```
import pandas as pd
from google.colab import drive
import matplotlib.pyplot as plt
import numpy as np
from sklearn.model_selection import train_test_split
import zipfile
from sklearn.tree import DecisionTreeClassifier
from sklearn.metrics import confusion_matrix,classification_report,plot_confusion_matrix, accuracy_score
from sklearn import tree
```

```
drive.mount('/content/gdrive/', force_remount = True)
```

```
Mounted at /content/gdrive/
```

```
zf = zipfile.ZipFile("/content/gdrive/MyDrive/stroke/us_perm_visas.csv.zip")
df = pd.read_csv(zf.open('us_perm_visas.csv'))
```

```
Exception ignored in: <function ZipFile.__del__ at 0x7f4999ede700>
Traceback (most recent call last):
  File "/usr/lib/python3.8/zipfile.py", line 1821, in __del__
  File "/usr/lib/python3.8/zipfile.py", line 1843, in close
  File "/usr/lib/python3.8/zipfile.py", line 1953, in _fpclose
OSError: [Errno 107] Transport endpoint is not connected
/usr/local/lib/python3.8/dist-packages/IPython/core/interactiveshell.py:3326: DtypeWarning: Columns (0,1,2,3,4,5,6,7,10,11,16,17,20,21,22,25,26,27,28,29,30,31,32,33,34
  exec(code_obj, self.user_global_ns, self.user_ns)
```

```
df
```

	add_these_pw_job_title_9089	agent_city	agent_firm_name	agent_state	application_type	case_no	case_number	case_received_date	case_status	class_of_admission
0	NaN	NaN	NaN	NaN	PERM	A-07323-97014	NaN	NaN	Certified	J-1
1	NaN	NaN	NaN	NaN	PERM	A-07332-99439	NaN	NaN	Denied	B-2
2	NaN	NaN	NaN	NaN	PERM	A-07333-99643	NaN	NaN	Certified	H-1B

```
df1 = df[['class_of_admission', 'us_economic_sector', 'wage_offer_from_9089', 'case_status' ]]
```

```
df2 = df1.dropna()
df2
```

	class_of_admission	us_economic_sector	wage_offer_from_9089	case_status
0	J-1	IT	75629.0	Certified
1	B-2	Other Economic Sector	37024.0	Denied
2	H-1B	Aerospace	47923.0	Certified
3	B-2	Other Economic Sector	10.97	Certified
4	L-1	Advanced Mfg	100000.0	Certified
...	...	...	...	...
20571	H-2B	Other Economic Sector	23.73	Certified
20572	EWI	Other Economic Sector	26.59	Withdrawn
20573	E-2	Aerospace	45.0	Withdrawn
20574	Not in USA	Agribusiness	8.1	Denied
20575	B-2	Transportation	39894.0	Certified-Expired

18741 rows × 4 columns

```
X = pd.get_dummies(df2.drop('case_status',axis=1),drop_first=True)
Y = df2['case_status']
```

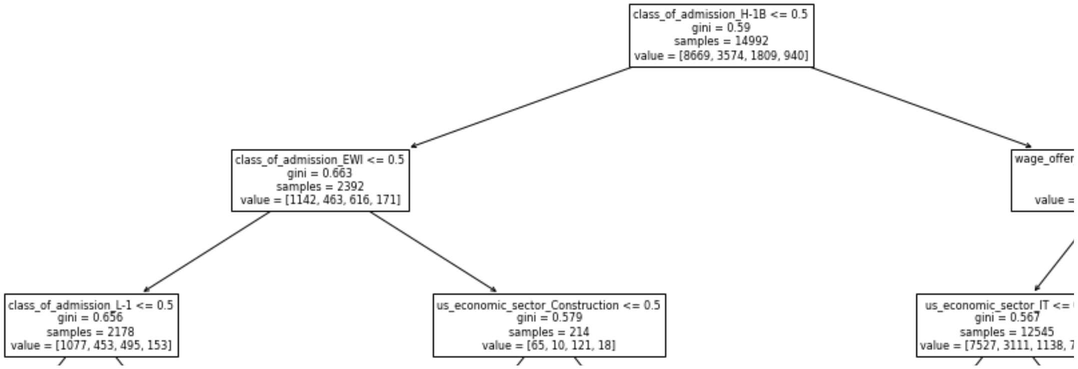
```
X_train, X_test, y_train, y_test = train_test_split(X, Y, test_size=0.20, random_state=100)
```

```
CLF = tree.DecisionTreeClassifier(max_depth=3)
CLF.fit(X_train,y_train)
```

```
DecisionTreeClassifier(max_depth=3)
```

```
plt.figure(figsize = (20,8))
tree.plot_tree(CLF,feature_names = X.columns )
```

```
[Text(0.5769230769230769, 0.875, 'class_of_admission_H-1B <= 0.5\ngini = 0.59\nsamples = 14992\nvalue =
[8669, 3574, 1809, 940]'),
 Text(0.3076923076923077, 0.625, 'class_of_admission_EWI <= 0.5\ngini = 0.663\nsamples = 2392\nvalue =
[1142, 463, 616, 171]'),
 Text(0.15384615384615385, 0.375, 'class_of_admission_L-1 <= 0.5\ngini = 0.656\nsamples = 2178\nvalue =
[1077, 453, 495, 153]'),
 Text(0.07692307692307693, 0.125, 'gini = 0.662\nsamples = 1739\nvalue = [834, 322, 455, 128]'),
 Text(0.23076923076923078, 0.125, 'gini = 0.593\nsamples = 439\nvalue = [243, 131, 40, 25]'),
 Text(0.46153846153846156, 0.375, 'us_economic_sector_Construction <= 0.5\ngini = 0.579\nsamples =
214\nvalue = [65, 10, 121, 18]'),
 Text(0.38461538461538464, 0.125, 'gini = 0.586\nsamples = 164\nvalue = [61, 6, 85, 12]'),
 Text(0.5384615384615384, 0.125, 'gini = 0.454\nsamples = 50\nvalue = [4, 4, 36, 6]'),
 Text(0.8461538461538461, 0.625, 'wage_offer_from_9089_44799.0 <= 0.5\ngini = 0.569\nsamples = 12600\nvalue
= [7527, 3111, 1193, 769]'),
 Text(0.7692307692307693, 0.375, 'us_economic_sector_IT <= 0.5\ngini = 0.567\nsamples = 12545\nvalue =
[7527, 3111, 1138, 769]'),
 Text(0.6923076923076923, 0.125, 'gini = 0.532\nsamples = 7666\nvalue = [4867, 1821, 602, 376]'),
 Text(0.8461538461538461, 0.125, 'gini = 0.614\nsamples = 4879\nvalue = [2660, 1290, 536, 393]'),
 Text(0.9230769230769231, 0.375, 'gini = 0.0\nsamples = 55\nvalue = [0, 0, 55, 0]')]
```


class_of_admission_H-1B <= 0.5
gini = 0.59
samples = 14992
value = [8669, 3574, 1809, 940]
class_of_admission_EWI <= 0.5
gini = 0.663
samples = 2392
value = [1142, 463, 616, 171]
wage_offer
value =
class_of_admission_L-1 <= 0.5
gini = 0.656
samples = 2178
value = [1077, 453, 495, 153]
us_economic_sector_Construction <= 0.5
gini = 0.579
samples = 214
value = [65, 10, 121, 18]
us_economic_sector_IT <=
gini = 0.567
samples = 12545
value = [7527, 3111, 1138, 7

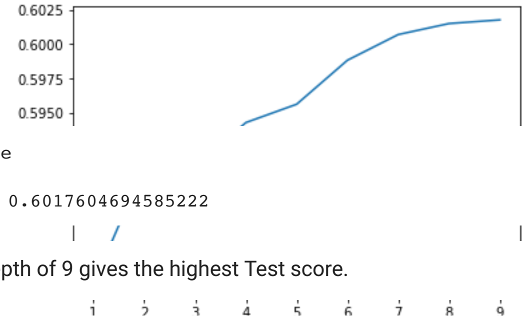
The largest leaf has a sample size of 7666, which is about 41% of our data. The gini for the largest leaf is closer to 1 which implies that there is missclassification in our predictions.

```
depth = []

for i in range(1,10):
    tree = DecisionTreeClassifier(max_depth=i)
    tree.fit(X_train,y_train)
    score = tree.score(X_test, y_test)
    depth.append(score)


plt.plot(range(1,10), depth)
```

```
[<matplotlib.lines.Line2D at 0x7f48f3676c40>]
```



```
score
```

```
0.6017604694585222
```

A depth of 9 gives the highest Test score.