialty	Husband	White	Male
48840	Private	Bachelors	13	Divorced	Adm-clerical	Own-child	Asian-Pac-Islander	Male
48841	Self-emp-	Bachelors	13	Married-civ-	Exec-	Husband	White	Male

In [63]:

```
data["CapitalGainCategory"]=pd.cut(data.CapitalGain, bins=[-1,1000,50000,100000,],labels
data["CapitalLossCategory"]=pd.cut(data.CapitalLoss, bins=[-1,1000,5000,10000,],labels=[
data["HoursPerWeekCategory"]=pd.cut(data.HoursPerWeek, bins=[-1,40,70,100,],labels=["<40
del data['CapitalGain']
del data['CapitalLoss']
del data['HoursPerWeek']
data
```

Out[63]:

	Workclass	Education	EducationNum	MaritalStatus	Occupation	Relationship	Rac
0	State-gov	Bachelors	13	Never-married	Adm-clerical	Not-in-family	Whit
1	Self-emp-not-inc	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband	Whit
2	Private	HS-grad	9	Divorced	Handlers-cleaners	Not-in-family	Whit
3	Private	11th	7	Married-civ-spouse	Handlers-cleaners	Husband	Blac
4	Private	Bachelors	13	Married-civ-spouse	Prof-specialty	Wife	Blac
...	...	...	...	...	...	...	...
48837	Private	Bachelors	13	Divorced	Prof-specialty	Not-in-family	Whit
48838	Private	HS-grad	9	Widowed	Prof-specialty	Other-relative	Blac
48839	Private	Bachelors	13	Married-civ-spouse	Prof-specialty	Husband	Whit
48840	Private	Bachelors	13	Divorced	Adm-clerical	Own-child	Asian Pac Islande
48841	Self-emp-inc	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband	Whit

48842 rows × 15 columns

In [64]:

```
data['CapitalGainCategory'].value_counts()
```

Out[64]:

```
<1000           44888
1000-50000       3710
50000-100000      244
Name: CapitalGainCategory, dtype: int64
```

In [65]:

```
from sklearn.preprocessing import LabelEncoder
enc=LabelEncoder()

data_num=pd.DataFrame()
data_num['AgeCategory']= enc.fit_transform(data['AgeCategory'])
data_num['Workclass']= enc.fit_transform(data['Workclass'])
data_num['Education']= enc.fit_transform(data['Education'])
data_num['EducationNum']= enc.fit_transform(data['EducationNum'])
data_num['MaritalStatus']= enc.fit_transform(data['MaritalStatus'])
data_num['Occupation']= enc.fit_transform(data['Occupation'])
data_num['Relationship']= enc.fit_transform(data['Relationship'])
data_num['Sex']= enc.fit_transform(data['Sex'])
data_num['NativeCountry']= enc.fit_transform(data['NativeCountry'])
data_num['Race']= enc.fit_transform(data['Race'])
data_num['FnlwgtCategory']= enc.fit_transform(data['FnlwgtCategory'])
data_num['CapitalGainCategory']= enc.fit_transform(data['CapitalGainCategory'])
data_num['CapitalLossCategory']= enc.fit_transform(data['CapitalLossCategory'])
data_num['HoursPerWeekCategory']= enc.fit_transform(data['HoursPerWeekCategory'])
data_num['Class']= enc.fit_transform(data['Class'])
```

In [66]:

```
data_num
```

Out[66]:

	AgeCategory	Workclass	Education	EducationNum	MaritalStatus	Occupation	Relati
0	3	6	9	12	4	0	
1	0	5	9	12	2	3	
2	3	3	11	8	0	5	
3	0	3	1	6	2	5	
4	3	3	9	12	2	9	
...	...	...	...	...	...	...	
48837	3	3	9	12	0	9	
48838	2	3	11	8	6	9	
48839	3	3	9	12	2	9	
48840	0	3	9	12	0	0	
48841	3	4	9	12	2	3	

48842 rows × 15 columns

In [68]:

```
data_num.to_excel(r"C:\Users\SHASHANK GAUTAM\Desktop\ML_ASSIGNMENT\Decision_Tree1\Combin
```

In [67]:

```
data_num['CapitalGainCategory'].value_counts()
```

Out[67]:

```
2    44888
0     3710
1      244
Name: CapitalGainCategory, dtype: int64
```

In [69]:

```
data_num['CapitalLossCategory'].value_counts()
```

Out[69]:

```
1    46605
0     2237
Name: CapitalLossCategory, dtype: int64
```

In [73]:

```
data_num['NativeCountry'].value_counts()
```

Out[73]:

```
38    44689
25      951
29      295
10      206
32      184
1       182
7       155
18      151
4       138
8       127
2       122
34      115
22      106
21      105
5       103
23       92
12       88
30       87
39       86
3        85
13       75
31       67
35       65
19       59
11       49
26       49
28       46
6        45
9        38
20       37
16       30
36       30
0        28
37       27
24       23
40       23
27       23
33       21
15       20
17       19
14        1
Name: NativeCountry, dtype: int64
```

In [ ]:

```

```