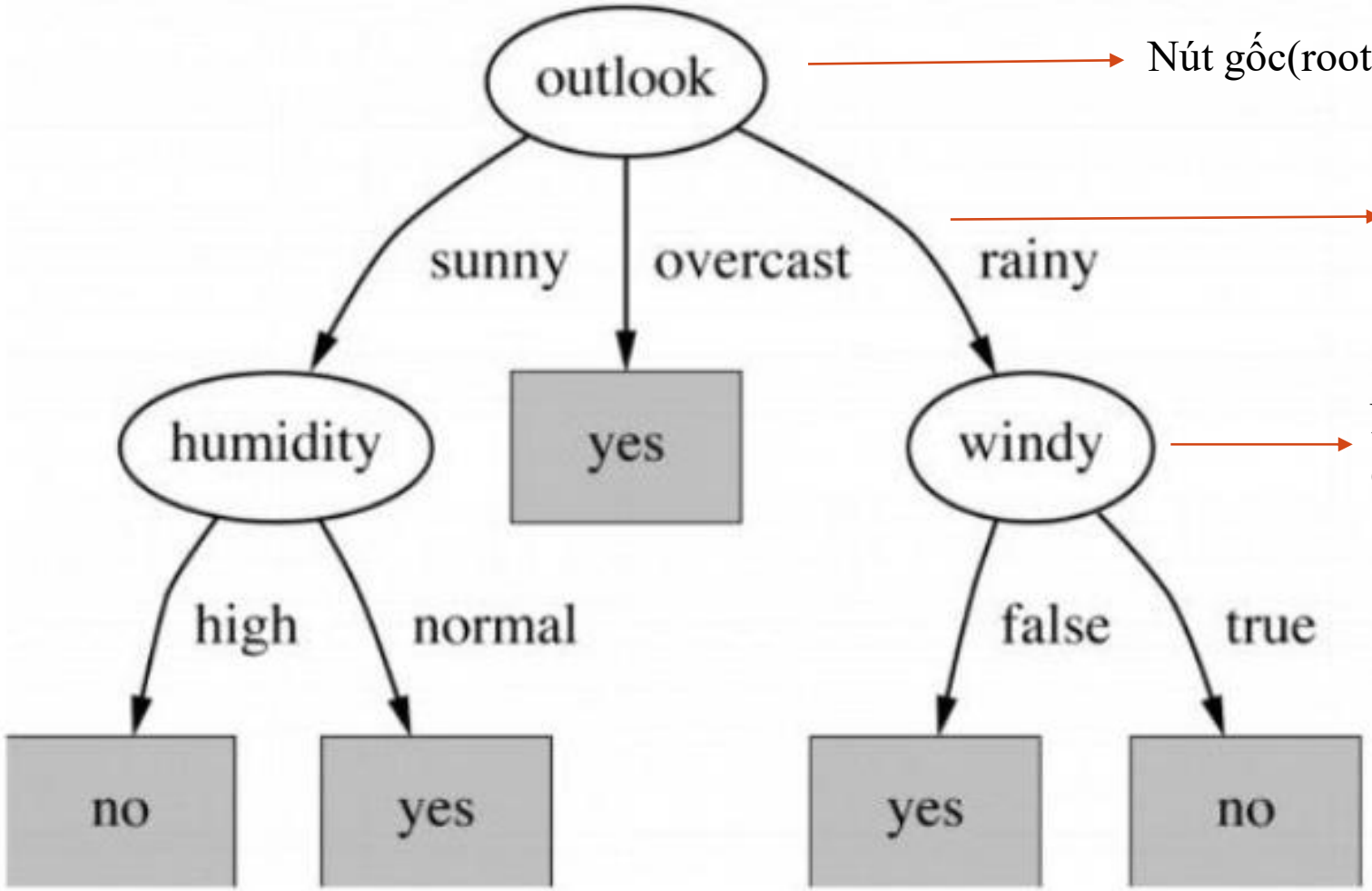
THUẬT TOÁN C4.5

Tác giả: Phan Anh Trúc - 3116410131

1. CÂY QUYẾT ĐỊNH (DECISION TREE)

- **Cây quyết định** là một kiểu mô hình dự báo (*predictive model*), nghĩa là một ánh xạ từ các quan sát về một sự vật/hiện tượng tới các kết luận về giá trị mục tiêu của sự vật/hiện tượng. Các cây quyết định được dùng để hỗ trợ quá trình ra quyết định.
- **Cây quyết định có hai loại:**
  - Cây hồi quy (Regression tree)
  - Cây phân loại (Classification tree)

- **Ưu điểm của cây quyết định:**
  - Đơn giản, dễ hiểu
  - Có thể xử lý dữ liệu số và dữ liệu có giá trị là tên loại
  - Mang lại kết quả dự báo có độ chính xác cao, dễ dàng thực hiện
- **Nhược điểm:**
  - Khó giải quyết được những vấn đề có dữ liệu phụ thuộc thời gian liên tục.
  - Dễ xảy ra lỗi khi có quá nhiều lớp chi phí tính toán để xây dựng mô hình cây quyết định cao
- **Các công thức:**
  - Gini impurity được sử dụng trong CART.
  - Entropy được dùng trong ID3, C4.5, C5.0



Nút gốc(root node): là nút trên cùng của cây dùng để phân nhánh

Nhánh(branch): thể hiện giá trị cụ thể của biến

Nút trong(internal node): là biến chứa các giá trị dùng cho phân nhánh tiếp theo

Nút lá(leaf node): là giá trị dự đoán cho biến mục tiêu

2. THUẬT TOÁN C4.5

- Thuật toán C4.5 (1993) là phần cải tiến vượt qua những giới hạn của thuật toán ID3.
- ID3 là một thuật toán được phát minh bởi Ross Quinlan (1986), được sử dụng để tạo ra cây quyết định từ bộ dữ liệu
- Thuật toán C5.0/See5 (C5.0 cho Unix, Linux, See5 cho Windows) cải thiện những hạn chế của thuật toán C4.5

- Cụ thể phần mở rộng:
  - Xử lý dữ liệu định lượng liên tục (continuous data) và dữ liệu định tính
  - Xử lý khi thiếu giá trị thuộc tính
  - Xử lý với các thuộc tính với các chi phí khác nhau
  - Kỹ thuật phân cành

Mã giả của thuật toán

- Kiểm tra những trường hợp cơ bản
- Với mỗi thuộc tính A, tìm Information Gain nhờ việc tách thuộc tính A
- Chọn A_best là thuộc tính mà Information Gain hoặc GainRatio lớn nhất
- Dùng A_best là thuộc tính cho node chia cắt cây
- Đệ quy các danh sách phụ được tạo ra bởi việc phân chia theo A_best và thêm các node này như là con của node

- ENTROPY (Độ hỗn loạn)
  - $Entropy(P) = -\sum_{i=1}^{n} p_i \times \log_2(p_i)$

Với: p_i là xác suất xuất hiện đối tượng dữ liệu mang thuộc tính i của bộ dữ liệu P

- INFORMATION GAIN (Lượng thông tin có được)
  - $Gain(p, T) = Entropy(p) - \sum_{i=1}^{n}(p_j \times Entropy(p_j))$

Với: p_j là tập giá trị có thể cho thuộc tính T, p là bộ dữ liệu

Ví dụ: Một bộ dữ liệu có 1 xanh lá, 2 tím, 3 đỏ: 🟢🟣🟣🔴🔴🔴

Ta áp dụng công thức:

- $Entropy(P) = -\sum_{i=1}^{n} p_i \times \log_2(p_i)$

- $Entropy(P) = -(p_{xanh} \times \log_2 p_{xanh} + p_{tím} \times \log_2 p_{tím} + p_{đỏ} \times \log_2 p_{đỏ})$

- $Entropy(P) = -(\frac{1}{6} \times \log_2\left(\frac{1}{6}\right) + \frac{2}{6} \times \log_2\left(\frac{2}{6}\right) + \frac{3}{6} \times \log_2(\frac{3}{6}))$

$= 1.46$

Trường hợp bộ dữ liệu chỉ có một màu, ví dụ bộ dữ liệu chỉ có màu xanh thì khi đó:

- $Entropy(P) = -(1 \times \log_2(1)) = 0$

Trước khi chia bộ dữ liệu:

$$E_{ban\ đầu} = -(0.5\ \log_2 0.5 + 0.5 \log_2 0.5) = 1$$

Sau khi chia bộ dữ liệu:

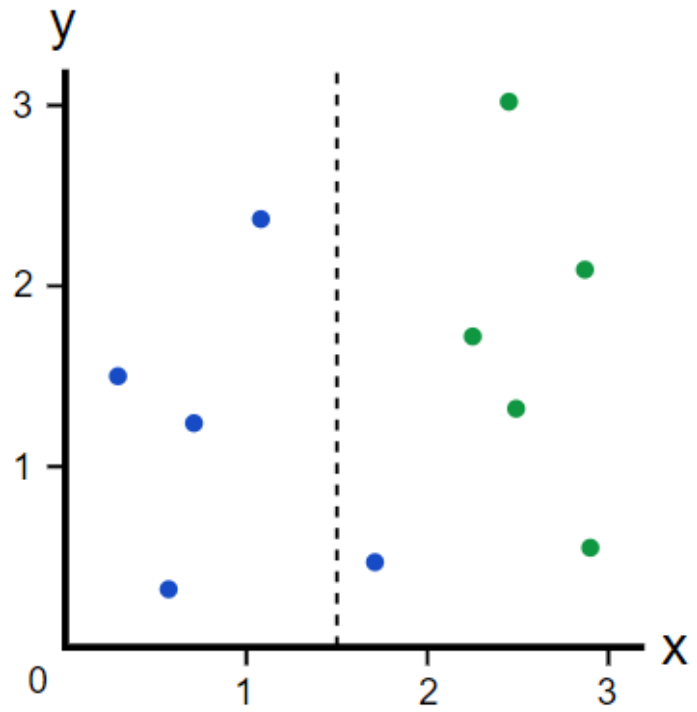
$$E_{trái} = -(1\ \times \log_2 1) = 0$$

$$E_{phải} = -\left(\frac{1}{6}\log_2\left(\frac{1}{6}\right) + \frac{5}{6}\log_2\left(\frac{5}{6}\right)\right) = 0.65$$

Ta áp dụng công thức:

- $Gain(p, T) = Entropy(p) - \sum_{i=1}^{n}(p_j \times Entropy(p_j))$

$$Gain(S, nhánh) = Entropy(S) - (p_{trái} \times E_{trái} + p_{phải} \times E_{phải})$$
$$= 1\ - (0.4 \times 0 + 0.6\ \times 0.65) = 0.61$$



Bộ dữ liệu S

Lượng thông tin có được ~ độ hỗn loạn mất đi

- # GAIN RATIO (Tỉ lệ đạt được)

$$GainRatio(p,T) = \frac{Gain(p,T)}{SplitInfo(p,T)}$$

Với SplitInfo(p,T):

$$SplitInfo(p,T) = -\sum P'\left(\frac{j}{p}\right) \times \log_2\left(P'\left(\frac{j}{p}\right)\right)$$

Với $P'\frac{j}{p}$ là tỷ lệ các thành phần ở vị trí p hiện tại

Ngày	Quang cảnh	Nhiệt độ	Độ ẩm	Độ gió	Quyết định
1	Nắng	Nóng	85	Yếu	Không
2	Nắng	Nóng	90	Mạnh	Không
3	Âm u	Nóng	78	Yếu	Có
4	Mưa	Ấm	96	Yếu	Có
5	Mưa	Lạnh	80	Yếu	Có
6	Mưa	Lạnh	70	Mạnh	Không
7	Âm u	Lạnh	65	Mạnh	Có
8	Nắng	Ấm	95	Yếu	Không
9	Nắng	Lạnh	70	Yếu	Có
10	Mưa	Ấm	80	Yếu	Có
11	Nắng	Ấm	70	Mạnh	Có
12	Âm u	Ấm	90	Mạnh	Có
13	Âm u	Nóng	75	Yếu	Có
14	Mưa	Ấm	80	Mạnh	Không

Bộ dữ liệu S

- **Ví dụ:** Câu hỏi có chơi tennis hay không? Quyết định đưa ra dựa trên các yếu tố về thời tiết: Quang cảnh, nhiệt độ, độ ẩm, độ gió.

Ta có tập các giá trị của thuộc tính sau:

Quang cảnh ={Nắng, Âm u, Mưa}

Nhiệt độ = {Nóng, Ấm, Lạnh}

Độ ẩm= {65,70,75,78,80,85,90,95,96}

Độ gió = {Yếu, Mạnh}

Bước 1: Tính Entropy của bộ dữ liệu S

$$Entropy(S) = -\sum_{i=1}^{n} p_i \times \log_2(p_i)$$

$$= -\left(\left(\frac{5}{14}\right) \times \log_2\left(\frac{5}{14}\right) + \left(\frac{9}{14}\right) \times \log_2\left(\frac{9}{14}\right)\right) = 0.94$$

Bước 2: Tính Gain Ratio của từng thuộc tính:

- Thuộc tính Quang cảnh = {Nắng, Âm u, Mưa}:

$$Entropy(S_{nắng}) = -\left(\left(\frac{2}{5}\right) \times \log_2\left(\frac{2}{5}\right) + \left(\frac{3}{5}\right) \times \log_2\left(\frac{3}{5}\right)\right) = 0.9710$$

$$Entropy(S_{mưa}) = -\left(\left(\frac{3}{5}\right) \times \log_2\left(\frac{3}{5}\right) + \left(\frac{2}{5}\right) \times \log_2\left(\frac{2}{5}\right)\right) = 0.9710$$

$$Entropy(S_{âm\ u}) = -\left(\left(\frac{4}{4}\right) \times \log_2\left(\frac{4}{4}\right) + (0) \times \log_2(0)\right) = 0$$

$$\boldsymbol{Gain}(\boldsymbol{S}, \textbf{Quang cảnh})$$

$$= Entropy(S) - \left(\frac{5}{14} \times Entropy(S_{nắng}) + \frac{5}{14} \times Entropy(S_{mưa}) + \frac{4}{14} \times Entropy(S_{âm\ u})\right)$$

$$= 0.94 - (\frac{5}{14} \times 0.9710 + \frac{5}{14} \times 0.9710 + \frac{4}{14} \times 0) = 0.246$$

$$\boldsymbol{SplitInfo}(\boldsymbol{S}, \textbf{quang cảnh}) = -\left(\frac{5}{14} \times \log_2\left(\frac{5}{14}\right) + \frac{5}{14} \times \log_2\left(\frac{5}{14}\right) + \frac{4}{14}\log_2\left(\frac{4}{14}\right)\right) = 1.577$$

$$\boldsymbol{GainRatio}(\boldsymbol{S}, \textbf{quang cảnh}) = \frac{Gain(S, quang\ cảnh)}{SplitInfo(S, quang\ cảnh)} = \frac{0.246}{1.577} = 0.156$$

➢ Thuộc tính Nhiệt độ = {Nóng, Ấm, Lạnh}

$$Entropy(S_{nóng}) = -\left(\left(\frac{2}{4}\right) \times \log_2\left(\frac{2}{4}\right) + \left(\frac{2}{4}\right) \times \log_2\left(\frac{2}{4}\right)\right) = 1$$

$$Entropy(S_{ấm}) = -\left(\left(\frac{4}{6}\right) \times \log_2\left(\frac{4}{6}\right) + \left(\frac{2}{6}\right) \times \log_2\left(\frac{2}{6}\right)\right) = 0.918$$

$$Entropy(S_{lạnh}) = -\left(\left(\frac{3}{4}\right) \times \log_2\left(\frac{3}{4}\right) + \left(\frac{1}{4}\right) \times \log_2\left(\frac{1}{4}\right)\right) = 0.811$$

$$\boldsymbol{Gain}(\boldsymbol{S}, \textbf{Nhiệt độ})$$
$$= Entropy(S) - \left(\frac{4}{14} \times Entropy(S_{nóng}) + \frac{6}{14} \times Entropy(S_{ấm}) + \frac{4}{14} \times Entropy(S_{lạnh})\right)$$
$$= 0.94 - (\frac{4}{14} \times 1 + \frac{6}{14} \times 0.918 + \frac{4}{14} \times 0.811) = 0.029$$

$$\boldsymbol{SplitInfo}(\boldsymbol{S}, \textbf{nhiệt độ}) = -\left(\frac{4}{14} \times \log_2\left(\frac{4}{14}\right) + \frac{6}{14} \times \log_2\left(\frac{6}{14}\right) + \frac{4}{14}\log_2\left(\frac{4}{14}\right)\right) = 1.5566$$

$$\boldsymbol{GainRatio}(\boldsymbol{S}, \textbf{nhiệt độ}) = \frac{Gain(S, \text{nhiệt độ})}{SplitInfo(S, \text{nhiệt độ})} = \frac{0.029}{1.5566} = 0.0186$$

➢ Thuộc tính Độ gió = {Yếu, Mạnh}:

$$Entropy(S_{yếu}) = -\left(\left(\frac{6}{8}\right) \times \log_2\left(\frac{6}{8}\right) + \left(\frac{2}{8}\right) \times \log_2\left(\frac{2}{8}\right)\right) = 0.811$$

$$Entropy(S_{mạnh}) = -\left(\left(\frac{3}{6}\right) \times \log_2\left(\frac{3}{6}\right) + \left(\frac{3}{6}\right) \times \log_2\left(\frac{3}{6}\right)\right) = 1$$

$$\boldsymbol{Gain(S, \text{Độ gió})}$$

$$= Entropy(S) - \left(\frac{5}{14} \times Entropy(S_{nắng}) + \frac{5}{14} \times Entropy(S_{mưa}) + \frac{4}{14} \times Entropy(S_{âm\ u})\right)$$

$$= 0.94 - (\frac{8}{14} \times 0.811 + \frac{6}{14} \times 1) = 0.048$$

$$\boldsymbol{SplitInfo(S, \text{độ gió})} = -\left(\frac{8}{14} \times \log_2\left(\frac{8}{14}\right) + \frac{6}{14} \times \log_2\left(\frac{6}{14}\right)\right) = 0.985$$

$$\boldsymbol{GainRatio(S, \text{độ gió})} = \frac{Gain(S, \textbf{độ gió})}{SplitInfo(S, \textbf{độ gió})} = \frac{0.048}{0.985} = 0.0487$$

➢ Thuộc tính Độ ẩm= {65,70,75,78,80,85,90,95,96}:

Xét ở giá trị ngưỡng 65:

$$Entropy(S_{\text{độ ẩm}\leq 65}) = -\left(\left(\frac{1}{1}\right)\times \log_2\left(\frac{1}{1}\right) + \left(\frac{0}{1}\right)\times \log_2\left(\frac{0}{1}\right)\right) = 0$$

$$Entropy(S_{\text{độ ẩm}>65}) = -\left(\left(\frac{8}{13}\right)\times \log_2\left(\frac{8}{13}\right) + \left(\frac{5}{13}\right)\times \log_2\left(\frac{5}{13}\right)\right) = 0.96$$

$$\boldsymbol{Gain(S, \text{Độ ẩm} <> 65)}$$

$$= Entropy(S) - \left(\frac{1}{14}\times Entropy(S_{\text{độ ẩm}\leq 65}) + \frac{13}{14}\times Entropy(S_{\text{độ ẩm}>65})\right)$$

$$= 0.94 - (\frac{1}{14}\times 0 + \frac{13}{14}\times 0.96) = 0.049$$

$$\boldsymbol{SplitInfo(S, \text{Độ ẩm} <> 65)} = -\left(\frac{1}{14}\times \log_2\left(\frac{1}{14}\right) + \frac{13}{14}\times \log_2\left(\frac{13}{14}\right)\right) = 0.371$$

$$\boldsymbol{GainRatio(S, \text{Độ ẩm} <> 65)} = \frac{Gain(S, \text{Độ ẩm} <> 65)}{SplitInfo(S, \text{Độ ẩm} <> 65)} = \frac{0.049}{0.371} = 0.13$$

Ngày	Độ ẩm	Quyết định
7	65	Có
6	70	Không
9	70	Có
11	70	Có
13	75	Có
3	78	Có
5	80	Có
10	80	Có
14	80	Không
1	85	Không
2	90	Không
12	90	Có
8	95	Không
4	96	Có

Ngày	Độ ẩm	Quyết định
7	65	Có
6	70	Không
9	70	Có
11	70	Có
13	75	Có
3	78	Có
5	80	Có
10	80	Có
14	80	Không
1	85	Không
2	90	Không
12	90	Có
8	95	Không
4	96	Có

- Thuộc tính Độ ẩm= {65,70,75,78,80,85,90,95,96}:

Xét ở giá trị ngưỡng 70:

$$Entropy(S_{\text{độ ẩm}\leq 70}) = -\left(\left(\frac{3}{4}\right)\times \log_2\left(\frac{3}{4}\right) + \left(\frac{1}{4}\right)\times \log_2\left(\frac{1}{4}\right)\right) = 0.811$$

$$Entropy(S_{\text{độ ẩm}>70}) = -\left(\left(\frac{6}{10}\right)\times \log_2\left(\frac{6}{10}\right) + \left(\frac{4}{10}\right)\times \log_2\left(\frac{4}{10}\right)\right) = 0.97$$

$$\boldsymbol{Gain(S, \text{Độ ẩm} <> 70)}$$

$$= Entropy(S) - \left(\frac{4}{14}\times Entropy(S_{\text{độ ẩm}\leq 70}) + \frac{10}{14}\times Entropy(S_{\text{độ ẩm}>70})\right)$$

$$= 0.94 - (\frac{4}{14}\times 0.811 + \frac{10}{14}\times 0.97) = 0.0154$$

$$\boldsymbol{SplitInfo(S, \text{Độ ẩm} <> 70)} = -\left(\frac{4}{14}\times \log_2\left(\frac{4}{14}\right) + \frac{10}{14}\times \log_2\left(\frac{10}{14}\right)\right) = 0.863$$

$$\boldsymbol{GainRatio(S, \text{Độ ẩm} <> 70)} = \frac{Gain(S, \text{Độ ẩm} <> 70)}{SplitInfo(S, \text{Độ ẩm} <> 70)} = \frac{0.0154}{0.863}$$

$$= 0.0178$$

➢ Thuộc tính Độ ẩm= {65,70,75,78,80,85,90,95,96}:

Xét ở giá trị ngưỡng 75:

$$Entropy(S_{\text{độ ẩm}\leq 75}) = -\left(\left(\frac{4}{5}\right)\times \log_2\left(\frac{4}{5}\right) + \left(\frac{1}{5}\right)\times \log_2\left(\frac{1}{5}\right)\right) = 0.722$$

$$Entropy(S_{\text{độ ẩm}>75}) = -\left(\left(\frac{5}{9}\right)\times \log_2\left(\frac{5}{9}\right) + \left(\frac{4}{9}\right)\times \log_2\left(\frac{4}{9}\right)\right) = 0.991$$

$$\boldsymbol{Gain(S, \text{Độ ẩm} <> 75)}$$

$$= Entropy(S) - \left(\frac{5}{14}\times Entropy(S_{\text{độ ẩm}\leq 75}) + \frac{9}{14}\times Entropy(S_{\text{độ ẩm}>75})\right)$$

$$= 0.94 - (\frac{5}{14}\times 0.722 + \frac{9}{14}\times 0.991) = 0.045$$

$$\boldsymbol{SplitInfo(S, \text{Độ ẩm} <> 75)} = -\left(\frac{5}{14}\times \log_2\left(\frac{5}{14}\right) + \frac{9}{14}\times \log_2\left(\frac{9}{14}\right)\right) = 0.94$$

$$\boldsymbol{GainRatio(S, \text{Độ ẩm} <> 75)} = \frac{Gain(S, \text{Độ ẩm} <> 75)}{SplitInfo(S, \text{Độ ẩm} <> 75)} = \frac{0.045}{0.94} = 0.0479$$

Ngày	Độ ẩm	Quyết định
7	65	Có
6	70	Không
9	70	Có
11	70	Có
13	75	Có
3	78	Có
5	80	Có
10	80	Có
14	80	Không
1	85	Không
2	90	Không
12	90	Có
8	95	Không
4	96	Có

- Thuộc tính Độ ẩm= {65,70,75,78,80,85,90,95,96}:

Xét ở giá trị ngưỡng 78:

$$Entropy(S_{\text{độ ẩm}\leq 78}) = -\left(\left(\frac{5}{6}\right)\times \log_2\left(\frac{5}{6}\right) + \left(\frac{1}{6}\right)\times \log_2\left(\frac{1}{6}\right)\right) = 0.65$$

$$Entropy(S_{\text{độ ẩm}>78}) = -\left(\left(\frac{4}{8}\right)\times \log_2\left(\frac{4}{8}\right) + \left(\frac{4}{8}\right)\times \log_2\left(\frac{4}{8}\right)\right) = 1$$

$$\boldsymbol{Gain(S, \text{Độ ẩm} <> 78)}$$

$$= Entropy(S) - \left(\frac{6}{14}\times Entropy(S_{\text{độ ẩm}\leq 78}) + \frac{8}{14}\times Entropy(S_{\text{độ ẩm}>78})\right)$$

$$= 0.94 - \left(\frac{6}{14}\times 0.65 + \frac{8}{14}\times 1\right) = 0.09$$

$$\boldsymbol{SplitInfo(S, \text{Độ ẩm} <> 78)} = -\left(\frac{6}{14}\times \log_2\left(\frac{6}{14}\right) + \frac{8}{14}\times \log_2\left(\frac{8}{14}\right)\right) = 0.985$$

$$\boldsymbol{GainRatio(S, \text{Độ ẩm} <> 78)} = \frac{Gain(S, \text{Độ ẩm} <> 78)}{SplitInfo(S, \text{Độ ẩm} <> 78)} = \frac{0.09}{0.985} = 0.091$$

Ngày	Độ ẩm	Quyết định
7	65	Có
6	70	Không
9	70	Có
11	70	Có
13	75	Có
3	78	Có
5	80	Có
10	80	Có
14	80	Không
1	85	Không
2	90	Không
12	90	Có
8	95	Không
4	96	Có

- Thuộc tính Độ ẩm= {65,70,75,78,80,85,90,95,96}:

Xét ở giá trị ngưỡng 80:

$$Entropy(S_{\text{độ ẩm}\leq 80}) = -\left(\left(\frac{7}{9}\right)\times \log_2\left(\frac{7}{9}\right) + \left(\frac{2}{9}\right)\times \log_2\left(\frac{2}{9}\right)\right) = 0.764$$

$$Entropy(S_{\text{độ ẩm}>80}) = -\left(\left(\frac{2}{5}\right)\times \log_2\left(\frac{2}{5}\right) + \left(\frac{3}{5}\right)\times \log_2\left(\frac{3}{5}\right)\right) = 0.97$$

$$\boldsymbol{Gain(S, \text{Độ ẩm} <> 80)}$$

$$= Entropy(S) - \left(\frac{9}{14}\times Entropy(S_{\text{độ ẩm}\leq 80}) + \frac{5}{14}\times Entropy(S_{\text{độ ẩm}>80})\right)$$

$$= 0.94 - (\frac{9}{14}\times 0.764 + \frac{5}{14}\times 0.97) = 0.102$$

$$\boldsymbol{SplitInfo(S, \text{Độ ẩm} <> 80)} = -\left(\frac{9}{14}\times \log_2\left(\frac{9}{14}\right) + \frac{5}{14}\times \log_2\left(\frac{5}{14}\right)\right) = 0.94$$

$$\boldsymbol{GainRatio(S, \text{Độ ẩm} <> 80)} = \frac{Gain(S, \text{Độ ẩm} <> 80)}{SplitInfo(S, \text{Độ ẩm} <> 80)} = \frac{0.102}{0.94} = 0.1085$$

Ngày	Độ ẩm	Quyết định
7	65	Có
6	70	Không
9	70	Có
11	70	Có
13	75	Có
3	78	Có
5	80	Có
10	80	Có
14	80	Không
1	85	Không
2	90	Không
12	90	Có
8	95	Không
4	96	Có

➢ Thuộc tính Độ ẩm= {65,70,75,78,80,85,90,95,96}:

Xét ở giá trị ngưỡng 85:

$$Entropy(S_{\text{độ ẩm}\leq 85}) = -\left(\left(\frac{7}{10}\right)\times \log_2\left(\frac{7}{10}\right) + \left(\frac{3}{10}\right)\times \log_2\left(\frac{3}{10}\right)\right) = 0.881$$

$$Entropy(S_{\text{độ ẩm}>85}) = -\left(\left(\frac{2}{4}\right)\times \log_2\left(\frac{2}{4}\right) + \left(\frac{2}{4}\right)\times \log_2\left(\frac{2}{4}\right)\right) = 1$$

$$\boldsymbol{Gain(S, \text{Độ ẩm} <> 85)}$$

$$= Entropy(S) - \left(\frac{5}{14}\times Entropy(S_{\text{độ ẩm}\leq 85}) + \frac{9}{14}\times Entropy(S_{\text{độ ẩm}>85})\right)$$

$$= 0.94 - (\frac{10}{14}\times 0.881 + \frac{4}{14}\times 1) = 0.025$$

$$\boldsymbol{SplitInfo(S, \text{Độ ẩm} <> 85)} = -\left(\frac{10}{14}\times \log_2\left(\frac{10}{14}\right) + \frac{4}{14}\times \log_2\left(\frac{4}{14}\right)\right) = 0.863$$

$$\boldsymbol{GainRatio(S, \text{Độ ẩm} <> 85)} = \frac{Gain(S, \text{Độ ẩm} <> 85)}{SplitInfo(S, \text{Độ ẩm} <> 85)} = \frac{0.025}{0.863} = 0.029$$

Ngày	Độ ẩm	Quyết định
7	65	Có
6	70	Không
9	70	Có
11	70	Có
13	75	Có
3	78	Có
5	80	Có
10	80	Có
14	80	Không
1	85	Không
2	90	Không
12	90	Có
8	95	Không
4	96	Có

- Thuộc tính Độ ẩm= {65,70,75,78,80,85,90,95,96}:

Xét ở giá trị ngưỡng 90:

$$Entropy(S_{\text{độ ẩm}\leq 90}) = -\left(\left(\frac{8}{12}\right)\times \log_2\left(\frac{8}{12}\right) + \left(\frac{4}{12}\right)\times \log_2\left(\frac{4}{12}\right)\right) = 0.918$$

$$Entropy(S_{\text{độ ẩm}>90}) = -\left(\left(\frac{1}{2}\right)\times \log_2\left(\frac{1}{2}\right) + \left(\frac{1}{2}\right)\times \log_2\left(\frac{1}{2}\right)\right) = 1$$

$$\boldsymbol{Gain(S, \text{Độ ẩm} <> 90)}$$

$$= Entropy(S) - \left(\frac{12}{14}\times Entropy(S_{\text{độ ẩm}\leq 90}) + \frac{2}{14}\times Entropy(S_{\text{độ ẩm}>90})\right)$$

$$= 0.94 - (\frac{12}{14}\times 0.918 + \frac{2}{14}\times 1) = 0.01$$

$$\boldsymbol{SplitInfo(S, \text{Độ ẩm} <> 90)} = -\left(\frac{12}{14}\times \log_2\left(\frac{12}{14}\right) + \frac{2}{14}\times \log_2\left(\frac{2}{14}\right)\right) = 0.592$$

$$\boldsymbol{GainRatio(S, \text{Độ ẩm} <> 90)} = \frac{Gain(S, \text{Độ ẩm} <> 90)}{SplitInfo(S, \text{Độ ẩm} <> 90)} = \frac{0.01}{0.592} = 0.017$$

Ngày	Độ ẩm	Quyết định
7	65	Có
6	70	Không
9	70	Có
11	70	Có
13	75	Có
3	78	Có
5	80	Có
10	80	Có
14	80	Không
1	85	Không
2	90	Không
12	90	Có
8	95	Không
4	96	Có

- Thuộc tính Độ ẩm= {65,70,75,78,80,85,90,95,96}:

Xét ở giá trị ngưỡng 95:

$$Entropy(S_{\text{độ ẩm}\leq 95}) = -\left(\left(\frac{8}{13}\right)\times \log_2\left(\frac{8}{13}\right) + \left(\frac{5}{13}\right)\times \log_2\left(\frac{5}{13}\right)\right) = 0.96$$

$$Entropy(S_{\text{độ ẩm}>95}) = -\left(\left(\frac{1}{1}\right)\times \log_2\left(\frac{1}{1}\right) + 0\right) = 0$$

$$\boldsymbol{Gain(S, \text{Độ ẩm} <> 95)}$$

$$= Entropy(S) - \left(\frac{13}{14}\times Entropy(S_{\text{độ ẩm}\leq 95}) + \frac{1}{14}\times Entropy(S_{\text{độ ẩm}>95})\right)$$

$$= 0.94 - (\frac{13}{14}\times 0.96 + \frac{1}{14}\times 0) = 0.049$$

$$\boldsymbol{SplitInfo(S, \text{Độ ẩm} <> 95)} = -\left(\frac{13}{14}\times \log_2\left(\frac{13}{14}\right) + \frac{1}{14}\times \log_2\left(\frac{1}{14}\right)\right) = 0.371$$

$$\boldsymbol{GainRatio(S, \text{Độ ẩm} <> 95)} = \frac{Gain(S, \text{Độ ẩm} <> 95)}{SplitInfo(S, \text{Độ ẩm} <> 95)} = \frac{0.049}{0.371} = 0.13$$

Ngày	Độ ẩm	Quyết định
7	65	Có
6	70	Không
9	70	Có
11	70	Có
13	75	Có
3	78	Có
5	80	Có
10	80	Có
14	80	Không
1	85	Không
2	90	Không
12	90	Có
8	95	Không
4	96	Có

- Thuộc tính Độ ẩm= {65,70,75,78,80,85,90,95,96}:

Xét ở giá trị ngưỡng 96:

$$Entropy(S_{\text{độ ẩm}\leq 96}) = -\left(\left(\frac{9}{14}\right)\times \log_2\left(\frac{9}{14}\right) + \left(\frac{5}{14}\right)\times \log_2\left(\frac{5}{14}\right)\right) = 0.94$$

$$Entropy(S_{\text{độ ẩm}>96}) = 0$$

$$\boldsymbol{Gain(S, \text{Độ ẩm} <> 96)}$$

$$= Entropy(S) - \left(\frac{14}{14}\times Entropy(S_{\text{độ ẩm}\leq 96}) + \frac{0}{14}\times Entropy(S_{\text{độ ẩm}>96})\right)$$

$$= 0.94 - (\frac{14}{14}\times 0.94 + 0) = 0$$

$$\boldsymbol{SplitInfo(S, \text{Độ ẩm} <> 96)} = -\left(\frac{14}{14}\times \log_2\left(\frac{14}{14}\right) + 0\right) = 0$$

$$\boldsymbol{GainRatio(S, \text{Độ ẩm} <> 96)} = \frac{Gain(S, \text{Độ ẩm} <> 96)}{SplitInfo(S, \text{Độ ẩm} <> 96)} = 0$$

Ngày	Độ ẩm	Quyết định
7	65	Có
6	70	Không
9	70	Có
11	70	Có
13	75	Có
3	78	Có
5	80	Có
10	80	Có
14	80	Không
1	85	Không
2	90	Không
12	90	Có
8	95	Không
4	96	Có

Giá trị ở ngưỡng	GainRatio
65	0.13
70	0.0178
75	0.0479
78	0.091
80	0.1085
85	0.029
90	0.017
95	0.13
96	0

Độ ẩm <> 65

Ngày	Độ ẩm	Quyết định
7	≤65	Có
6	> 65	Không
9	> 65	Có
11	> 65	Có
13	> 65	Có
3	> 65	Có
5	> 65	Có
10	> 65	Có
14	> 65	Không
1	> 65	Không
2	> 65	Không
12	> 65	Có
8	> 65	Không
4	> 65	Có

- Thuộc tính Độ ẩm= {65,70,75,78,80,85,90,95,96}:

$$Entropy(S_{\text{độ ẩm}\leq 65}) = -\left(\left(\frac{1}{1}\right)\times \log_2\left(\frac{1}{1}\right) + \left(\frac{0}{1}\right)\times \log_2\left(\frac{0}{1}\right)\right) = 0$$

$$Entropy(S_{\text{độ ẩm}>65}) = -\left(\left(\frac{8}{13}\right)\times \log_2\left(\frac{8}{13}\right) + \left(\frac{5}{13}\right)\times \log_2\left(\frac{5}{13}\right)\right) = 0.96$$

$$\boldsymbol{Gain(S, \text{Độ ẩm} <> 65)}$$

$$= Entropy(S) - \left(\frac{1}{14}\times Entropy(S_{\text{độ ẩm}\leq 65}) + \frac{13}{14}\times Entropy(S_{\text{độ ẩm}>65})\right)$$

$$= 0.94 - (\frac{1}{14}\times 0 + \frac{13}{14}\times 0.96) = 0.049$$

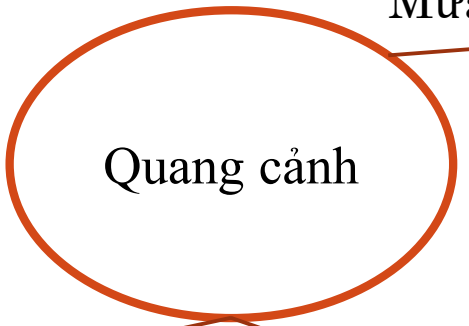
$$\boldsymbol{SplitInfo(S, \text{Độ ẩm} <> 65)} = -\left(\frac{1}{14}\times \log_2\left(\frac{1}{14}\right) + \frac{13}{14}\times \log_2\left(\frac{13}{14}\right)\right) = 0.371$$

$$\boldsymbol{GainRatio(S, \text{Độ ẩm} <> 65)} = \frac{Gain(S, \text{Độ ẩm} <> 65)}{SplitInfo(S, \text{Độ ẩm} <> 65)} = \frac{0.049}{0.371} = 0.13$$

Ngày	Độ ẩm	Quyết định
7	≤65	Có
6	> 65	Không
9	> 65	Có
11	> 65	Có
13	> 65	Có
3	> 65	Có
5	> 65	Có
10	> 65	Có
14	> 65	Không
1	> 65	Không
2	> 65	Không
12	> 65	Có
8	> 65	Không
4	> 65	Có

Thuộc tính	GainRatio
Quang cảnh	0.156
Nhiệt độ	0.0186
Độ ẩm	0.13
Độ gió	0.0487

Quang cảnh



Mưa

Ngày	Quang cảnh	Nhiệt độ	Độ ẩm	Độ gió	Quyết định
4	Mưa	Ấm	> 65	Yếu	Có
5	Mưa	Lạnh	> 65	Yếu	Có
6	Mưa	Lạnh	> 65	Mạnh	Không
10	Mưa	Ấm	> 65	Yếu	Có
14	Mưa	Ấm	> 65	Mạnh	Không

Nắng

Ngày	Quang cảnh	Nhiệt độ	Độ ẩm	Độ gió	Quyết định
1	Nắng	Nóng	> 65	Yếu	Không
2	Nắng	Nóng	> 65	Mạnh	Không
8	Nắng	Ấm	> 65	Yếu	Không
9	Nắng	Lạnh	> 65	Yếu	Có
11	Nắng	Ấm	> 65	Mạnh	Có

Âm u

Ngày	Quang cảnh	Nhiệt độ	Độ ẩm	Độ gió	Quyết định
3	Âm u	Nóng	> 65	Yếu	Có
7	Âm u	Lạnh	≤65	Mạnh	Có
12	Âm u	Ấm	> 65	Mạnh	Có
13	Âm u	Nóng	> 65	Yếu	Có

Bước 1: Tính Entropy của bảng $S_{\text{Quang cảnh = Nắng}}$:

$$Entropy(S_{\text{Quang cảnh = Nắng}}) = -\sum_{i=1}^{n} p_i \times \log_2(p_i)$$

$$= -\left(\left(\frac{2}{5}\right) \times \log_2\left(\frac{2}{5}\right) + \left(\frac{3}{5}\right) \times \log_2\left(\frac{3}{5}\right)\right) = 0.97$$

Ngày	Quang cảnh	Nhiệt độ	Độ ẩm	Độ gió	Quyết định
1	Nắng	Nóng	> 65	Yếu	Không
2	Nắng	Nóng	> 65	Mạnh	Không
8	Nắng	Ấm	> 65	Yếu	Không
9	Nắng	Lạnh	> 65	Yếu	Có
11	Nắng	Ấm	> 65	Mạnh	Có

Bước 2: Tính Gain Ratio của thuộc tính trong $S_{\text{Quang cảnh = Nắng}}$:

- Thuộc tính Nhiệt độ = {Nóng, Ấm, Lạnh}

$$Entropy(S_{nóng}) = -\left(\left(\frac{2}{2}\right) \times \log_2\left(\frac{2}{2}\right) + 0\right) = 0$$

$$Entropy(S_{lạnh}) = -\left(\left(\frac{1}{1}\right) \times \log_2\left(\frac{1}{1}\right) + 0\right) = 0$$

$$Entropy(S_{ấm}) = -\left(\left(\frac{1}{2}\right) \times \log_2\left(\frac{1}{2}\right) + \left(\frac{1}{2}\right) \times \log_2\left(\frac{1}{2}\right)\right) = 1$$

$$\boldsymbol{SplitInfo}(\boldsymbol{S}_{\textbf{Quang cảnh = Nắng}}, \textbf{nhiệt độ})$$

$$= -\left(\frac{2}{5} \times \log_2\left(\frac{2}{5}\right) + \frac{1}{5} \times \log_2\left(\frac{1}{5}\right) + \frac{2}{5}\log_2\left(\frac{2}{5}\right)\right)$$

$$= 1.522$$

$$\boldsymbol{Gain}(\boldsymbol{S}_{\textbf{Quang cảnh = Nắng}}, \textbf{Nhiệt độ})$$

$$= Entropy(S) - \left(\frac{2}{5} \times Entropy(S_{nóng}) + \frac{2}{5} \times Entropy(S_{ấm}) + \frac{1}{5} \times Entropy(S_{lạnh})\right) = 0.97 - (0 + \frac{2}{5} \times 1 + 0)$$

$$= 0.57$$

$$\boldsymbol{GainRatio}(\boldsymbol{S}_{\textbf{Quang cảnh = Nắng}}, \textbf{nhiệt độ}) = \frac{0.57}{1.522} = 0.375$$

➢ Thuộc tính Độ ẩm= {>65}:

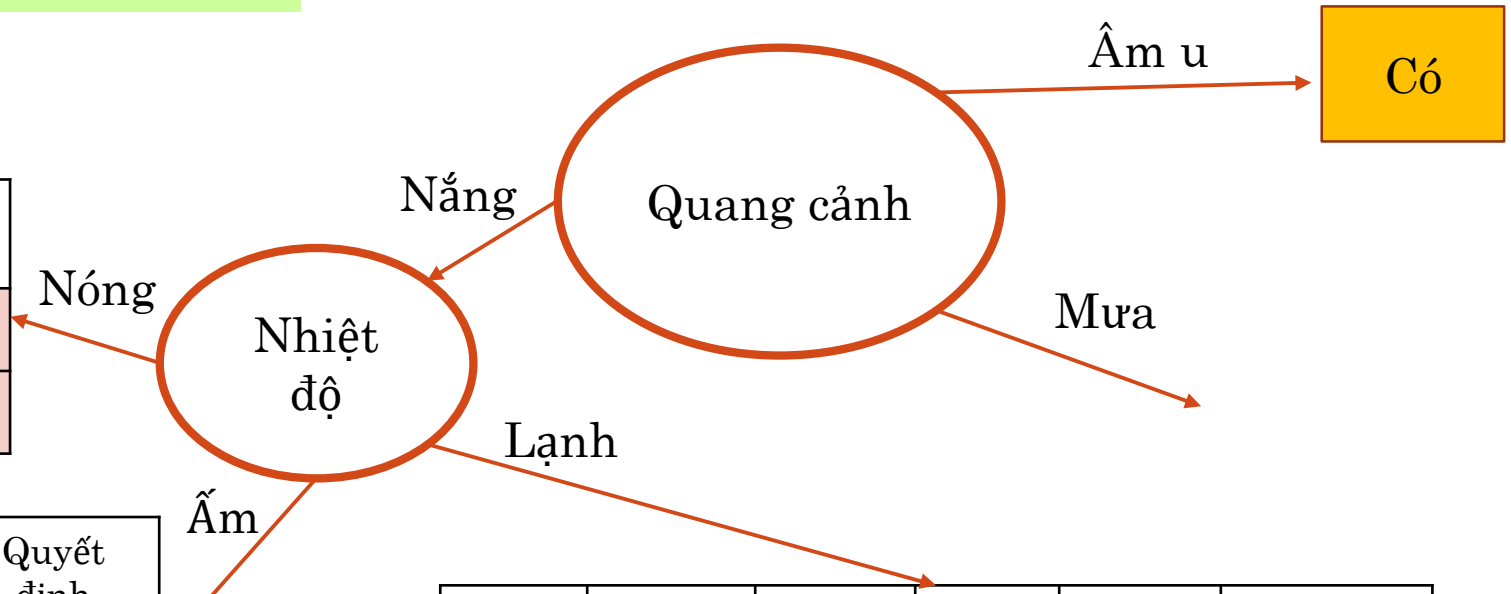
$$GainRatio\left(S_{\text{Quang cảnh = Nắng}}, \text{Độ ẩm} > 65\right) = 0$$

➢ Thuộc tính Độ gió = {Yếu, Mạnh}:

$$GainRatio(S_{\text{Quang cảnh = Nắng}}, \text{Độ gió}) = 0.0185$$

Ngày	Quang cảnh	Nhiệt độ	Độ ẩm	Độ gió	Quyết định
1	Nắng	Nóng	> 65	Yếu	Không
2	Nắng	Nóng	> 65	Mạnh	Không
8	Nắng	Ấm	> 65	Yếu	Không
9	Nắng	Lạnh	> 65	Yếu	Có
11	Nắng	Ấm	> 65	Mạnh	Có

→ Nhiệt độ



Ngày	Quang cảnh	Nhiệt độ	Độ ẩm	Độ gió	Quyết định
1	Nắng	Nóng	> 65	Yếu	Không
2	Nắng	Nóng	> 65	Mạnh	Không

Ngày	Quang cảnh	Nhiệt độ	Độ ẩm	Độ gió	Quyết định
8	Nắng	Ấm	> 65	Yếu	Không
11	Nắng	Ấm	> 65	Mạnh	Có

Ngày	Quang cảnh	Nhiệt độ	Độ ẩm	Độ gió	Quyết định
9	Nắng	Lạnh	> 65	Yếu	Có

Xét bảng $S_{\text{Quang cảnh = Nắng,Nhiệt độ=ấm}}$

Ngày	Quang cảnh	Nhiệt độ	Độ ẩm	Độ gió	Quyết định
8	Nắng	Ấm	> 65	Yếu	Không
11	Nắng	Ấm	> 65	Mạnh	Có

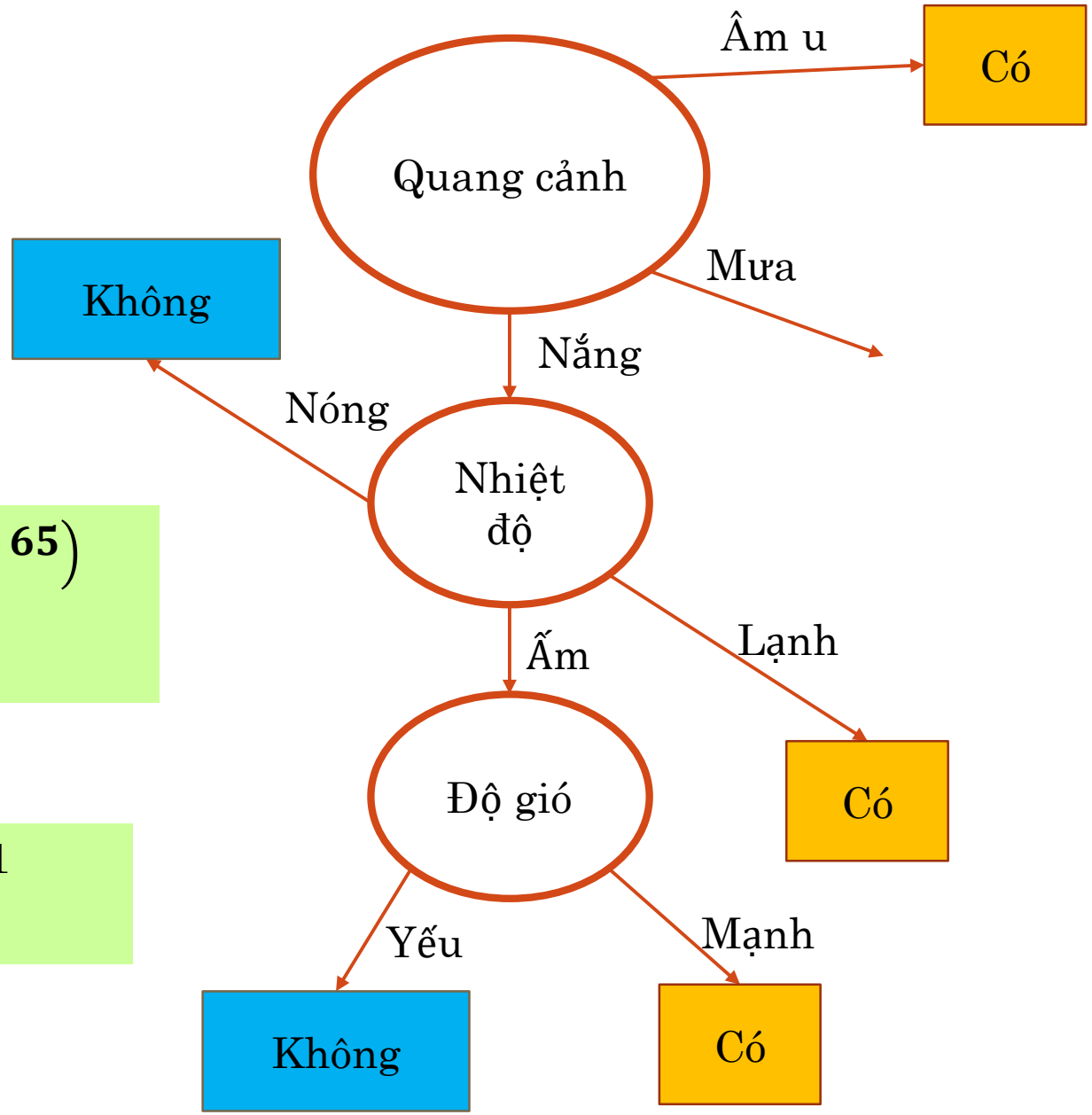
- Thuộc tính Độ ẩm= {>65}:

$$GainRatio\left(S_{\text{Quang cảnh = Nắng,Nhiệt độ=ấm}}, \text{Độ ẩm} > 65\right) = 0$$

- Thuộc tính Độ gió = {Yếu, Mạnh}:

$$GainRatio(S_{\text{Quang cảnh =Nắng,Nhiệt độ=ấm}}, \text{Độ gió}) = 1$$

→ Độ gió



Bước 1: Tính Entropy của bảng $S_{\text{Quang cảnh = Mưa}}$:

$$Entropy\left(S_{\text{Quang cảnh = Mưa}}\right) = 0.97$$

Bước 2: Tính Gain Ratio của thuộc tính trong $S_{\text{Quang cảnh = Mưa}}$:

- Thuộc tính Nhiệt độ = {Ấm, Lạnh}

$$\boldsymbol{GainRatio}\left(\boldsymbol{S_{\text{Quang cảnh = Mưa}}, \text{nhiệt độ}}\right) = 0.0186$$

- Thuộc tính Độ ẩm= {>65}:

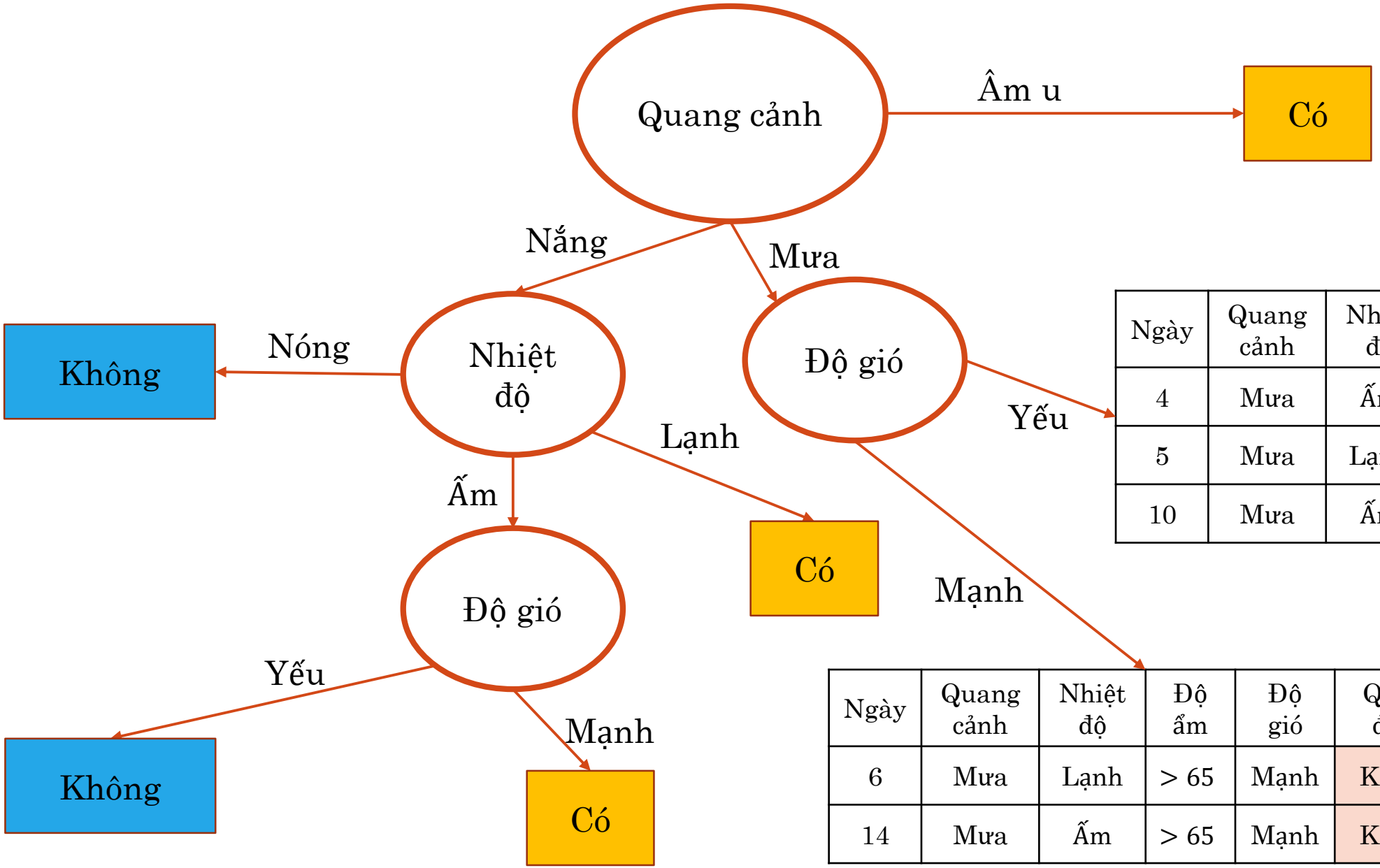
$$\boldsymbol{GainRatio}\left(\boldsymbol{S_{\text{Quang cảnh = Mưa}}, \text{độ ẩm>65}}\right) = 0$$

- Thuộc tính Độ gió = {Yếu, Mạnh}:

$$\boldsymbol{GainRatio}\left(\boldsymbol{S_{\text{Quang cảnh = Mưa}}, \text{độ gió}}\right) = 1$$

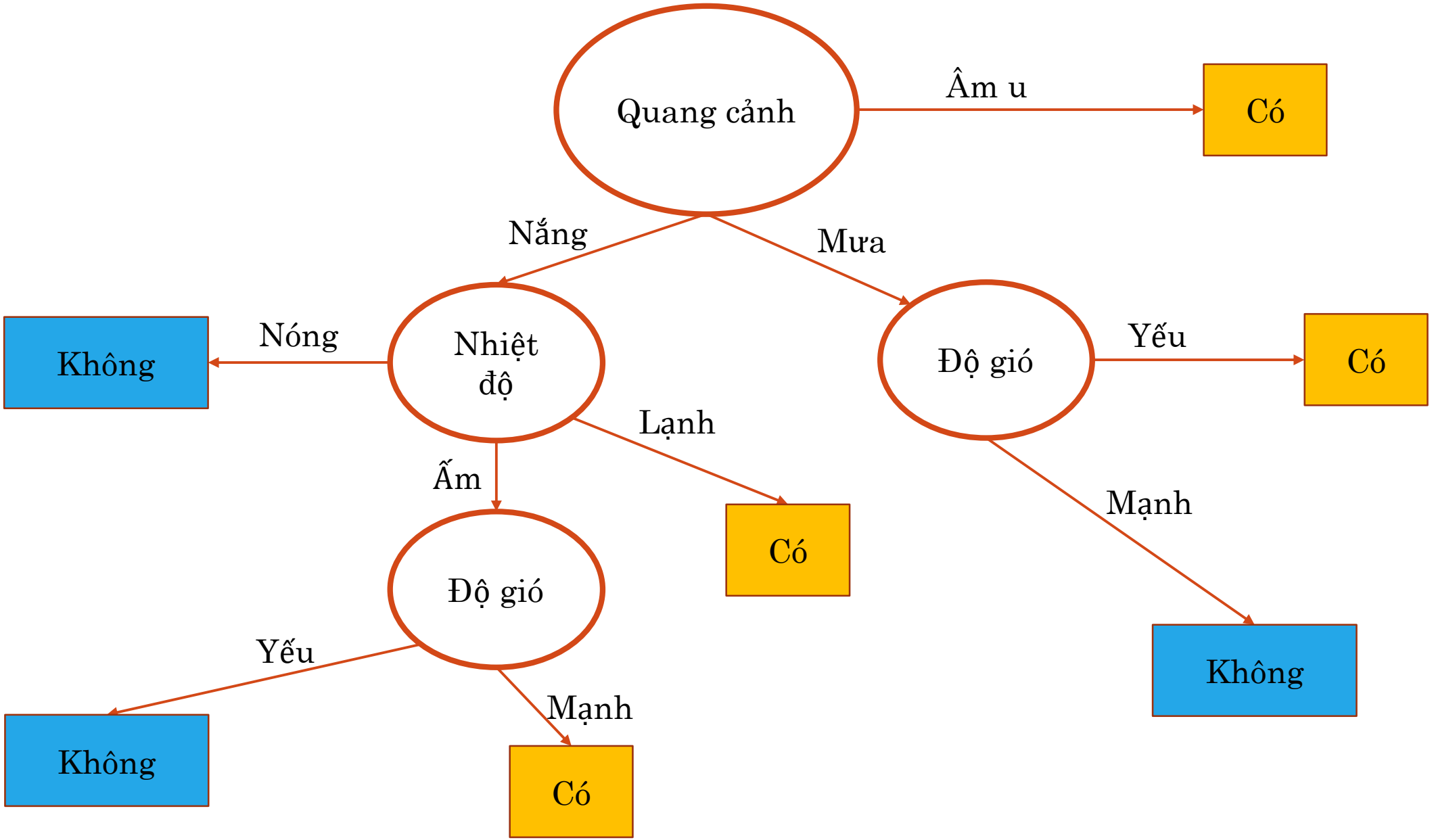
→ Độ gió

Ngày	Quang cảnh	Nhiệt độ	Độ ẩm	Độ gió	Quyết định
4	Mưa	Ấm	> 65	Yếu	Có
5	Mưa	Lạnh	> 65	Yếu	Có
6	Mưa	Lạnh	> 65	Mạnh	Không
10	Mưa	Ấm	> 65	Yếu	Có
14	Mưa	Ấm	> 65	Mạnh	Không



Ngày	Quang cảnh	Nhiệt độ	Độ ẩm	Độ gió	Quyết định
4	Mưa	Ấm	> 65	Yếu	Có
5	Mưa	Lạnh	> 65	Yếu	Có
10	Mưa	Ấm	> 65	Yếu	Có

Ngày	Quang cảnh	Nhiệt độ	Độ ẩm	Độ gió	Quyết định
6	Mưa	Lạnh	> 65	Mạnh	Không
14	Mưa	Ấm	> 65	Mạnh	Không



Cây quyết định hoàn chỉnh

Luật từ cây quyết định:

- Nếu quang cảnh âm u thì đánh tennis
- Nếu quang cảnh mưa và gió yếu thì đánh tennis
- Nếu quang cảnh mưa và gió mạnh thì không đánh tennis
- Nếu quang cảnh nắng và nhiệt độ nóng thì không đánh tennis
- Nếu quang cảnh nắng và nhiệt độ lạnh thì đánh tennis
- Nếu quang cảnh nắng và nhiệt độ ấm và độ gió mạnh thì đánh tennis
- Nếu quang cảnh nắng và nhiệt độ ấm và độ gió yếu thì không đánh tennis

CÁC TẬP KIỂU GIÁ TRỊ THUỘC TÍNH