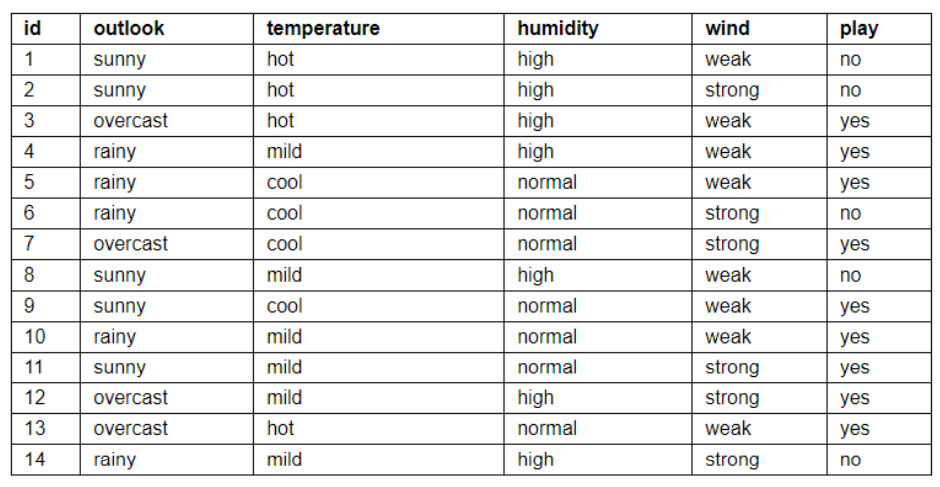
**Họ và tên: Hà Huy Sơn**

**MSSV: 18574802010055**

**BÀI TẬP BUỔI THỰC HÀNH 2**



**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}
* |Cyes| = 9, |Cno| = 5

Tính Entropy:

Info(D) =

**Xét thuộc tính outlook:**

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

- Infooutlook(D) = + + = 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 |

- Infotemperature (D) = + + = 0.911bits

 Gain (temperature) = 0.940bits 0.911bits = 0.029bits

**Xét thuộc tính humidity:**

|  |  |  |
| --- | --- | --- |
| high | yes | 3 |
| no | 4 |
| normal | yes | 6 |
| no | 1 |

- Infohumidity (D) = + = 0.788bits

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

**Xét thuộc tính wind:**

|  |  |  |
| --- | --- | --- |
| weak | yes | 6 |
| no | 2 |
| strong | yes | 3 |
| no | 3 |

- Infowind (D) = + = 0.892bits

 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:**

outlook

sunny

rainy

overcast

humidity

yes

wind

high

normal

no

yes

yes

no

weak

strong

**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 |