Họ và tên: Mai Thuý Ngọc

MSSV: 1755248020100038

BÀI TẬP BUỔI THỰC HÀNH 2

id	outlook	temperature	humidity	wind	play
1	sunny	hot	high	weak	no
2	sunny	hot	high	strong	no
3	overcast	hot	high	weak	yes
4	rainy	mild	high	weak	yes
5	rainy	cool	normal	weak	yes
6	rainy	cool	normal	strong	no
7	overcast	cool	normal	strong	yes
8	sunny	mild	high	weak	no
9	sunny	cool	normal	weak	yes
10	rainy	mild	normal	weak	yes
11	sunny	mild	normal	strong	yes
12	overcast	mild	high	strong	yes
13	overcast	hot	normal	weak	yes
14	rainy	mild	high	strong	no

1. Trả lời:

- Bộ dữ liệu có **14** mẫu dữ liệu.
- Có **4** thuộc tính: outlook, temperature, humidity, wind.
- Có **7** mẫu dữ liệu có thuộc tính Humidity: “high”
- Có **6** mẫu dữ liệu có thuộc tính Wind: “strong”
- Có **2** nhãn lớp: yes, no

2. Hãy xây dựng mô hình phân lớp dựa trên giải thuật Cây quyết định theo 2 cách:

❖ **Cách 1: Tính toán bằng tay**

Theo bảng dữ liệu ta có:

- $|D| = 14$
- $C = \{yes, no\}$
- $|C_{yes}| = 9$, $|C_{no}| = 5$

Tính Entropy:

$$\text{Info}(D) = -\frac{9}{14} \times \log_2 \frac{9}{14} - \frac{5}{14} \times \log_2 \frac{5}{14} = 0.940\text{bits}$$

Xét thuộc tính outlook:

sunny	yes	2
	no	3
overcast	yes	4
	no	0
rainy	yes	3
	no	2

- $Info_{outlook}(D) = \frac{5}{14} \times (-\frac{2}{5} \times \log_2 \frac{2}{5} - \frac{3}{5} \times \log_2 \frac{3}{5}) + \frac{4}{14} \times (-\frac{4}{4} \times \log_2 \frac{4}{4}) + \frac{5}{14} \times (-\frac{3}{5} \times \log_2 \frac{3}{5} - \frac{2}{5} \times \log_2 \frac{2}{5}) = 0.694bits$

➔ Gain (outlook) = 0.940 bits – 0.694bits = 0.246bits

Xét thuộc tính temperature:

hot	yes	2
	no	2
mid	yes	4
	no	2
cool	yes	3
	no	1

- $Info_{temperature}(D) = \frac{4}{14} \times (-\frac{2}{4} \times \log_2 \frac{2}{4} - \frac{2}{4} \times \log_2 \frac{2}{4}) + \frac{6}{14} \times (-\frac{4}{6} \times \log_2 \frac{4}{6} - \frac{2}{6} \times \log_2 \frac{2}{6}) + \frac{4}{14} \times (-\frac{3}{4} \times \log_2 \frac{3}{4} - \frac{1}{4} \times \log_2 \frac{1}{4}) = 0.911bits$

➔ Gain (temperature) = $0.940bits - 0.911bits =$ 0.029bits

Xét thuộc tính humidity:

high	yes	3
	no	4
normal	yes	6
	no	1

- $\text{Info}_{\text{humidity}}(D) = \frac{7}{14} \times (-\frac{3}{7} \times \log_2 \frac{3}{7} - \frac{4}{7} \times \log_2 \frac{4}{7}) + \frac{7}{14} \times (-\frac{6}{7} \times \log_2 \frac{6}{7} - \frac{1}{7} \times \log_2 \frac{1}{7}) = 0.788\text{bits}$

➔ Gain (humidity) = 0.940bits – 0.788bits = 0.152bits

Xét thuộc tính wind:

weak	yes	6
	no	2
strong	yes	3
	no	3

- $\text{Info}_{\text{wind}}(D) = \frac{8}{14} \times (-\frac{6}{8} \times \log_2 \frac{6}{8} - \frac{2}{8} \times \log_2 \frac{2}{8}) + \frac{6}{14} \times (-\frac{3}{6} \times \log_2 \frac{3}{6} - \frac{3}{6} \times \log_2 \frac{3}{6}) = 0.892\text{bits}$

➔ Gain (wind) = 0.940bits – 0.892bits = 0.048bits

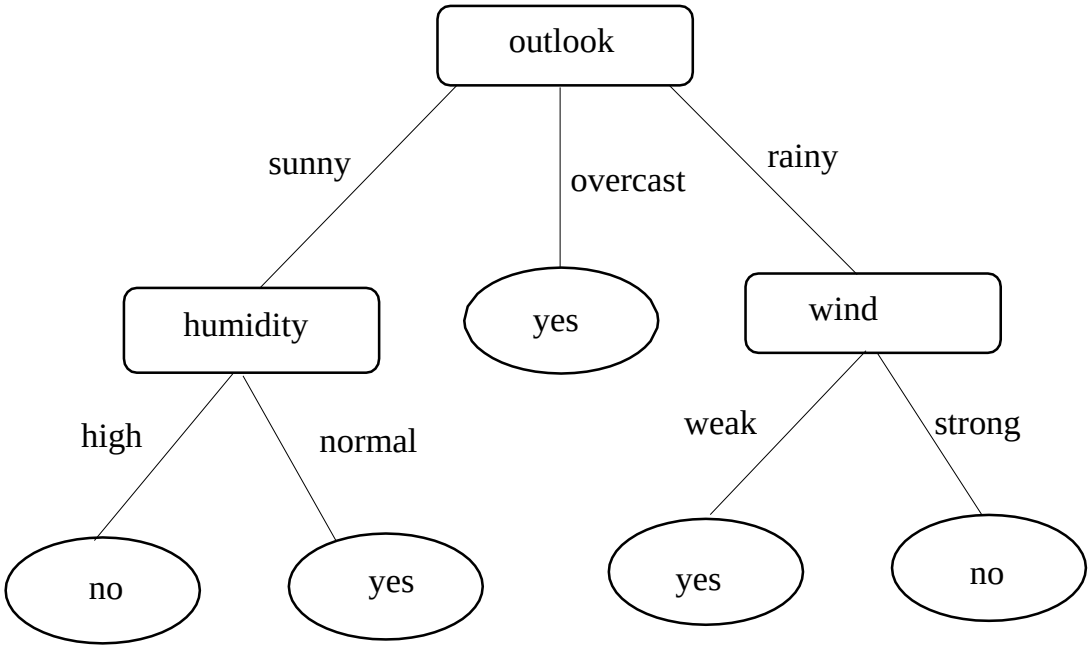
❖ **Sử dụng thư viện scikit-learn**

```
import numpy as np
import pandas as pd
from sklearn import tree
from sklearn.preprocessing import LabelEncoder

df = pd.read_csv('weather.csv')

y = df.iloc[:, -1]

columns = ['outlook', 'temperature', 'humidity', 'wind', 'play']
for a in columns:
    label = LabelEncoder()
    df[a] = label.fit_transform(df[a])

X = df.iloc[:, :-1]
```

```
clf = tree.DecisionTreeClassifier()
clf = clf.fit(X, y)
y_pred = clf.predict(X)
print(y_pred)
```

Vì thuộc tính outlook có độ lợi thông tin lớn nhất (0.247bits) nên outlook là thuộc tính được chọn để rẽ nhánh.

Ta có cây quyết định:



3. Đưa ra kết quả của giải thuật dự đoán với các mẫu dữ liệu có thuộc tính như sau:

id	outlook	temperature	humidity	wind	Play (tự tính)	Play (máy)
1	rainy	cool	high	weak	yes	yes
2	rainy	mild	normal	strong	no	no
3	overcast	cool	normal	strong	yes	yes
4	sunny	hot	high	weak	no	no
5	overcast	hot	normal	strong	yes	yes