# BÀI TẬP CÂY QUYẾT ĐỊNH – DECISION TREE – NHÓM 3

1. **Xây dựng Decision Tree theo 2 thuật toán ID3 (cần trình bày Information Gain) và CART (không cần trình bày Information Gain) cho bộ dữ liệu thời tiết sau:**

| day | outlook | temperature | humidity | wind | decision |
|---|---|---|---|---|---|
| 1 | sunny | hot | high | weak | no |
| 2 | sunny | hot | high | strong | no |
| 3 | overcast | hot | high | weak | yes |
| 4 | rainfall | mild | high | weak | yes |
| 5 | rainfall | cool | normal | weak | yes |
| 6 | rainfall | cool | normal | strong | no |
| 7 | overcast | cool | normal | strong | yes |
| 8 | sunny | mild | high | weak | no |
| 9 | sunny | cool | normal | weak | yes |
| 10 | rainfall | mild | normal | weak | yes |
| 11 | sunny | mild | normal | strong | yes |
| 12 | overcast | mild | high | strong | yes |
| 13 | overcast | hot | normal | weak | yes |
| 14 | rainfall | mild | high | strong | no |

a. ID3 (Sử dụng Hàm số Entropy để đo lường mức độ tinh khiết):

i. Lựa chọn thuộc tính lần 1:

$$\boldsymbol{E(S)} = -\sum_{c=1}^{2}\left(\frac{N_c}{N}\log\left(\frac{N_c}{N}\right)\right) = -\frac{9}{14}\log\left(\frac{9}{14}\right) - \frac{5}{14}\log\left(\frac{5}{14}\right) \approx 0.65$$

**<u>Tính toán Entropy cho thuộc tính "outlook":</u>**

| day | outlook | temperature | humidity | wind | decision |
|---|---|---|---|---|---|
| 1 | sunny | hot | high | weak | no |
| 2 | sunny | hot | high | strong | no |
| 8 | sunny | mild | high | weak | no |
| 9 | sunny | cool | normal | weak | yes |
| 11 | sunny | mild | normal | strong | yes |
| 4 | rainfall | mild | high | weak | yes |
| 5 | rainfall | cool | normal | weak | yes |
| 6 | rainfall | cool | normal | strong | no |
| 10 | rainfall | mild | normal | weak | yes |
| 14 | rainfall | mild | high | strong | no |
| 3 | overcast | hot | high | weak | yes |
| 7 | overcast | cool | normal | strong | yes |
| 12 | overcast | mild | high | strong | yes |
| 13 | overcast | hot | normal | weak | yes |

# BÀI TẬP CÂY QUYẾT ĐỊNH – DECISION TREE – NHÓM 3

$$E('sunny') = -\sum_{c=1}^{2}\left(\frac{N_c}{N}\log\left(\frac{N_c}{N}\right)\right) = -\frac{2}{5}\log\left(\frac{2}{5}\right) - \frac{3}{5}\log\left(\frac{3}{5}\right) \approx 0.67$$

$$E('rainfal') = -\sum_{c=1}^{2}\left(\frac{N_c}{N}\log\left(\frac{N_c}{N}\right)\right) = -\frac{3}{5}\log\left(\frac{3}{5}\right) - \frac{2}{5}\log\left(\frac{2}{5}\right) \approx 0.67$$

$$E('overcast') = -\sum_{c=1}^{2}\left(\frac{N_c}{N}\log\left(\frac{N_c}{N}\right)\right) = -\frac{4}{4}\log\left(\frac{4}{4}\right) + 0 = 0$$

$$E(outlook, S) = \frac{5}{14}E('sunny') + \frac{5}{14}E('rainfall') + \frac{4}{14}E('overcast') \approx 0.48$$

**Tính toán Entropy cho thuộc tính "temperature":**

| day | outlook | temperature | humidity | wind | decision |
|---|---|---|---|---|---|
| 8 | sunny | mild | high | weak | no |
| 11 | sunny | mild | normal | strong | yes |
| 4 | rainfall | mild | high | weak | yes |
| 10 | rainfall | mild | normal | weak | yes |
| 14 | rainfall | mild | high | strong | no |
| 12 | overcast | mild | high | strong | yes |
| 1 | sunny | hot | high | weak | no |
| 2 | sunny | hot | high | strong | no |
| 3 | overcast | hot | high | weak | yes |
| 13 | overcast | hot | normal | weak | yes |
| 9 | sunny | cool | normal | weak | yes |
| 5 | rainfall | cool | normal | weak | yes |
| 6 | rainfall | cool | normal | strong | no |
| 7 | overcast | cool | normal | strong | yes |

$$E('mild') = -\sum_{c=1}^{2}\left(\frac{N_c}{N}\log\left(\frac{N_c}{N}\right)\right) = -\frac{4}{6}\log\left(\frac{4}{6}\right) - \frac{2}{6}\log\left(\frac{2}{6}\right) \approx 0.64$$

$$E('hot') = -\sum_{c=1}^{2}\left(\frac{N_c}{N}\log\left(\frac{N_c}{N}\right)\right) = -\frac{2}{4}\log\left(\frac{2}{4}\right) - \frac{2}{4}\log\left(\frac{2}{4}\right) \approx 0.69$$

$$E('cool') = -\sum_{c=1}^{2}\left(\frac{N_c}{N}\log\left(\frac{N_c}{N}\right)\right) = -\frac{3}{4}\log\left(\frac{3}{4}\right) - \frac{1}{4}\log\left(\frac{1}{4}\right) \approx 0.56$$

$$E(temperature, S) = \frac{4}{14}E('mild') + \frac{6}{14}E('hot') + \frac{4}{14}E('cool') \approx 0.63$$

**Tính toán Entropy cho thuộc tính "humidity":**

| day | outlook | temperature | humidity | wind | decision |
|---|---|---|---|---|---|
| 8 | sunny | mild | high | weak | no |

# BÀI TẬP CÂY QUYẾT ĐỊNH – DECISION TREE – NHÓM 3

| | | | | | |
|---|---|---|---|---|---|
| 14 | rainfall | mild | high | strong | no |
| 1 | sunny | hot | high | weak | no |
| 2 | sunny | hot | high | strong | no |
| 4 | rainfall | mild | high | weak | yes |
| 12 | overcast | mild | high | strong | yes |
| 3 | overcast | hot | high | weak | yes |
| 6 | rainfall | cool | normal | strong | no |
| 11 | sunny | mild | normal | strong | yes |
| 10 | rainfall | mild | normal | weak | yes |
| 13 | overcast | hot | normal | weak | yes |
| 9 | sunny | cool | normal | weak | yes |
| 5 | rainfall | cool | normal | weak | yes |
| 7 | overcast | cool | normal | strong | yes |

$$\boldsymbol{E('high')} = -\sum_{c=1}^{2}\left(\frac{N_c}{N}\log\left(\frac{N_c}{N}\right)\right) = -\frac{3}{7}\log\left(\frac{3}{7}\right) - \frac{4}{7}\log\left(\frac{4}{7}\right) \approx 0.68$$

$$\boldsymbol{E('normal')} = -\sum_{c=1}^{2}\left(\frac{N_c}{N}\log\left(\frac{N_c}{N}\right)\right) = -\frac{6}{7}\log\left(\frac{6}{7}\right) - \frac{1}{7}\log\left(\frac{1}{7}\right) \approx 0.41$$

$$\boldsymbol{E(humidity, S)} = \frac{7}{14}\boldsymbol{E('high')} + \frac{7}{14}\boldsymbol{E('normal')} \approx 0.545$$

**Tính toán Entropy cho thuộc tính "wind":**

| day | outlook | temperature | humidity | wind | decision |
|---|---|---|---|---|---|
| 14 | rainfall | mild | high | strong | no |
| 2 | sunny | hot | high | strong | no |
| 6 | rainfall | cool | normal | strong | no |
| 12 | overcast | mild | high | strong | yes |
| 11 | sunny | mild | normal | strong | yes |
| 7 | overcast | cool | normal | strong | yes |
| 8 | sunny | mild | high | weak | no |
| 1 | sunny | hot | high | weak | no |
| 4 | rainfall | mild | high | weak | yes |
| 3 | overcast | hot | high | weak | yes |
| 10 | rainfall | mild | normal | weak | yes |
| 13 | overcast | hot | normal | weak | yes |
| 9 | sunny | cool | normal | weak | yes |
| 5 | rainfall | cool | normal | weak | yes |

$$\boldsymbol{E('strong')} = -\sum_{c=1}^{2}\left(\frac{N_c}{N}\log\left(\frac{N_c}{N}\right)\right) = -\frac{3}{6}\log\left(\frac{3}{6}\right) - \frac{3}{6}\log\left(\frac{3}{6}\right) \approx 0.69$$

# BÀI TẬP CÂY QUYẾT ĐỊNH – DECISION TREE – NHÓM 3

$$\boldsymbol{E('weak')} = -\sum_{c=1}^{2}\left(\frac{N_c}{N}\log\left(\frac{N_c}{N}\right)\right) = -\frac{6}{8}\log\left(\frac{6}{8}\right)-\frac{2}{8}\log\left(\frac{2}{8}\right) \approx 0.56$$

$$\boldsymbol{E(wind, S)} = \frac{6}{14}\boldsymbol{E('strong')} + \frac{8}{14}\boldsymbol{E('weak')} \approx 0.616$$

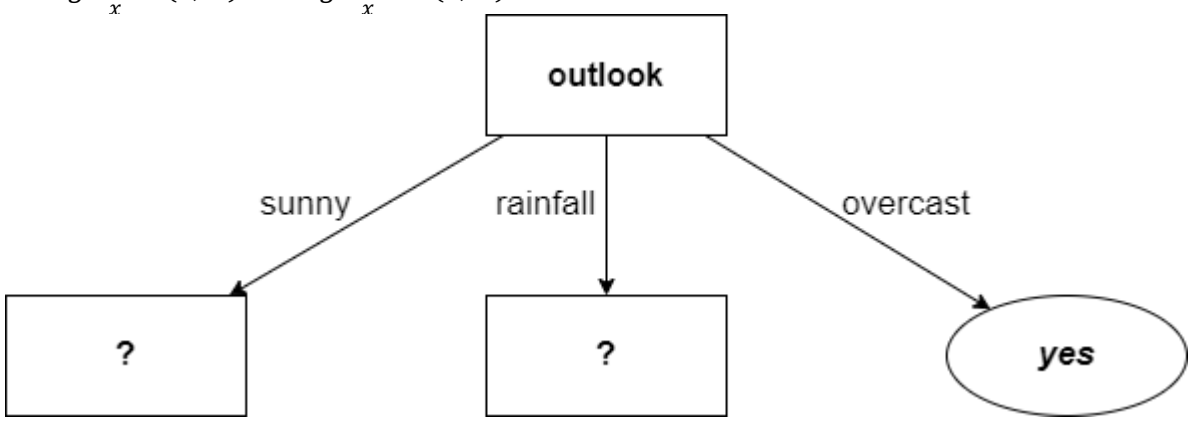
$\boldsymbol{IG(outlook, S)} = \boldsymbol{E(S)} - \boldsymbol{E(outlook, S)} = 0.17$

$\boldsymbol{IG(temperature, S)} = \boldsymbol{E(S)} - \boldsymbol{E(temperature, S)} \approx 0.02$

$\boldsymbol{IG(humidity, S)} = \boldsymbol{E(S)} - \boldsymbol{E(humidity, S)} \approx 0.105$

$\boldsymbol{IG(wind, S)} = \boldsymbol{E(S)} - \boldsymbol{E(wind, S)} \approx 0.034$

$x^* = arg\min_x E(x, \boldsymbol{S}) = arg\max_x IG(x, \boldsymbol{S}) \Rightarrow x = \boldsymbol{outlook}$




ii. Lựa chọn thuộc tính lần 2:

$$\boldsymbol{E(S')} = -\sum_{c=1}^{2}\left(\frac{N_c}{N}\log\left(\frac{N_c}{N}\right)\right) = -\frac{2}{5}\log\left(\frac{2}{5}\right)-\frac{3}{5}\log\left(\frac{3}{5}\right) \approx 0.292$$

**Tính toán Entropy cho thuộc tính "temperature":**

| day | outlook | temperature | humidity | wind | decision |
|---|---|---|---|---|---|
| 9 | sunny | cool | normal | weak | yes |
| 1 | sunny | hot | high | weak | no |
| 2 | sunny | hot | high | strong | no |
| 8 | sunny | mild | high | weak | no |
| 11 | sunny | mild | normal | strong | yes |

$$\boldsymbol{E('cool')} = -\sum_{c=1}^{2}\left(\frac{N_c}{N}\log\left(\frac{N_c}{N}\right)\right) = -\frac{1}{1}\log\left(\frac{1}{1}\right) - 0 = 0$$

$$\boldsymbol{E('hot')} = -\sum_{c=1}^{2}\left(\frac{N_c}{N}\log\left(\frac{N_c}{N}\right)\right) = -0-\frac{2}{2}\log\left(\frac{2}{2}\right) = 0$$

$$\boldsymbol{E('mild')} = -\sum_{c=1}^{2}\left(\frac{N_c}{N}\log\left(\frac{N_c}{N}\right)\right) = -\frac{1}{2}\log\left(\frac{1}{2}\right)-\frac{1}{2}\log\left(\frac{1}{2}\right) \approx 0.301$$

# BÀI TẬP CÂY QUYẾT ĐỊNH – DECISION TREE – NHÓM 3

$$\boldsymbol{E(temperature,\ S')} = \frac{1}{5}\boldsymbol{E('cool')} + \frac{2}{5}\boldsymbol{E('hot')} + \frac{2}{5}\boldsymbol{E('mild')} \approx 0.1204$$

**Tính toán Entropy cho thuộc tính "humidity":**

| day | outlook | temperature | humidity | wind | decision |
|---|---|---|---|---|---|
| 1 | sunny | hot | high | weak | no |
| 2 | sunny | hot | high | strong | no |
| 8 | sunny | mild | high | weak | no |
| 9 | sunny | cool | normal | weak | yes |
| 11 | sunny | mild | normal | strong | yes |

$$\boldsymbol{E('high')} = -\sum_{c=1}^{2}\left(\frac{N_c}{N}\log\left(\frac{N_c}{N}\right)\right) = -0 - \frac{3}{3}\log\left(\frac{3}{3}\right) = 0$$

$$\boldsymbol{E('normal')} = -\sum_{c=1}^{2}\left(\frac{N_c}{N}\log\left(\frac{N_c}{N}\right)\right) = -\frac{2}{2}\log\left(\frac{2}{2}\right) - 0 = 0$$

$$\boldsymbol{E(humidity,\ S')} = \frac{3}{5}\boldsymbol{E('high')} + \frac{2}{5}\boldsymbol{E('normal')} = 0$$

**Tính toán Entropy cho thuộc tính "wind":**

| day | outlook | temperature | humidity | wind | decision |
|---|---|---|---|---|---|
| 2 | sunny | hot | high | strong | no |
| 11 | sunny | mild | normal | strong | yes |
| 1 | sunny | hot | high | weak | no |
| 8 | sunny | mild | high | weak | no |
| 9 | sunny | cool | normal | weak | yes |

$$\boldsymbol{E('strong')} = -\sum_{c=1}^{2}\left(\frac{N_c}{N}\log\left(\frac{N_c}{N}\right)\right) = -\frac{1}{2}\log\left(\frac{1}{2}\right) - \frac{1}{2}\log\left(\frac{1}{2}\right) \approx 0.301$$

$$\boldsymbol{E('weak')} = -\sum_{c=1}^{2}\left(\frac{N_c}{N}\log\left(\frac{N_c}{N}\right)\right) = -\frac{1}{3}\log\left(\frac{1}{3}\right) - \frac{2}{3}\log\left(\frac{2}{3}\right) \approx 0.276$$

$$\boldsymbol{E(wind,\ S')} = \frac{2}{5}\boldsymbol{E('strong')} + \frac{3}{5}\boldsymbol{E('weak')} \approx 0.286$$

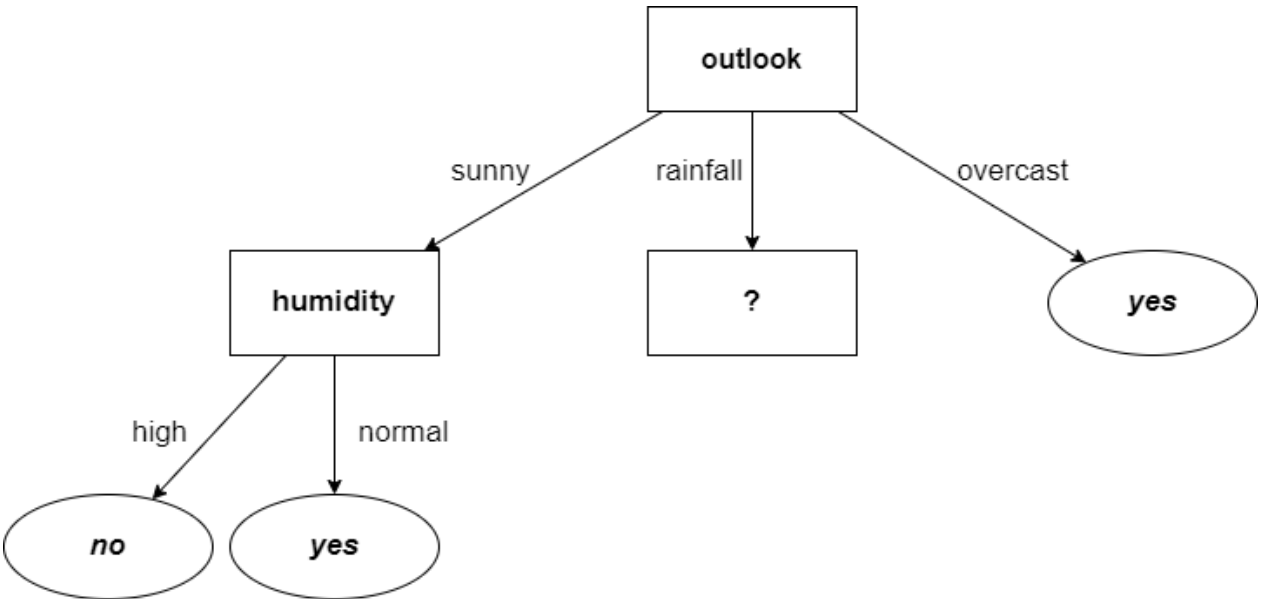
$$\boldsymbol{IG(temperature,\ S')} = \boldsymbol{E(S')} - \boldsymbol{E(temperature,\ S')} \approx 0.1716$$

$$\boldsymbol{IG(humidity,\ S')} = \boldsymbol{E(S')} - \boldsymbol{E(humidity,\ S')} \approx 0.292$$

$$\boldsymbol{IG(wind,\ S')} = \boldsymbol{E(S')} - \boldsymbol{E(wind,\ S')} \approx 0.006$$

$$x^* = arg\min_x E(x,\ \boldsymbol{S'}) = arg\max_x IG(x,\ \boldsymbol{S'}) \Rightarrow x = \boldsymbol{humidity}$$

# BÀI TẬP CÂY QUYẾT ĐỊNH – DECISION TREE – NHÓM 3




iii. Lựa chọn thuộc tính lần 3:

$$\boldsymbol{E(S'')} = -\sum_{c=1}^{2}\left(\frac{N_c}{N}\log\left(\frac{N_c}{N}\right)\right) = -\frac{3}{5}\log\left(\frac{3}{5}\right) - \frac{2}{5}\log\left(\frac{2}{5}\right) \approx 0.292$$

**Tính toán Entropy cho thuộc tính "temperature":**

| day | outlook | temperature | humidity | wind | decision |
|---|---|---|---|---|---|
| 5 | rainfall | cool | normal | weak | yes |
| 6 | rainfall | cool | normal | strong | no |
| 4 | rainfall | mild | high | weak | yes |
| 10 | rainfall | mild | normal | weak | yes |
| 14 | rainfall | mild | high | strong | no |

$$\boldsymbol{E('cool')} = -\sum_{c=1}^{2}\left(\frac{N_c}{N}\log\left(\frac{N_c}{N}\right)\right) = -\frac{1}{2}\log\left(\frac{1}{2}\right) - \frac{1}{2}\log\left(\frac{1}{2}\right) \approx 0.301$$

$$\boldsymbol{E('mild')} = -\sum_{c=1}^{2}\left(\frac{N_c}{N}\log\left(\frac{N_c}{N}\right)\right) = -\frac{2}{3}\log\left(\frac{2}{3}\right) - \frac{1}{3}\log\left(\frac{1}{3}\right) \approx 0.276$$

$$\boldsymbol{E(temperature, S'')} = \frac{2}{5}\boldsymbol{E('cool')} + \frac{3}{5}\boldsymbol{E('mild')} \approx 0.286$$

**Tính toán Entropy cho thuộc tính "humidity":**

| day | outlook | temperature | humidity | wind | decision |
|---|---|---|---|---|---|
| 4 | rainfall | mild | high | weak | yes |
| 14 | rainfall | mild | high | strong | no |
| 5 | rainfall | cool | normal | weak | yes |
| 6 | rainfall | cool | normal | strong | no |
| 10 | rainfall | mild | normal | weak | yes |

# BÀI TẬP CÂY QUYẾT ĐỊNH – DECISION TREE – NHÓM 3

$$\boldsymbol{E('high')} = -\sum_{c=1}^{2}\left(\frac{N_c}{N}\log\left(\frac{N_c}{N}\right)\right) = -\frac{1}{2}\log\left(\frac{1}{2}\right) - \frac{1}{2}\log\left(\frac{1}{2}\right) \approx 0.301$$

$$\boldsymbol{E('normal')} = -\sum_{c=1}^{2}\left(\frac{N_c}{N}\log\left(\frac{N_c}{N}\right)\right) = -\frac{2}{3}\log\left(\frac{2}{3}\right) - \frac{1}{3}\log\left(\frac{1}{3}\right) \approx 0.276$$

$$\boldsymbol{E(humidity, S'')} = \frac{2}{5}\boldsymbol{E('high')} + \frac{3}{5}\boldsymbol{E('normal')} \approx 0.286$$

**Tính toán Entropy cho thuộc tính "wind":**

| day | outlook | temperature | humidity | wind | decision |
|---|---|---|---|---|---|
| 14 | rainfall | mild | high | strong | no |
| 6 | rainfall | cool | normal | strong | no |
| 4 | rainfall | mild | high | weak | yes |
| 5 | rainfall | cool | normal | weak | yes |
| 10 | rainfall | mild | normal | weak | yes |

$$\boldsymbol{E('strong')} = -\sum_{c=1}^{2}\left(\frac{N_c}{N}\log\left(\frac{N_c}{N}\right)\right) = -0 - \frac{2}{2}\log\left(\frac{2}{2}\right) = 0$$

$$\boldsymbol{E('weak')} = -\sum_{c=1}^{2}\left(\frac{N_c}{N}\log\left(\frac{N_c}{N}\right)\right) = -\frac{3}{3}\log\left(\frac{3}{3}\right) - 0 = 0$$

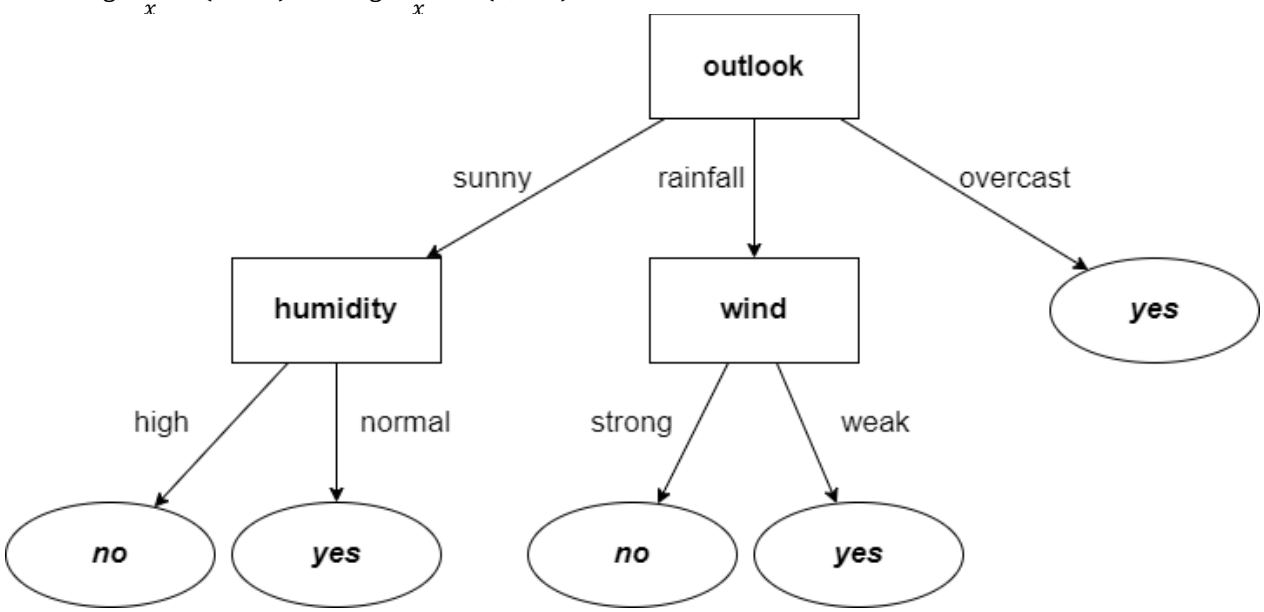
$$\boldsymbol{E(wind, S'')} = \frac{2}{5}\boldsymbol{E('strong')} + \frac{3}{5}\boldsymbol{E('weak')} = 0$$

$$\boldsymbol{IG(temperature, S'')} = \boldsymbol{E(S'')} - \boldsymbol{E(temperature, S'')} \approx 0.006$$

$$\boldsymbol{IG(humidity, S'')} = \boldsymbol{E(S'')} - \boldsymbol{E(humidity, S'')} \approx 0.006$$

$$\boldsymbol{IG(wind, S'')} = \boldsymbol{E(S'')} - \boldsymbol{E(wind, S'')} \approx 0.292$$

$$x^* = arg\min_x E(x, \boldsymbol{S''}) = arg\max_x IG(x, \boldsymbol{S''}) \Rightarrow x = \boldsymbol{wind}$$



# BÀI TẬP CÂY QUYẾT ĐỊNH – DECISION TREE – NHÓM 3

b. CART:

i. Lựa chọn thuộc tính lần 1:

**Thuộc tính "outlook":**

| day | outlook | temperature | humidity | wind | decision |
|---|---|---|---|---|---|
| 1 | sunny | hot | high | weak | no |
| 2 | sunny | hot | high | strong | no |
| 8 | sunny | mild | high | weak | no |
| 9 | sunny | cool | normal | weak | yes |
| 11 | sunny | mild | normal | strong | yes |
| 4 | rainfall | mild | high | weak | yes |
| 5 | rainfall | cool | normal | weak | yes |
| 6 | rainfall | cool | normal | strong | no |
| 10 | rainfall | mild | normal | weak | yes |
| 14 | rainfall | mild | high | strong | no |
| 3 | overcast | hot | high | weak | yes |
| 7 | overcast | cool | normal | strong | yes |
| 12 | overcast | mild | high | strong | yes |
| 13 | overcast | hot | normal | weak | yes |

# BÀI TẬP CÂY QUYẾT ĐỊNH – DECISION TREE – NHÓM 3

| | | | |
|---|---|---|---|
| sunny | 5 | yes | 2 |
| | | no | 3 |
| rainfall | 5 | yes | 3 |
| | | no | 2 |
| overcast | 4 | yes | 4 |
| | | no | 0 |

| Rule | Error | Total Error |
|---|---|---|
| Sunny → no | 2/5 | 4/14 |
| Rainfall → yes | 2/5 | |
| Overcast → yes | 0/4 | |

**Thuộc tính "temperature":**

| day | outlook | temperature | humidity | wind | decision |
|---|---|---|---|---|---|
| 8 | sunny | mild | high | weak | no |
| 11 | sunny | mild | normal | strong | yes |
| 4 | rainfall | mild | high | weak | yes |
| 10 | rainfall | mild | normal | weak | yes |
| 14 | rainfall | mild | high | strong | no |
| 12 | overcast | mild | high | strong | yes |
| 1 | sunny | hot | high | weak | no |
| 2 | sunny | hot | high | strong | no |
| 3 | overcast | hot | high | weak | yes |
| 13 | overcast | hot | normal | weak | yes |
| 9 | sunny | cool | normal | weak | yes |
| 5 | rainfall | cool | normal | weak | yes |
| 6 | rainfall | cool | normal | strong | no |
| 7 | overcast | cool | normal | strong | yes |

| | | | |
|---|---|---|---|
| Mild | 6 | yes | 4 |
| | | no | 2 |
| Hot | 4 | yes | 2 |
| | | no | 2 |
| Cool | 4 | yes | 3 |
| | | no | 1 |

| Rule | Error | Total Error |
|---|---|---|
| Mild → yes | 2/6 | 5/14 |
| Hot → yes | 2/4 | |
| Cool → yes | 1/4 | |

**Thuộc tính "humidity":**

| day | outlook | temperature | humidity | wind | decision |
|---|---|---|---|---|---|

# BÀI TẬP CÂY QUYẾT ĐỊNH – DECISION TREE – NHÓM 3

| | | | | | |
|---|---|---|---|---|---|
| 8 | sunny | mild | high | weak | no |
| 14 | rainfall | mild | high | strong | no |
| 1 | sunny | hot | high | weak | no |
| 2 | sunny | hot | high | strong | no |
| 4 | rainfall | mild | high | weak | yes |
| 12 | overcast | mild | high | strong | yes |
| 3 | overcast | hot | high | weak | yes |
| 6 | rainfall | cool | normal | strong | no |
| 11 | sunny | mild | normal | strong | yes |
| 10 | rainfall | mild | normal | weak | yes |
| 13 | overcast | hot | normal | weak | yes |
| 9 | sunny | cool | normal | weak | yes |
| 5 | rainfall | cool | normal | weak | yes |
| 7 | overcast | cool | normal | strong | yes |

| | | | |
|---|---|---|---|
| High | 7 | yes | 3 |
| | | no | 4 |
| Normal | 7 | yes | 6 |
| | | no | 1 |

| **Rule** | **Error** | **Total Error** |
|---|---|---|
| High → no | 3/7 | 4/14 |
| Normal → yes | 1/7 | |

**Thuộc tính "wind":**

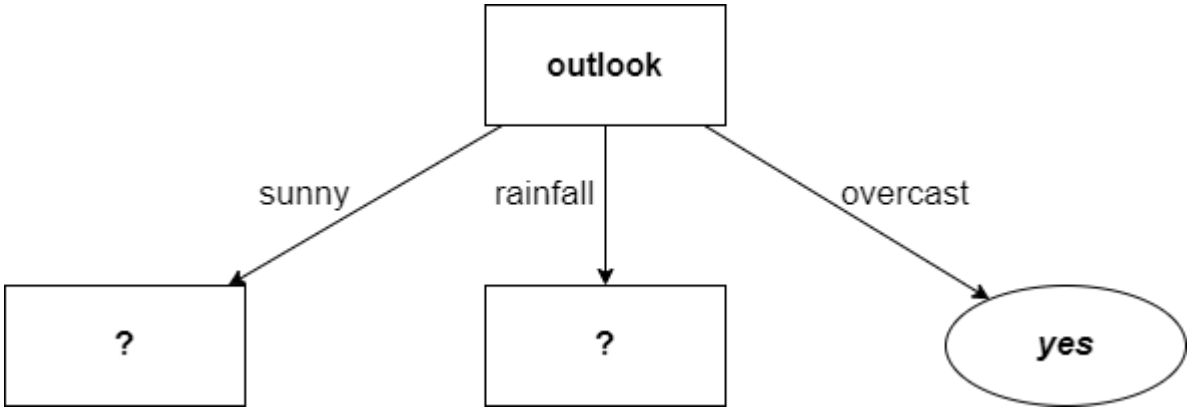
| day | outlook | temperature | humidity | wind | decision |
|---|---|---|---|---|---|
| 14 | rainfall | mild | high | strong | no |
| 2 | sunny | hot | high | strong | no |
| 6 | rainfall | cool | normal | strong | no |
| 12 | overcast | mild | high | strong | yes |
| 11 | sunny | mild | normal | strong | yes |
| 7 | overcast | cool | normal | strong | yes |
| 8 | sunny | mild | high | weak | no |
| 1 | sunny | hot | high | weak | no |
| 4 | rainfall | mild | high | weak | yes |
| 3 | overcast | hot | high | weak | yes |
| 10 | rainfall | mild | normal | weak | yes |
| 13 | overcast | hot | normal | weak | yes |
| 9 | sunny | cool | normal | weak | yes |
| 5 | rainfall | cool | normal | weak | yes |

| | | | |
|---|---|---|---|
| Strong | 6 | yes | 3 |

| | | | |
|---|---|---|---|
| | | no | 3 |
| Weak | 8 | yes | 6 |
| | | no | 2 |

| Rule | Error | Total Error |
|---|---|---|
| Strong → yes | 3/6 | 5/14 |
| Weak → yes | 2/8 | |

| Attribute | Total Error |
|---|---|
| Outlook | 4/14 |
| Temperature | 5/14 |
| Humidity | 4/14 |
| Wind | 5/14 |
| Outlook/Humidity → Outlook | |




ii. Lựa chọn thuộc tính lần 2:

**Thuộc tính "temperature":**

| day | outlook | temperature | humidity | wind | decision |
|---|---|---|---|---|---|
| 9 | sunny | cool | normal | weak | yes |
| 1 | sunny | hot | high | weak | no |
| 2 | sunny | hot | high | strong | no |
| 8 | sunny | mild | high | weak | no |
| 11 | sunny | mild | normal | strong | yes |

| | | | |
|---|---|---|---|
| Mild | 2 | yes | 1 |
| | | no | 1 |
| Hot | 2 | yes | 0 |
| | | no | 2 |
| Cool | 1 | yes | 1 |
| | | no | 0 |

# BÀI TẬP CÂY QUYẾT ĐỊNH – DECISION TREE – NHÓM 3

| Rule | Error | Total Error |
|---|---|---|
| Mild → yes | 1/2 | 1/5 |
| Hot → no | 0/2 | |
| Cool → yes | 0/1 | |

**Thuộc tính "humidity":**

| day | outlook | temperature | humidity | wind | decision |
|---|---|---|---|---|---|
| 1 | sunny | hot | high | weak | no |
| 2 | sunny | hot | high | strong | no |
| 8 | sunny | mild | high | weak | no |
| 9 | sunny | cool | normal | weak | yes |
| 11 | sunny | mild | normal | strong | yes |

| | | | |
|---|---|---|---|
| High | 3 | yes | 0 |
| | | no | 3 |
| Normal | 2 | yes | 2 |
| | | no | 0 |

| Rule | Error | Total Error |
|---|---|---|
| High → no | 0/3 | 0/5 |
| Normal → yes | 0/2 | |

**Thuộc tính "wind":**

| day | outlook | temperature | humidity | wind | decision |
|---|---|---|---|---|---|
| 2 | sunny | hot | high | strong | no |
| 11 | sunny | mild | normal | strong | yes |
| 1 | sunny | hot | high | weak | no |
| 8 | sunny | mild | high | weak | no |
| 9 | sunny | cool | normal | weak | yes |

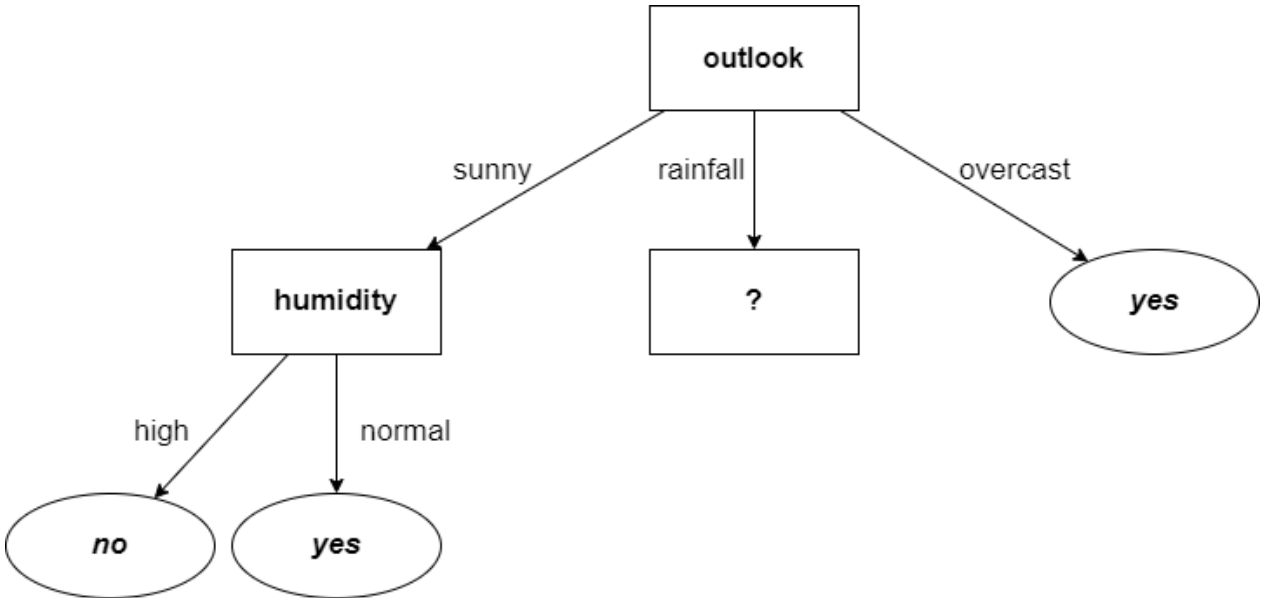
| | | | |
|---|---|---|---|
| Strong | 2 | yes | 1 |
| | | no | 1 |
| Weak | 3 | yes | 1 |
| | | no | 2 |

| Rule | Error | Total Error |
|---|---|---|
| Strong → yes | 1/2 | 2/5 |
| Weak → no | 1/3 | |

| Attribute | Total Error |
|---|---|

# BÀI TẬP CÂY QUYẾT ĐỊNH – DECISION TREE – NHÓM 3

| | |
|---|---|
| Temperature | 1/5 |
| Humidity | 0/5 |
| Wind | 2/5 |
| Humidity | |




iii. Lựa chọn thuộc tính lần 3:

**Thuộc tính "temperature":**

| day | outlook | temperature | humidity | wind | decision |
|---|---|---|---|---|---|
| 5 | rainfall | cool | normal | weak | yes |
| 6 | rainfall | cool | normal | strong | no |
| 4 | rainfall | mild | high | weak | yes |
| 10 | rainfall | mild | normal | weak | yes |
| 14 | rainfall | mild | high | strong | no |

| | | | |
|---|---|---|---|
| Mild | 3 | yes | 2 |
| | | no | 1 |
| Cool | 2 | yes | 1 |
| | | no | 1 |

| Rule | Error | Total Error |
|---|---|---|
| Mild → yes | 1/3 | 2/5 |
| Cool → yes | 1/2 | |

**Thuộc tính "humidity":**

| day | outlook | temperature | humidity | wind | decision |
|---|---|---|---|---|---|
| 4 | rainfall | mild | high | weak | yes |
| 14 | rainfall | mild | high | strong | no |

# BÀI TẬP CÂY QUYẾT ĐỊNH – DECISION TREE – NHÓM 3

| | | | | | |
|---|---|---|---|---|---|
| 5 | rainfall | cool | normal | weak | yes |
| 6 | rainfall | cool | normal | strong | no |
| 10 | rainfall | mild | normal | weak | yes |

| | | | |
|---|---|---|---|
| High | 2 | yes | 1 |
| | | no | 1 |
| Normal | 3 | yes | 2 |
| | | no | 1 |

| Rule | Error | Total Error |
|---|---|---|
| High → yes | 1/2 | 2/5 |
| Normal → yes | 1/3 | |

**Thuộc tính "wind":**

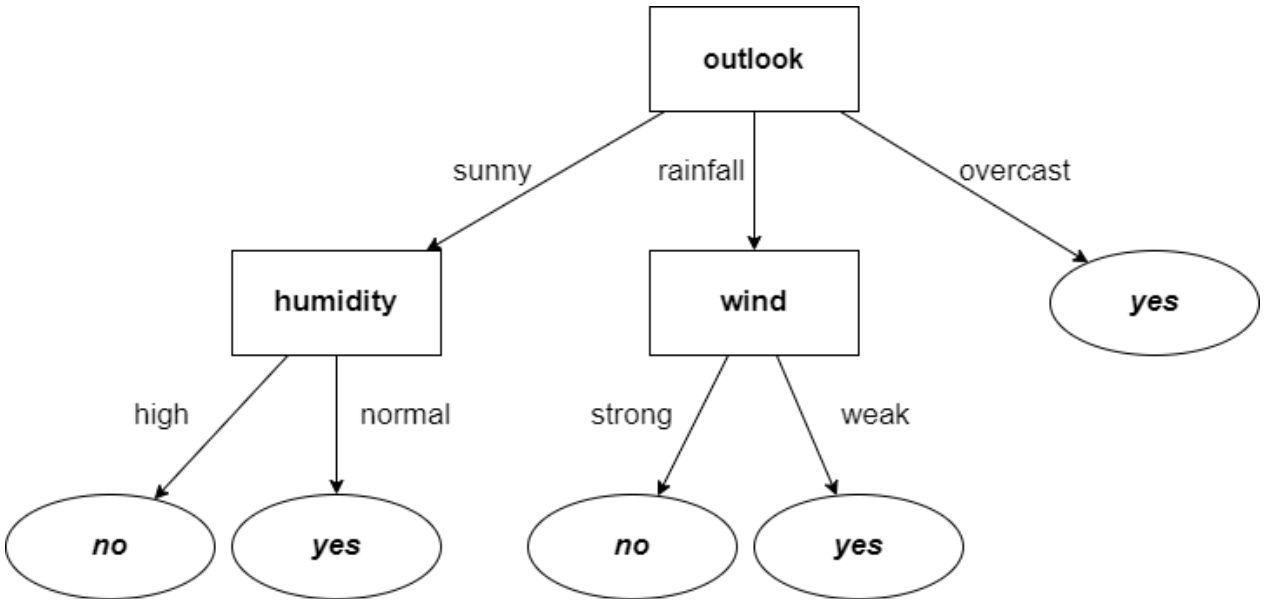
| day | outlook | temperature | humidity | wind | decision |
|---|---|---|---|---|---|
| 14 | rainfall | mild | high | strong | no |
| 6 | rainfall | cool | normal | strong | no |
| 4 | rainfall | mild | high | weak | yes |
| 5 | rainfall | cool | normal | weak | yes |
| 10 | rainfall | mild | normal | weak | yes |

| | | | |
|---|---|---|---|
| Strong | 2 | yes | 0 |
| | | no | 2 |
| Weak | 3 | yes | 3 |
| | | no | 0 |

| Rule | Error | Total Error |
|---|---|---|
| Strong → no | 0/2 | 0/5 |
| Weak → yes | 0/3 | |

| Attribute | Total Error |
|---|---|
| Temperature | 2/5 |
| Humidity | 2/5 |
| Wind | 0/5 |
| Wind | |

# BÀI TẬP CÂY QUYẾT ĐỊNH – DECISION TREE – NHÓM 3




## 2. Trong model Tree Classification nếu model bị Overfitting thì nên điều chỉnh tham số nào dưới đây để tránh bị Overfitting.

```
DecisionTreeClassifier(criterion="entropy",max_depth=3)
```

*Vấn đề Overfitting là vấn đề xảy ra khi Classification Tree học trên một bộ dữ liệu quá lớn. Sau khi học, Classification Tree cho ra kết quả rất tốt trên khi test trên tập train nhưng lại cho ra kết quả với độ chính xác thấp khi áp dụng vào thực tế.*

*Vấn đề này xảy ra tương tự với con người, gọi là học khớp/thuộc lòng. Classification Tree lúc này không còn mang tính tổng quát mà nữa mà nó sẽ phụ thuộc vào tập dữ liệu train, không thể áp dụng vào thực tế được nữa. Khi Classification Tree overfitting thì đặc điểm dễ nhận biết là cây sẽ quá phức tạp.*

*Để khắc phục vấn đề này, ta có 2 hướng giải quyết là **tiền xử lý** và **hậu xử lý**.*

*Tiền xử lý (can thiệp, quy định, đặt ra một số luật/quy định cho cây trước khi xây dựng cây):*

- *Tạo node lá sau khi đạt một số lượng điểm dữ liệu nào đó.*
- *Giới hạn độ sâu cây.*
- *Giới hạn tổng số node lá của cây.*
- *Luật khi chọn thuộc tính để phân nhánh, nếu việc phân nhánh tại một node không làm thay đổi quá nhiều thì ta có thể tiến hành phân lớp ngay lập tức và chấp nhận sai sót.*

*Hậu xử lý (can thiệp vào cây sau khi xây dựng hoàn tất): Tỉa cây (biến non-leaf thành leaf)*

*Câu trả lời:* Theo dòng mã trên thì ta sẽ điều chỉnh tham số độ sâu (max_depth) của cây để tránh trường hợp Overfitting xảy ra.

# BÀI TẬP CÂY QUYẾT ĐỊNH – DECISION TREE – NHÓM 3

**BÀI TẬP CÂY QUYẾT ĐỊNH – DECISION TREE – NHÓM 3**

THÀNH VIÊN NHÓM:


1. Phạm Bùi Nhật Huy – 20521410
2. Nguyễn Thị Như Vân – 20520855
3. Nguyễn Minh Trí – 19522388