

**Bài 5.**

Với bộ dữ liệu của bài 4, hãy xây dựng cây quyết định sử dụng thuật toán C4.5 (dùng Gain Ratio thay cho Information Gain)

**Bài 6.**

Cho tập dữ liệu huấn luyện với 4 thuộc tính như trong bảng sau:

Day	Outlook	Tempreture	Humidity	Wind	Play tennis
D <sub>i</sub>	Sunny	Hot	High	Weak	No

D <sub>1</sub>	Sunny	Hot	High	Strong	No
D <sub>2</sub>	Overcast	Hot	High	Strong	Yes
D <sub>3</sub>	Rain	Mild	Normal	Weak	Yes
D <sub>4</sub>	Rain	Cool	Normal	Weak	Yes
D <sub>5</sub>	Rain	Cool	Normal	Strong	No
D <sub>6</sub>	Overcast	Cool	Normal	Strong	No
D <sub>7</sub>	Sunny	Mild	High	Weak	No
D <sub>8</sub>	Sunny	Cool	Normal	Strong	Yes
D <sub>9</sub>	Rain	Mild	High	Weak	Yes
D <sub>10</sub>	Sunny	Mild	Normal	Strong	Yes
D <sub>11</sub>	Overcast	Mild	Normal	Strong	Yes
D <sub>12</sub>	Overcast	Hot	Normal	Weak	Yes
D <sub>13</sub>	Rain	Mild	High	Strong	No
D <sub>14</sub>	Sunny	Mild	Normal	Weak	Yes
D <sub>15</sub>	Overcast	Cool	High	Strong	No
D <sub>16</sub>	Overcast	Cool	High	Strong	No

a. Hãy xây dựng cây quyết định T1 tương ứng với bộ dữ liệu trên sử dụng thuật toán ID3.

b. Hãy xây dựng cây quyết định T2 tương ứng với bộ dữ liệu trên sử dụng thuật toán C4.5 (dùng Gain Ratio thay cho Information Gain)

c. Cho mẫu quan sát mới D=<Sunny, Mild, High, Strong>, tìm kết quả “Choi” hay “Không chơi” tennis tương ứng với cây T1 và T2 vừa xây dựng được

Bài 6

Đưa vào đề bài ta có 2 nhánh lớp chính

C<sub>1</sub>= yes (gấp nhánh)

C<sub>2</sub>= No (2 nhánh)

$$H(S) = -\frac{9}{16} \log_2 \frac{9}{16} - \frac{7}{16} \log_2 \frac{7}{16}$$

$$= 0,9886$$

$$IG(S, \text{Outlook}) = \text{Entropy}(S) - \frac{6}{16} \text{Entropy}(S_{\text{Sunny}}) - \frac{5}{16} \text{Entropy}(S_{\text{Overcast}})$$

$$- \frac{5}{16} \text{Entropy}(S_{\text{Rain}})$$

$$= 0,9886 - \frac{6}{16} \left( -\frac{3}{6} \log_2 \frac{3}{6} - \frac{3}{6} \log_2 \frac{3}{6} \right) -$$

$$\frac{5}{16} \left( -\frac{3}{5} \log_2 \frac{3}{5} - \frac{2}{5} \log_2 \frac{2}{5} \right) -$$

$$\frac{5}{16} \left( -\frac{3}{5} \log_2 \frac{3}{5} - \frac{2}{5} \log_2 \frac{2}{5} \right)$$

$$= 6,855 \cdot 10^{-3}$$

IG(S, Temperature)

$$= 0,9886 - \frac{4}{16} \text{Entropy}(S_{\text{Hot}}) - \frac{7}{16} \text{Entropy}(S_{\text{Mild}})$$

$$- \frac{5}{16} \text{Entropy}(S_{\text{Cool}})$$

$$= 0,9886 - \frac{4}{16} \left( -\frac{2}{4} \log_2 \frac{2}{4} - \frac{2}{4} \log_2 \frac{2}{4} \right)$$

$$- \frac{7}{16} \left( -\frac{5}{7} \log_2 \frac{5}{7} - \frac{2}{7} \log_2 \frac{2}{7} \right)$$

$$- \frac{5}{16} \left( -\frac{2}{5} \log_2 \frac{2}{5} - \frac{3}{5} \log_2 \frac{3}{5} \right)$$

$$= 0,0525$$

IG(S, Humidity)

$$= 0,9886 - \frac{7}{16} \left( -\frac{2}{7} \log_2 \frac{2}{7} - \frac{5}{7} \log_2 \frac{5}{7} \right)$$

$$- \frac{9}{16} \left( -\frac{2}{9} \log_2 \frac{2}{9} - \frac{2}{9} \log_2 \frac{2}{9} \right)$$

$$= 0,18111$$

IG(S, Wind)

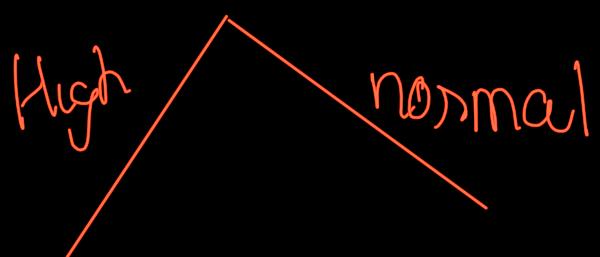
$$= 0,9886 - \frac{9}{16} \left( -\frac{4}{9} \log_2 \frac{4}{9} - \frac{5}{9} \log_2 \frac{5}{9} \right)$$

$$- \frac{7}{16} \left( -\frac{2}{7} \log_2 \frac{2}{7} - \frac{5}{7} \log_2 \frac{5}{7} \right)$$

$$= 0,0525$$

⇒ Vô cùng Humidity là lớn nhất

Humidity



T<sub>1</sub> Tập 1

T<sub>1</sub> Tập 1

g Tập

T<sub>2</sub> Tập 2

Xét Tập 1 =>   
 Yes : 1 Tập  
 No : 5 Tập

$$\text{Entropy}(T_1) = -\frac{5}{7} \log_2 \frac{5}{7} - \frac{2}{7} \log_2 \frac{2}{2}$$

$$= 0,86312$$

$$IG(T_1, \text{outlook}) = 0,86312 - \frac{3}{7} \left( -\frac{1}{3} \log_2 \frac{1}{3} \right)$$

$$- \frac{2}{7} \left( -\frac{1}{2} \log_2 \frac{1}{2} - \frac{1}{2} \log_2 \frac{1}{2} \right)$$

$$- \frac{2}{7} \left( -\frac{1}{2} \log_2 \frac{1}{2} - \frac{1}{2} \log_2 \frac{1}{2} \right)$$

$$= 0,2916$$

$$IG(T_1, \text{temp}) = 0,86312 - \frac{3}{2} \left( -\frac{2}{3} \log_2 \frac{2}{3} - \frac{1}{3} \log_2 \frac{1}{3} \right)$$

$$- \frac{3}{2} \left( -\frac{2}{3} \log_2 \frac{2}{3} - \frac{1}{3} \log_2 \frac{1}{3} \right)$$

$$- \frac{1}{2} \left( -\frac{1}{1} \log_2 \frac{1}{1} \right)$$

$$= 0,076$$

$$IG(T_1, \text{Wind}) = 0,86312 - \frac{4}{2} \left( -\frac{3}{4} \log_2 \frac{3}{4} - \frac{1}{4} \log_2 \frac{1}{4} \right)$$

$$-\frac{3}{2} \left( -\frac{2}{3} \log_2 \frac{2}{3} - \frac{1}{3} \log_2 \frac{1}{3} \right)$$

$$= 5,9771 \cdot 10^{-3}$$

$\Rightarrow$  Outlook là phán của bằng  $T_1$

Tương tự thì ta sẽ được ô bằng  $T_2$ .  
(normal)

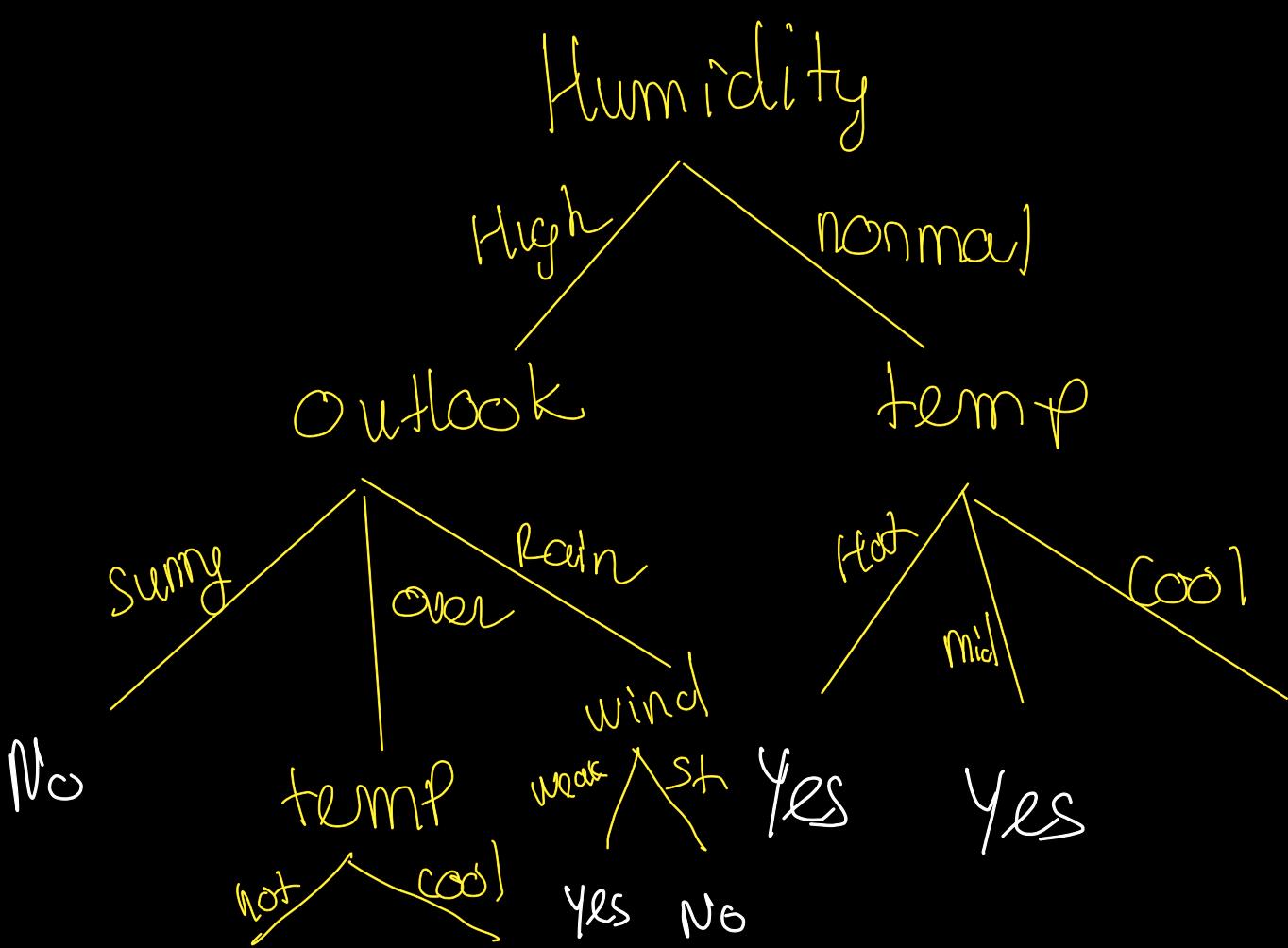
$$\text{Entropy } C(T_2) = 0,2642$$

$$IG(T_2, \text{Outlook}) = 0,152$$

$$IG(T_2, \text{Temp}) = 0,31975$$

$$IG(T_2, \text{Wind}) = 0,2242$$

→ Temp là phẩn hép úc của ô bằng  $T_2$



Yes

No

Mild	weak	yes	fair
Mild	Strong	No	

Rain	weak	yes	cold
Rain	Strong	No	
overcast	Strong	No	
Sunny	Strong	yes	

Bài 1 Sử dụng precision, recall, F1  
(tài liệu pdf)