BÀI TẬP KHAI PHÁ DỮ LIỆU

Vẽ 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

- 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 thuộc tính màu tóc:

Đen	Bị rám	2
	Không	2
Rám	Bị rám	0
	Không	3
Bạc	Bị rám	1
	Không	0

$\text{Info}_{\text{màu tóc}}$ (D) = $\frac{4}{8} \times \left(-\frac{2}{4} \times \log_2 \frac{2}{4} - \frac{2}{4} \times \log_2 \frac{2}{4}\right) + \frac{3}{8} \times \left(-\frac{3}{3} \times \log_2 \frac{3}{3}\right) + \frac{1}{8} \times (-1 \times \log_2 1) = 0.5\text{bits}$

➔ Gain (màu tóc) = 0.954bits – 0.5bits = 0.454bits

Xét thuộc tính chiều cao:

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

$\text{Info}_{\text{chiều cao}}$ (D) = $\frac{2}{8} \times \left(-\frac{2}{2} \times \log_2 \frac{2}{2}\right) + \frac{3}{8} \times \left(-\frac{2}{3} \times \log_2 \frac{2}{3} - \frac{1}{3} \times \log_2 \frac{1}{3}\right) + \frac{3}{8} \times \left(-\frac{1}{3} \times \log_2 \frac{1}{3} - \frac{2}{3} \times \log_2 \frac{2}{3}\right) = 0.689\text{bits}$

➔ Gain (chiều cao) = 0.954bits – 0.689bits = 0.265bits

Xét thuộc tính cân nặng:

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

$\text{Info}_{\text{cân nặng}}(D) = \frac{2}{8} \times \left(-\frac{1}{2} \times \log_2 \frac{1}{2} - \frac{1}{2} \times \log_2 \frac{1}{2}\right) + \frac{3}{8} \times \left(-\frac{1}{3} \times \log_2 \frac{1}{3} - \frac{2}{3} \times \log_2 \frac{2}{3}\right) + \frac{3}{8} \times \left(-\frac{1}{3} \times \log_2 \frac{1}{3} - \frac{2}{3} \times \log_2 \frac{2}{3}\right) = 0.939\text{bits}$

➔ Gain (cân nặng) = 0.954bits – 0.939bits = 0.015bits

Xét thuộc tính dùng thuốc:

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

$\text{Info}_{\text{dùng thuốc}}(D) = \frac{5}{8} \times \left(-\frac{3}{5} \times \log_2 \frac{3}{5} - \frac{2}{5} \times \log_2 \frac{2}{5}\right) + \frac{3}{8} \times \left(-\frac{3}{3} \times \log_2 \frac{3}{3}\right) = 0.607\text{bits}$

➔ Gain (dùng thuốc) = 0.954bits – 0.607bits = 0.347bit

Vì thuộc tính *màu tóc* có độ lợi thông tin lớn nhất (0.454bits) nên *màu tóc* là thuộc tính được chọn để phân tách (rẽ nhánh)

Ta có cây quyết định: