No. .....
Date .....

Bài tập về cây quyết định

Theo bảng dữ liệu ta có:

- $|D| = 8$
- $C = \{$Bị rám, không$\}$
- $|C_{\text{bị rám}}| = 3$, $|C_{\text{không}}| = 5$

Tính Entropy

$$\text{Info}(D) = -\frac{3}{8} \times \left(\log_2 \frac{3}{8}\right) - \frac{5}{8} \times \left(\log_2 \frac{5}{8}\right) = 0,954 \text{ bits}$$

Xét các thuộc tính màu tóc

Đen	Bị rám	2
	không	2
Râm	Bị rám	0
	Không	3
Bạc	Bị rám	1
	không	0

$$\text{Info}_{\text{màu tóc}}(D) = \frac{4}{8} \times \left(-\frac{2}{4}\log_2 \frac{2}{4} - \frac{2}{4} \times \log_2 \frac{2}{4}\right) + \frac{3}{8} \times \left(-\frac{3}{8}\log_2 \frac{3}{3}\right) + \frac{1}{8} \times \left(-1 \log_2 1\right)$$

$$= 0,5 \text{ bits}$$

$\Rightarrow$ Gain (màu tóc) $= 0,954 \text{ bits} - 0,5 \text{ bits} = 0,454 \text{ bits}$

Được quét bằng CamScanner

No. .....
Date .....

Xét thuộc tính chiều cao

Cao	Bị rám	0
	Không	2
Thấp	Bị rám	2
	Không	1
Tầm thước	Bị rám	1
	Không	2

$$\text{Info}_{\text{chiều cao}}(D) = \frac{2}{8} \times \left(-\frac{2}{2} \times \log_2 \frac{2}{2}\right) + \frac{3}{8}\left(-\frac{2}{3}\log_2 \frac{2}{3} - \frac{1}{3}\log_2 \frac{1}{3}\right)$$
$$+ \frac{3}{8} \times \left(-\frac{1}{3} \times \log_2 \frac{1}{3} - \frac{2}{3} \times \log_2 \frac{2}{3}\right)$$
$$= 0,689 \text{ bits}$$

→ $\text{Gian}_{\text{(chiều cao)}} = 0,954 \text{ bits} - 0,689 \text{ bit} = 0,265 \text{ bits}$

Được quét bằng CamScanner

NO. .....
Date .....

Xét thuộc tính cân nặng

Nhẹ	Bị rám	1
	Không	1
Nặng	Bị rám	1
	Không	2
Vừa phải	Bị rám	1
	Không	2

$$\text{Info cân nặng}(D) = \frac{2}{8} \times \left(-\frac{1}{2} \times \log_2 \frac{1}{2} - \frac{1}{2} \times \log_2 \frac{1}{2}\right)$$
$$+ \frac{3}{8} \times \left(-\frac{1}{3} \log_2 \frac{1}{3} - \frac{2}{3} \log_2 \frac{2}{3}\right)$$
$$+ \frac{3}{8} \times \left(-\frac{1}{3} \log_2 \frac{1}{3} - \frac{2}{3} \log_2 \frac{2}{3}\right)$$
$$= 0,939 \text{ bits}$$

$\Rightarrow$ Gain (cân nặng) $= 0,954 - 0,939 = 0,015$ bits

Được quét bằng CamScanner

Xét thuộc tính dùng thuốc

Không	Bị rám	3
	Không	2
Có	Bị rám	0
	Không	3

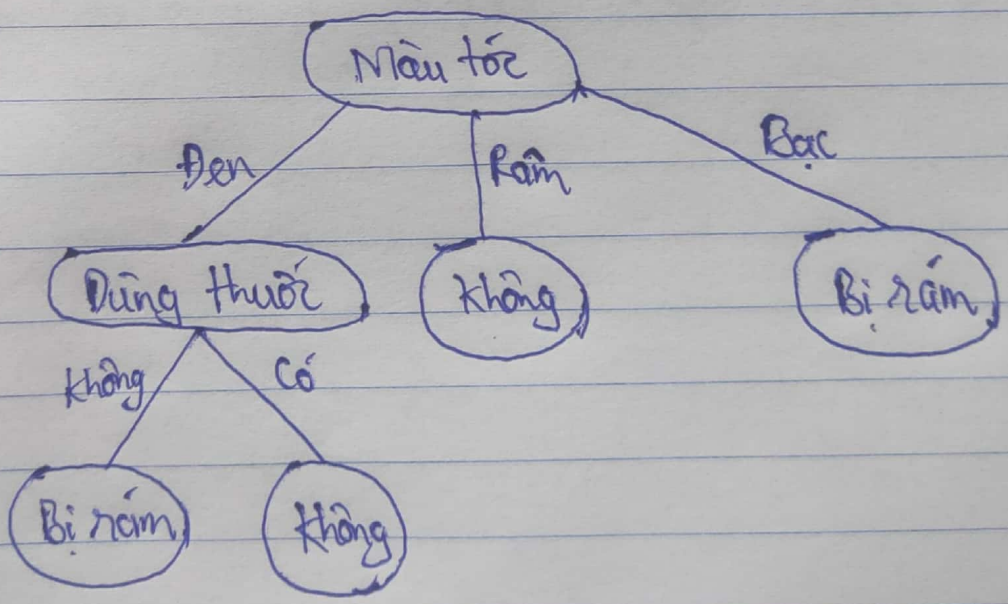
$$\text{Info}_{\text{dùng thuốc}}(D) = \frac{5}{8} \times \left(-\frac{3}{5} \times \log_2 \frac{3}{5} - \frac{2}{5} \times \log_2 \frac{2}{5}\right) + \frac{3}{8} \times \left(\frac{3}{3} \times \log_2 \frac{3}{3}\right)$$

$$= 0.607 \text{ bits}$$

$\Rightarrow$ Gain(dùng thuốc) = 0,954 - 0,607 = 0,347 bits

Vì thuộc tính màu tóc có độ thông tin lớn nhất (0,454 bits) nên màu tóc là thuộc tính được chọn để phân tách (rẽ nhánh)

* Ta có cây quyết định:



Được quét bằng CamScanner