Name : Prathamesh Rajbhoj

Roll no : 53

Batch : A2

Practical 04

Aim: To implement decision tree classifier on play tennis dataset and display generated tree.

```
from google.colab import drive
drive.mount('/content/gdrive')
```

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

```
import pandas as pd
import numpy as np
from math import log2
```

```
df = pd.read_csv('/content/gdrive/MyDrive/play_tennis.csv')
df
```

	day	outlook	temp	humidity	wind	play
0	D1	Sunny	Hot	High	Weak	No
1	D2	Sunny	Hot	High	Strong	No
2	D3	Overcast	Hot	High	Weak	Yes
3	D4	Rain	Mild	High	Weak	Yes
4	D5	Rain	Cool	Normal	Weak	Yes
5	D6	Rain	Cool	Normal	Strong	No
6	D7	Overcast	Cool	Normal	Strong	Yes
7	D8	Sunny	Mild	High	Weak	No
8	D9	Sunny	Cool	Normal	Weak	Yes
9	D10	Rain	Mild	Normal	Weak	Yes
10	D11	Sunny	Mild	Normal	Strong	Yes
11	D12	Overcast	Mild	High	Strong	Yes
12	D13	Overcast	Hot	Normal	Weak	Yes
13	D14	Rain	Mild	High	Strong	No

```
y=0
n=0
j = len(df)
for i in df['play']:
  if i == 'Yes':
    y+=1
  else:
    n+=1
e_total = - (y/j)*log2(y/j) - (n/j)*log2(n/j)
print(round(e_total,2))
```

```
0.94
```

```
ans = ''
max = -999
for col in df.columns:
  if col != 'play' and col != 'day':
    sum = 0
    for attVal in df[col].unique():
      sy = 0
      sn = 0
      for i in range(len(df)):
        if df[col][i] == attVal:
          if df['play'][i] == 'Yes':
            sy+=1
          else:
            sn+=1
      t = sy + sn
      if sy !=0 and sn != 0:
        en = - (sy/t)*log2(sy/t) - (sn/t)*log2(sn/t)
        sum += ((t/len(df)) * en)

    gain = e_total - round(sum,2)
    print(f"Information Gain for {col} = {gain}")
    if gain > max:
      ans = col
      max = gain
print("\n"+ans +'\tis root node')
```

```
Information Gain for outlook = 0.2502859586706312
Information Gain for temp = 0.030285958670631108
Information Gain for humidity = 0.1502859586706311
Information Gain for wind = 0.050285958670631126

outlook is root node
```