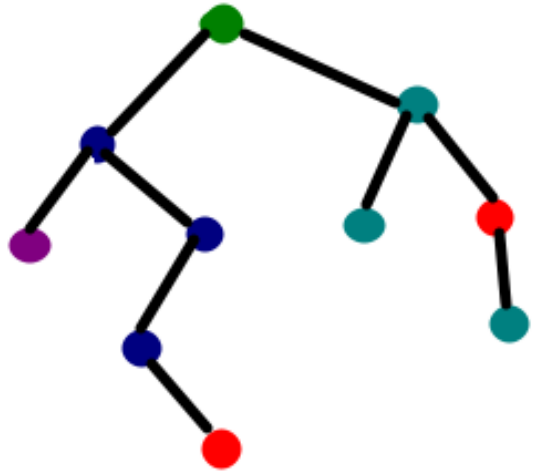
Decision Tree

Khoa Khoa học và Kỹ thuật thông tin
Bộ môn Khoa học dữ liệu

Cây quyết định

- Là mô hình máy học dự đoán câu trả lời bằng việc ra quyết định dựa trên các luật.
  + Các luật ở đây sẽ được biểu diễn bằng dạng cây.
- Các thành phần của 1 cây quyết định:
  + Nút không phải nút lá (non-leaf node).
  + Nút con (child node).
  + Nút gốc (root node).
  + Nút lá (leaf node / terminal node).
  + Đường đi (path)



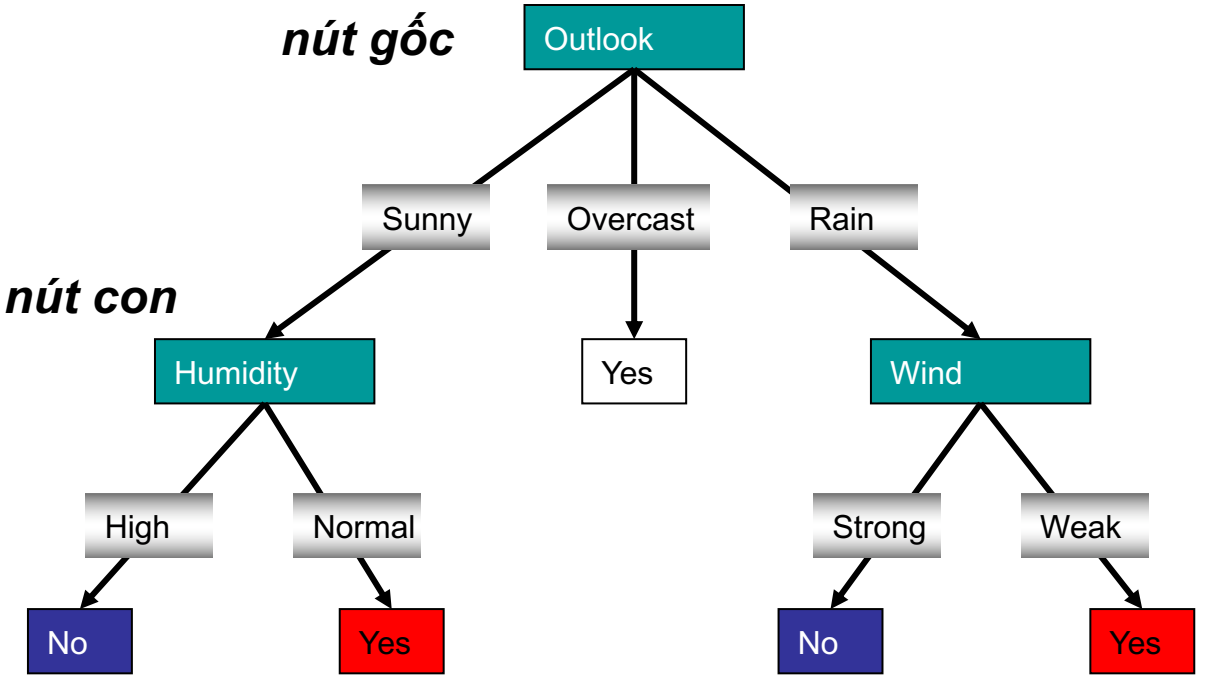
Ví dụ

David là quản lý của một câu lạc bộ đánh golf. Anh nhận thấy: Có ngày đông người muốn chơi golf nhưng số nhân viên câu lạc bộ lại không đủ phục vụ, Có hôm lại quá ít (hoặc không có) người đến chơi dẫn đến câu lạc bộ lại thừa nhân viên phục vụ, và việc này rõ ràng bị ảnh hưởng lớn từ yếu tố thời tiết.

➔ Có quy luật nào về thời tiết ảnh hưởng đến số lượng người đến câu lạc bộ golf hay không?

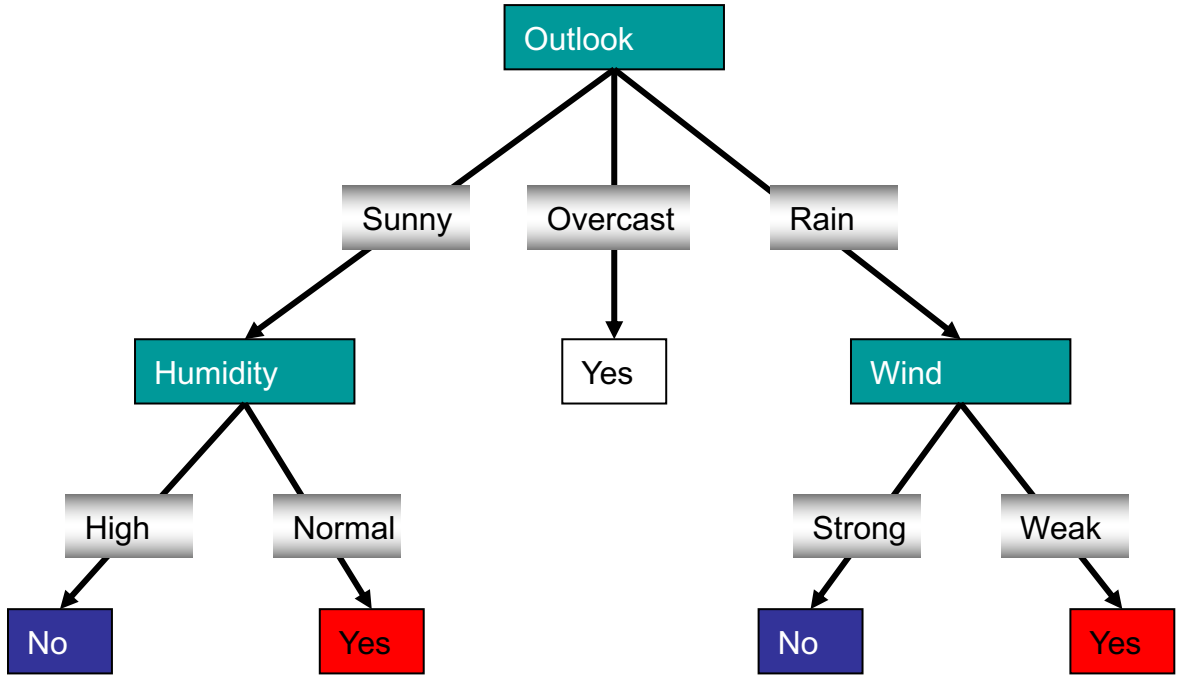
Biểu diễn cây quyết định

Day	Outlook	Temp.	Humidity	Wind	Play?
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
3	Overcast	Hot	High	Weak	Yes
4	Rain	Mild	High	Weak	Yes
5	Rain	Cool	Normal	Weak	No
6	Rain	Cool	Normal	Strong	Yes
7	Overcast	Cool	Normal	Weak	Yes
8	Sunny	Mild	High	Weak	Yes
9	Sunny	Cold	Normal	Weak	Yes
10	Rain	Mild	Normal	Strong	Yes
11	Sunny	Mild	Normal	Strong	Yes
12	Overcast	Mild	High	Strong	Yes
13	Overcast	Hot	Normal	Weak	Yes
14	Rain	Mild	High	Strong	No



(Outlook=Sunny ∧ Humidity=Normal)
∨ Outlook=Overcast
∨ (Outlook=Rain ∧ Wind=Weak) → YES

Biến đổi cây quyết định thành luật



$\wedge$ = AND = và
$\vee$ = OR = hoặc

R_1: If (Outlook=Sunny) $\wedge$ (Humidity=High) ➔ Play=No

R_2: If (Outlook=Sunny) $\wedge$ (Humidity=Normal) ➔ Play=Yes

R_3: If (Outlook=Overcast) ➔ Play=Yes

R_4: If (Outlook=Rain) $\wedge$ (Wind=Strong) ➔ Play=No

R_5: If (Outlook=Rain) $\wedge$ (Wind=Weak) ➔ Play=Yes

Xây dựng cây quyết định

1. Cây được thiết lập từ trên xuống dưới.
2. Rời rạc hóa các thuộc tính dạng phi số.
3. Các mẫu huấn luyện nằm ở gốc của cây.
4. **Chọn một thuộc tính để phân chia thành các nhánh.**
5. Tiếp tục lặp lại việc xây dựng cây quyết định cho các nhánh.
6. Điều kiện dừng:
   + Tất cả các mẫu rơi vào một nút thuộc về cùng một lớp (nút lá)
   + Không còn thuộc tính nào có thể dùng để phân chia mẫu nữa
   + Không còn lại mẫu nào tại nút

Lựa chọn thuộc tính

- Thuộc tính được chọn là thuộc tính **có lợi nhất** cho quá trình phân chia các giá trị về các lớp.
  + Mục tiêu: Cây quyết định **càng đơn giản càng tốt**.
- Tính toán độ lợi thông tin như thế nào?
- Có 2 độ đo thường dùng
  + Độ lợi thông tin (Information gain).
  + Chỉ số Gini (Gini index).

Thuật toán xây dựng Decision Tree

- Độ đo Gini:
  + Thuật toán CART (Classification And Regression Tree, Breiman et al., 1984)
- Độ đo Information Gain (IG):
  + Thuật toán ID3 (Interative Dichotomiser, R. Quilan, 1983).

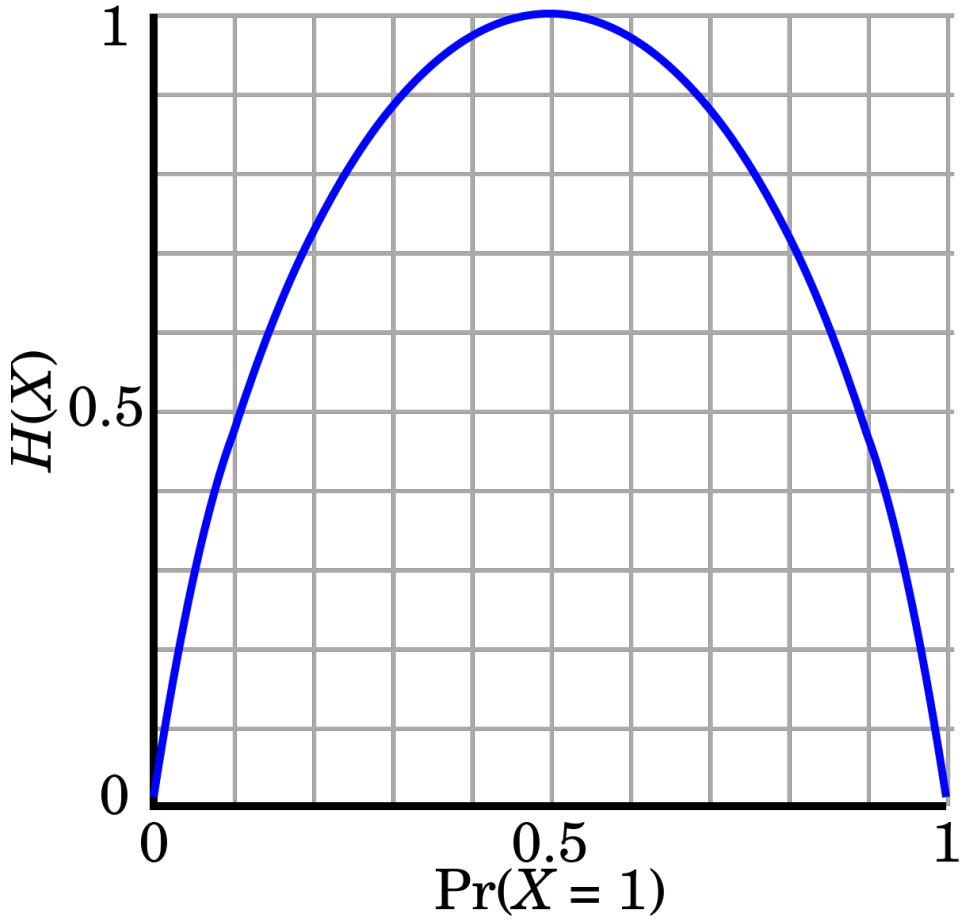
Độ lợi thông tin

- Độ lợi thông tin được xây dựng dựa trên khái niệm về entropy thông tin.
- Khái niệm **entropy** chỉ mức độ hỗn loạn của thông tin mang trong dữ liệu.
- Nếu một sự kiện **ngẫu nhiên rời rạc** x, có thể nhận các giá trị là 1..n, thì entropy của nó là:

$$H(\mathrm{x}) = -\sum_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{p}(\mathrm{i}) * \log_2(\mathrm{p}(\mathrm{i}))$$

- Trong đó: p(i) là xác suất xảy ra giá trị i.

Entropy thông tin



- Với p = 0 hoặc p = 1 thì H = 0 ➔ thông tin ít nhiễu loạn.
- Với p = 0.5 thì H = 1 ➔ thông tin nhiễu loạn.

➔ Khi cần ra quyết định, thì chọn thông tin nhiễu loạn hay ít nhiễu loạn ??

Hàm mất mát cho ID3

- Tính entropy cho một node S (gồm C class):

$$H(S) = -\sum_{i=1}^{C} \frac{N_c}{N} \log_2\left(\frac{N_c}{N}\right)$$

- Tính entropy thuộc tính x của node S. Mỗi dữ liệu trong node S được phân ra thành K node con m_1, m_2, ..., m_k theo thuộc tính x.

$$H(x, S) = \sum_{k=1}^{K} H(S_k) \frac{m_K}{N}$$

- Độ lợi thông tin được định nghĩa như sau: **G(x,s) = H(S) – H(x,S)**
- Thuộc tính x được chọn khi G(x,S) lớn nhất (độ lợi thông tin lớn nhất).

$$\boldsymbol{x = argmax_x G(x, S)}$$

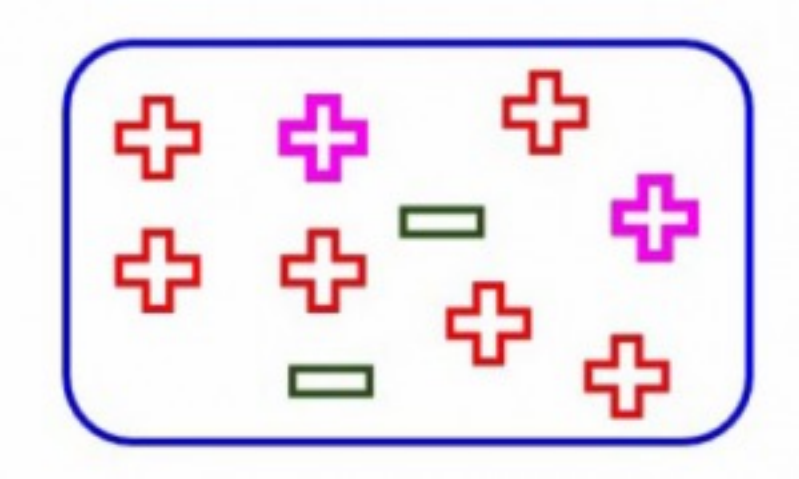
Độ đo Gini

- Độ đo Gini thể hiện mức độ phân loại sai khi chọn ngẫu nhiên 1 phần tử từ tập dữ liệu.
- Công thức tính độ đo Gini:

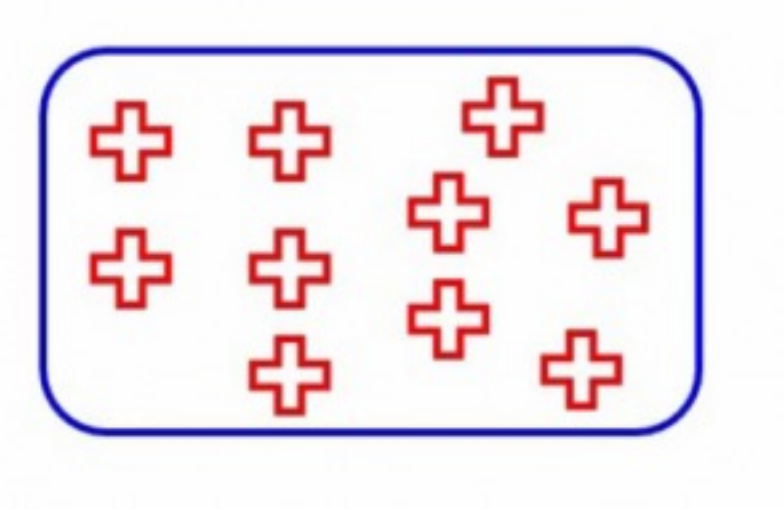
$$G(\mathrm{x}) = 1 - \sum_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{p}(\mathrm{i})^2$$

p(i) xác suất một phần tử ngẫu nhiên x thuộc lớp i.

Gini Impurity



Nếu lấy ngẫu nhiên 1 dữ liệu từ tập dữ liệu, xác suất lấy đúng dữ liêu thuộc lớp (+) là 0.8 ➔ có khoảng 20% lấy nhầm sang lớp (-)
➔ Impurity



Nếu lấy ngẫu nhiên 1 dữ liệu từ tập dữ liệu, xác suất lấy đúng dữ liêu thuộc lớp (+) là 1.
➔ Pured

Hàm mất mát cho CART

- Gini_split $= \sum_{\mathrm{i}=1}^{\mathrm{n}} \frac{k_i}{k} G(\mathrm{i})^2$
- k_i là số điểm dữ liệu trong ”child node” (node của nhánh được phân), k là số điểm dữ liệu của ”parent node” (node được dùng để phân nhánh).
- Gini_split càng nhỏ thì phân nhánh càng tối ưu

Các kỹ thuật giảm overfit

- Pruning (cắt tỉa): tạo ra một tập dữ liệu phát triển (validation), sau đó, đi ngược lên từ leaf-node và cắt tỉa các sibling node (giá trị) sao cho độ chính xác trên tập phát triển cải thiện hơn ➔ đang điều chỉnh lại tham số.
- Regularization: Cộng thêm một đại lượng λK vào hàm mất mát, với K là số lớp (hay là số nút lá).

$$H(x, S) = \sum_{k=1}^{K} H(S_k) \frac{m_K}{N} + \lambda K$$

Một số nhận xét

- Hai thuật toán như nhau trong đa số trường hợp.
- Gini thường chạy nhanh hơn nên được mặc định trong sklearn.
- Entropy thường cho cây cân bằng hơn.

Bài tập: Xây dựng cây quyết định

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

Xây dựng cây quyết định để tìm ra quy luật: với điều kiện thời tiết nào thì người chơi sẽ chơi golf ?

- *Xây dựng theo giải thuật ID3 và độ đo Information gain (IG)*
- *Xây dựng theo giải thuật CART và độ đo Gini*

Sử dụng chỉ số Information Gain (IG)

Tính toán cho thuộc tính Play

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

- Tổng số lượng dữ liệu: S = 14
- Số thuộc tính thuộc nhãn Play = yes: 9
- Số thuộc tính thuộc nhãn Play = no: 5

➔ $H(S) = -\left(\frac{9}{14}\log_2\left(\frac{9}{14}\right) + \frac{5}{14}\log_2\left(\frac{5}{14}\right)\right)$

$= 0.94$

Tính toán cho thuộc tính Outlook

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

	Yes	No	Total
Sunny	3	2	5
Overcast	4	0	4
Rainy	3	2	5

$$\mathrm{H(S, outlook)} = -\frac{5}{14}\left(\frac{3}{5}\log_2\left(\frac{3}{5}\right) + \frac{2}{5}\log_2\left(\frac{2}{5}\right)\right) - \frac{4}{14}\left(\frac{4}{4}\log_2\left(\frac{4}{4}\right)\right)$$
$$-\frac{5}{14}\left(\frac{3}{5}\log_2\left(\frac{3}{5}\right) + \frac{2}{5}\log_2\left(\frac{2}{5}\right)\right) = 0.69$$

$$G(S, outlook) = 0.94 - 0.69 = 0.25$$

Tính toán cho thuộc tính Temp

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

	Yes	No	Total
Hot	2	2	4
Mild	4	2	6
Cool	3	1	4

$$\mathrm{H}(\mathrm{S}, temp) = -\frac{4}{14}\left(\frac{2}{4}\log_2\left(\frac{2}{4}\right) + \frac{2}{4}\log_2\left(\frac{2}{4}\right)\right) - \frac{6}{14}\left(\frac{4}{6}\log_2\left(\frac{4}{6}\right) + \frac{2}{6}\log_2\left(\frac{2}{6}\right)\right) - \frac{4}{14}\left(\frac{3}{4}\log_2\left(\frac{3}{4}\right) + \frac{1}{4}\log_2\left(\frac{1}{4}\right)\right)$$

= 0.91

$$G(S, Temp) = 0.94 - 0.91 = 0.03$$

Tính toán cho thuộc tính Humidity

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

	Yes	No	Total
High	3	4	7
Normal	6	1	7

$$\mathrm{H}(S, humidity) = -\frac{7}{14}\left(\frac{3}{7}\log_2\left(\frac{3}{7}\right) + \frac{4}{7}\log_2\left(\frac{4}{7}\right)\right) - \frac{7}{14}\left(\frac{6}{7}\log_2\left(\frac{6}{7}\right) + \frac{1}{7}\log_2\left(\frac{1}{7}\right)\right)$$

= 0.78

$$G(S, humidity) = 0.94 - 0.78 = 0.16$$

Tính toán cho thuộc tính Wind

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

	Yes	No	Total
Weak	6	2	8
Strong	3	3	6

$$\mathrm{H}(S, Wind) = -\frac{8}{14}\left(\frac{6}{8}\log_2\left(\frac{6}{8}\right) + \frac{2}{8}\log_2\left(\frac{2}{8}\right)\right) - \frac{6}{14}\left(\frac{3}{6}\log_2\left(\frac{3}{6}\right) + \frac{3}{6}\log_2\left(\frac{3}{6}\right)\right)$$

= 0.89

$$G(S, Wind) = 0.94 - 0.89 = 0.05$$

Tổng hợp

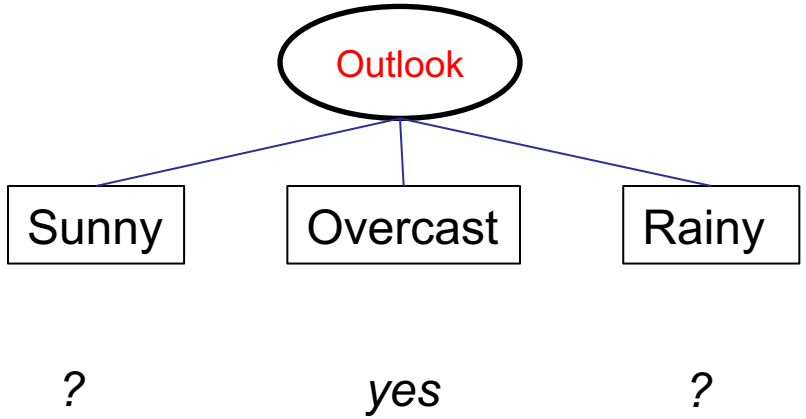
day	outlook	temperature	humidity	wind	play
1	sunny	hot	high	weak	no
2	sunny	hot	high	strong	no
3	overcast	hot	high	weak	yes
4	rainy	mild	high	weak	yes
5	rainy	cool	normal	weak	yes
6	rainy	cool	normal	strong	no
7	overcast	cool	normal	strong	yes
8	sunny	mild	high	weak	no
9	sunny	cool	normal	weak	yes
10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

- Outlook: G(S, Outlook) = 0.25
- Temp: G(S, Temp) = 0.03
- Humidity: G(S, humidity) = 0.16
- Wind: G(S, Wind) = 0.05

➔ Chọn Outlook làm gốc.

Tổng hợp

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



Xét nhánh Outlook = Sunny

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

- Tổng số lượng dữ liệu: S = 5
- Số thuộc tính thuộc nhãn Play = yes: 3
- Số thuộc tính thuộc nhãn Play = no: 2

➔ $H(S) = -\left(\frac{3}{5}\log_2\left(\frac{3}{5}\right) + \frac{2}{5}\log_2\left(\frac{2}{5}\right)\right)$

$= 0.97$

Tính toán cho thuộc tính Temp

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

	Yes	No	Total
Hot	0	2	2
Mild	1	1	2
Cool	1	0	1

$$\mathrm{H}(\mathrm{S}, temp) = -\frac{2}{5}\left(\frac{2}{2}\log_2\left(\frac{2}{2}\right)\right) - \frac{2}{5}\left(\frac{1}{2}\log_2\left(\frac{1}{2}\right) + \frac{1}{2}\log_2\left(\frac{1}{2}\right)\right) - \frac{1}{5}\left(\frac{1}{1}\log_2\left(\frac{1}{1}\right)\right)$$

= 0.4

$$G(S, Temp) = 0.97 - 0.4 = 0.57$$

Tính toán cho thuộc tính Humidity

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

	Yes	No	Total
High	0	3	3
Normal	2	0	2

$$\mathrm{H}(S, humidity) = -\frac{3}{5}\left(\frac{3}{3}\log_2\left(\frac{3}{3}\right)\right) - \frac{2}{5}\left(\frac{2}{2}\log_2\left(\frac{2}{2}\right)\right)$$

$$= 0$$

$$G(S, humidity) = 0.97 - 0 = 0.97$$

Tính toán cho thuộc tính Wind

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

	Yes	No	Total
Weak	1	2	3
Strong	1	1	2

$$\mathrm{H}(S, Wind) = -\frac{3}{5}\left(\frac{1}{3}\log_2\left(\frac{1}{3}\right) + \frac{2}{3}\log_2\left(\frac{2}{3}\right)\right) - \frac{2}{5}\left(\frac{1}{2}\log_2\left(\frac{1}{2}\right) + \frac{1}{2}\log_2\left(\frac{1}{2}\right)\right)$$

= 0.95

$$G(S, Wind) = 0.97 - 0.95 = 0.02$$

Tổng hợp

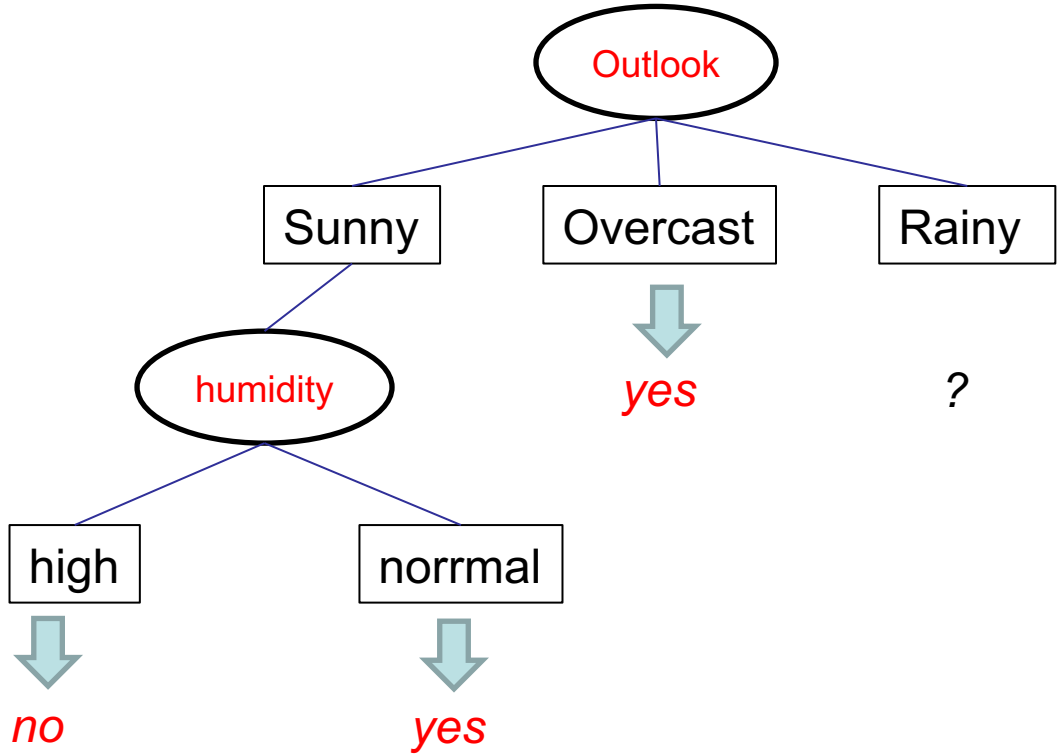
day	outlook	temperature	humidity	wind	play
1	sunny	hot	high	weak	no
2	sunny	hot	high	strong	no
~~3~~	~~overcast~~	~~hot~~	~~high~~	~~weak~~	~~yes~~
~~4~~	~~rainy~~	~~mild~~	~~high~~	~~weak~~	~~yes~~
~~5~~	~~rainy~~	~~cool~~	~~normal~~	~~weak~~	~~yes~~
~~6~~	~~rainy~~	~~cool~~	~~normal~~	~~strong~~	~~no~~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

- Temp: G(S, Temp) = 0.57
- Humidity: G(S, humidity) = 0.97
- Wind: G(S, Wind) = 0.02

➔ Chọn Humidity làm gốc.

Tổng hợp

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



Xét nhánh Outlook = Rainy

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

- Tổng số lượng dữ liệu: S = 5
- Số thuộc tính thuộc nhãn Play = yes: 3
- Số thuộc tính thuộc nhãn Play = no: 2

➔ $H(S) = -\left(\frac{3}{5}\log_2\left(\frac{3}{5}\right) + \frac{2}{5}\log_2\left(\frac{2}{5}\right)\right)$

$= 0.97$

Tính toán cho thuộc tính Temp

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

	Yes	No	Total
Mild	2	1	3
Cool	1	1	2

$$\mathrm{H}(\mathrm{S}, temp) = -\frac{3}{5}\left(\frac{2}{3}\log_2\left(\frac{2}{3}\right) + \frac{1}{3}\log_2\left(\frac{1}{3}\right)\right) - \frac{2}{5}\left(\frac{1}{2}\log_2\left(\frac{1}{2}\right) + \frac{1}{2}\log_2\left(\frac{1}{2}\right)\right)$$

= 0.95

$$G(S, Temp) = 0.97 - 0.95 = 0.02$$

Tính toán cho thuộc tính Humidity

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

	Yes	No	Total
High	1	1	2
Normal	2	1	3

$$\mathrm{H}(S, humidity) = -\frac{2}{5}\left(\frac{1}{2}\log_2\left(\frac{1}{2}\right) + \frac{1}{2}\log_2\left(\frac{1}{2}\right)\right) - \frac{3}{5}\left(\frac{2}{3}\log_2\left(\frac{2}{3}\right) + \frac{1}{3}\log_2\left(\frac{1}{3}\right)\right)$$

= 0.55

$$G(S, humidity) = 0.97 - 0.95 = 0.02$$

Tính toán cho thuộc tính Wind

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

	Yes	No	Total
Weak	3	0	3
Strong	2	0	2

$$\mathrm{H}(S, Wind) = -\frac{3}{5}\left(\frac{3}{3}\log_2\left(\frac{3}{3}\right)\right) - \frac{2}{5}\left(\frac{2}{2}\log_2\left(\frac{2}{2}\right)\right)$$
$$= 0$$

$$G(S, Wind) = 0.97 - 0 = 0.97$$

Tổng hợp

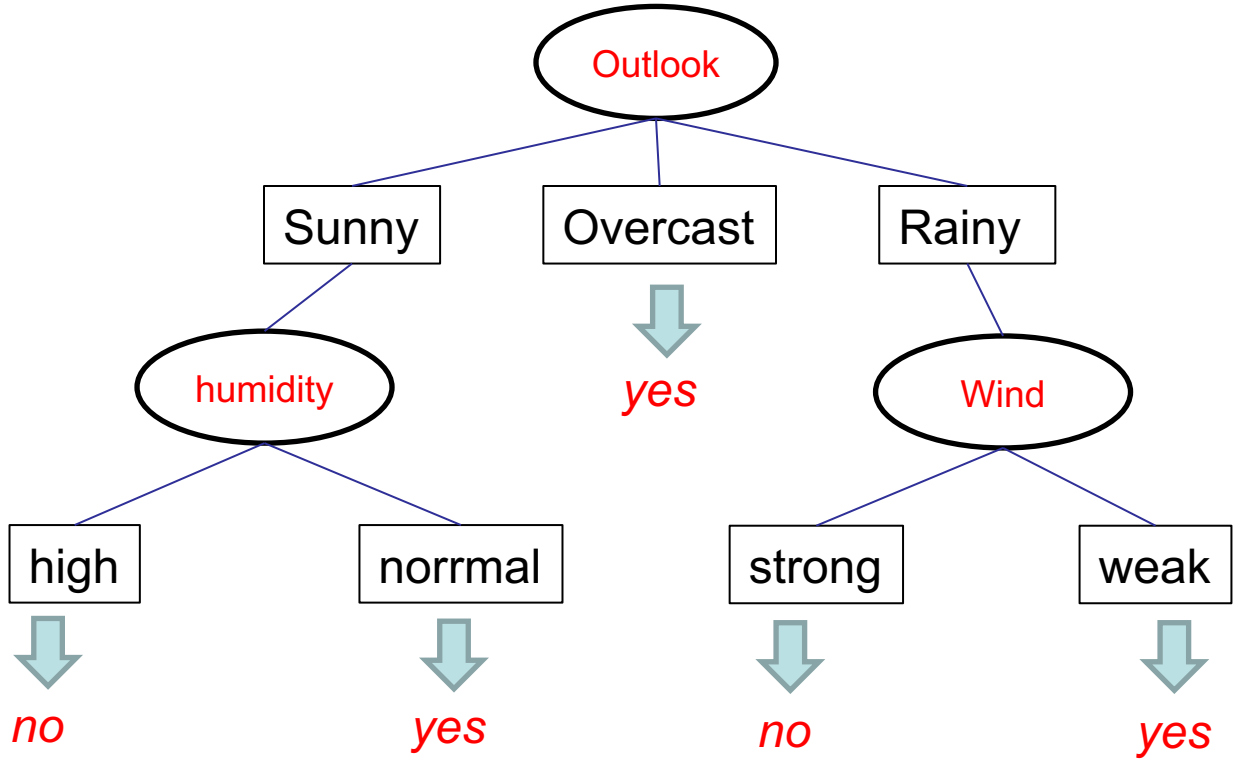
day	outlook	temperature	humidity	wind	play
~~1~~	~~sunny~~	~~hot~~	~~high~~	~~weak~~	~~no~~
~~2~~	~~sunny~~	~~hot~~	~~high~~	~~strong~~	~~no~~
~~3~~	~~overcast~~	~~hot~~	~~high~~	~~weak~~	~~yes~~
4	rainy	mild	high	weak	yes
5	rainy	cool	normal	weak	yes
6	rainy	cool	normal	strong	no
~~7~~	~~overcast~~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

- Temp: G(S, Temp) = 0.02
- Humidity: G(S, humidity) = 0.02
- Wind: G(S, Wind) = 0.97

➔ Chọn Wind làm node.

Tổng hợp

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



Dừng !!!

Sử dụng chỉ số Gini

Tính toán cho thuộc tính Play

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

- Tổng số lượng dữ liệu: S = 14
- Số thuộc tính thuộc nhãn Play = yes: 9
- Số thuộc tính thuộc nhãn Play = no: 5

Tính toán cho thuộc tính Outlook

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

	Yes	No	Total
Sunny	3	2	5
Overcast	4	0	4
Rainy	3	2	5

$$G(\text{outlook} = sunny) = 1 - \left(\frac{3}{5}\right)^2 - \left(\frac{2}{5}\right)^2 = 0.48$$

$$G(\text{outlook} = overcast) = 1 - \left(\frac{4}{4}\right)^2 - \left(\frac{0}{4}\right)^2 = 0$$

$$G(\text{outlook} = sunny) = 1 - \left(\frac{3}{5}\right)^2 - \left(\frac{2}{5}\right)^2 = 0.48$$

$$L(outlook) = \frac{5}{14} * 0.48 + \frac{4}{14} * 0 + \frac{5}{14} * 0.48 = 0.342$$

Tính toán cho thuộc tính Temp

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

	Yes	No	Total
Hot	2	2	4
Mild	4	2	6
Cool	3	1	4

$$G(temp = hot) = 1 - \left(\frac{2}{4}\right)^2 - \left(\frac{2}{4}\right)^2 = 0.5$$

$$G(temp = cool) = 1 - \left(\frac{3}{4}\right)^2 - \left(\frac{1}{4}\right)^2 = 0.375$$

$$G(temp = mild) = 1 - \left(\frac{4}{6}\right)^2 - \left(\frac{3}{6}\right)^2 = 0.445$$

$$L(temp) = \frac{4}{14} * 0.5 + \frac{6}{14} * 0.445 + \frac{4}{14} * 0.375 = 0.439$$

Tính toán cho thuộc tính Humidity

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

	Yes	No	Total
High	3	4	7
Normal	6	1	7

$$G(humidity = high) = 1 - \left(\frac{3}{7}\right)^2 - \left(\frac{4}{7}\right)^2 = 0.489$$

$$G(humidity = normal) = 1 - \left(\frac{6}{7}\right)^2 - \left(\frac{1}{7}\right)^2 = 0.244$$

$$L(Humidity) = \frac{7}{14} * 0.489 + \frac{7}{14} * 0.244 = 0.367$$

Tính toán cho thuộc tính Wind

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

	Yes	No	Total
Weak	6	2	8
Strong	3	3	6

$$G(wind = weak) = 1 - \left(\frac{6}{8}\right)^2 - \left(\frac{2}{8}\right)^2 = 0.375$$

$$G(wind = strong) = 1 - \left(\frac{3}{6}\right)^2 - \left(\frac{3}{6}\right)^2 = 0.5$$

$$L(Wind) = \frac{8}{14} * 0.375 + \frac{6}{14} * 0.5 = 0.428$$

Tổng hợp

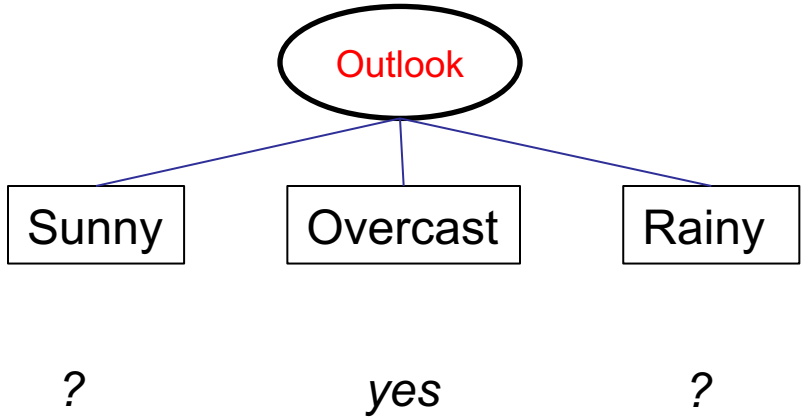
day	outlook	temperature	humidity	wind	play
1	sunny	hot	high	weak	no
2	sunny	hot	high	strong	no
3	overcast	hot	high	weak	yes
4	rainy	mild	high	weak	yes
5	rainy	cool	normal	weak	yes
6	rainy	cool	normal	strong	no
7	overcast	cool	normal	strong	yes
8	sunny	mild	high	weak	no
9	sunny	cool	normal	weak	yes
10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

- Outlook: L(Outlook) = 0.342
- Temp: L(Temp) = 0.439
- Humidity: L(Humidity) = 0.367
- Wind: L(Wind) = 0.428

➔ Chọn Outlook làm gốc (loss có giá trị nhỏ nhất).

Tổng hợp

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



Xét nhánh Outlook = Sunny

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

- Tổng số lượng dữ liệu: S = 5
- Số thuộc tính thuộc nhãn Play = yes: 2
- Số thuộc tính thuộc nhãn Play = no: 3

Tính toán cho thuộc tính Temp

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

	Yes	No	Total
Hot	0	2	2
Mild	1	1	2
Cool	1	0	1

$$G(temp = hot) = 1 - \left(\frac{0}{2}\right)^2 - \left(\frac{2}{2}\right)^2 = 0$$

$$G(temp = cool) = 1 - \left(\frac{1}{2}\right)^2 - \left(\frac{1}{2}\right)^2 = 0.5$$

$$G(temp = mild) = 1 - \left(\frac{1}{1}\right)^2 - \left(\frac{0}{1}\right)^2 = 0$$

$$L(temp) = \frac{2}{5} * 0 + \frac{2}{5} * 0.5 + \frac{1}{5} * 0 = 0.2$$

Tính toán cho thuộc tính Humidity

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

	Yes	No	Total
High	0	3	3
Normal	2	0	2

$$G(humidity = high) = 1 - \left(\frac{0}{3}\right)^2 - \left(\frac{3}{3}\right)^2 = 0$$

$$G(humidity = normal) = 1 - \left(\frac{2}{2}\right)^2 - \left(\frac{0}{2}\right)^2 = 0$$

$$L(Humidity) = \frac{3}{5} * 0 + \frac{2}{5} * 0 = 0$$

Tính toán cho thuộc tính Wind

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

	Yes	No	Total
Weak	1	2	3
Strong	1	1	2

$$G(wind = weak) = 1 - \left(\frac{1}{3}\right)^2 - \left(\frac{2}{3}\right)^2 = 0.266$$

$$G(wind = strong) = 1 - \left(\frac{1}{2}\right)^2 - \left(\frac{1}{2}\right)^2 = 0.2$$

$$L(Wind) = \frac{3}{5} * 0.266 + \frac{2}{5} * 0.2 = 0.466$$

Tổng hợp

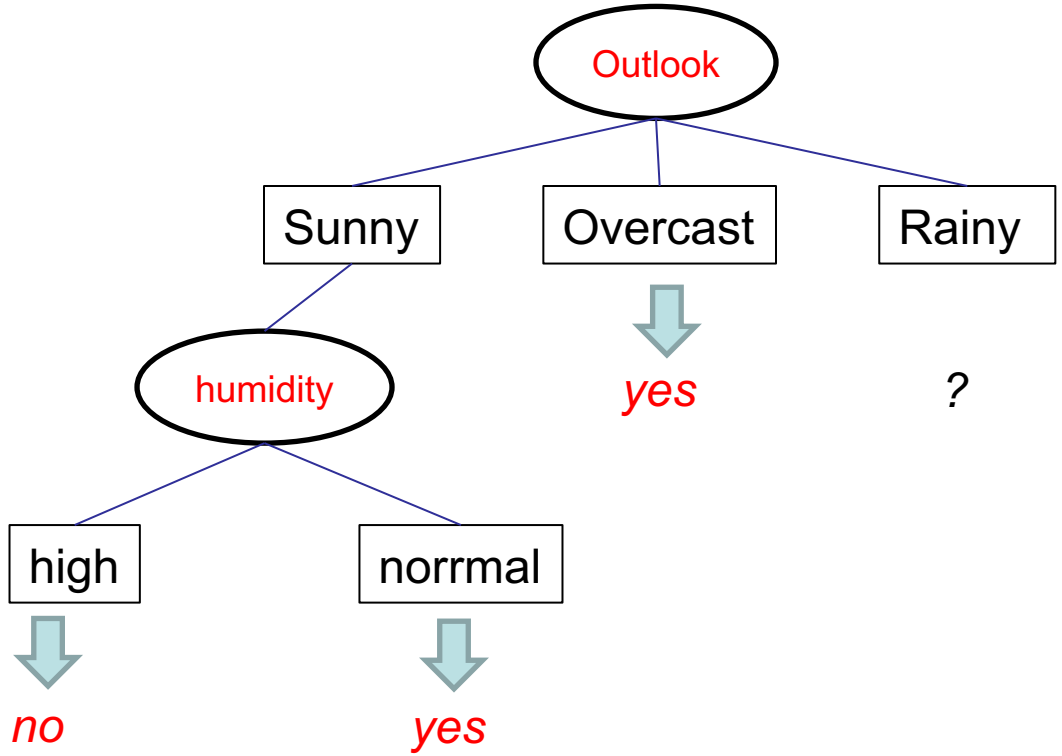
day	outlook	temperature	humidity	wind	play
1	sunny	hot	high	weak	no
2	sunny	hot	high	strong	no
~~3~~	~~overcast~~	~~hot~~	~~high~~	~~weak~~	~~yes~~
~~4~~	~~rainy~~	~~mild~~	~~high~~	~~weak~~	~~yes~~
~~5~~	~~rainy~~	~~cool~~	~~normal~~	~~weak~~	~~yes~~
~~6~~	~~rainy~~	~~cool~~	~~normal~~	~~strong~~	~~no~~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

- Temp: L(Temp) = 0.2
- Humidity: L(Humidity) = 0
- Wind: L(Wind) = 0.466

➔ Chọn Humidity làm node (loss nhỏ nhất).

Tổng hợp

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



Xét nhánh Outlook = Rainy

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

- Tổng số lượng dữ liệu: S = 5
- Số thuộc tính thuộc nhãn Play = yes: 3
- Số thuộc tính thuộc nhãn Play = no: 2

Tính toán cho thuộc tính Temp

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

	Yes	No	Total
Mild	2	1	3
Cool	1	1	2

$$G(temp = cool) = 1 - \left(\frac{1}{2}\right)^2 - \left(\frac{1}{2}\right)^2 = 0.5$$

$$G(temp = mild) = 1 - \left(\frac{2}{3}\right)^2 - \left(\frac{1}{3}\right)^2 = 0.444$$

$$L(temp) = \frac{2}{5} * 0.5 + \frac{3}{5} * 0.444 = 0.466$$

Tính toán cho thuộc tính Humidity

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

	Yes	No	Total
High	1	1	2
Normal	2	1	3

$$G(humidity = high) = 1 - \left(\frac{1}{2}\right)^2 - \left(\frac{1}{2}\right)^2 = 0.5$$

$$G(humidity = normal) = 1 - \left(\frac{2}{3}\right)^2 - \left(\frac{1}{3}\right)^2 = 0.444$$

$$L(Humidity) = \frac{2}{5} * 0.5 + \frac{3}{5} * 0.444 = 0.466$$

Tính toán cho thuộc tính Wind

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

	Yes	No	Total
Weak	3	0	3
Strong	0	2	2

$$G(wind = weak) = 1 - \left(\frac{3}{3}\right)^2 - \left(\frac{0}{3}\right)^2 = 0$$

$$G(wind = strong) = 1 - \left(\frac{0}{2}\right)^2 - \left(\frac{2}{2}\right)^2 = 0$$

$$L(Wind) = \frac{3}{5} * 0 + \frac{2}{5} * 0 \ = 0$$

Tổng hợp

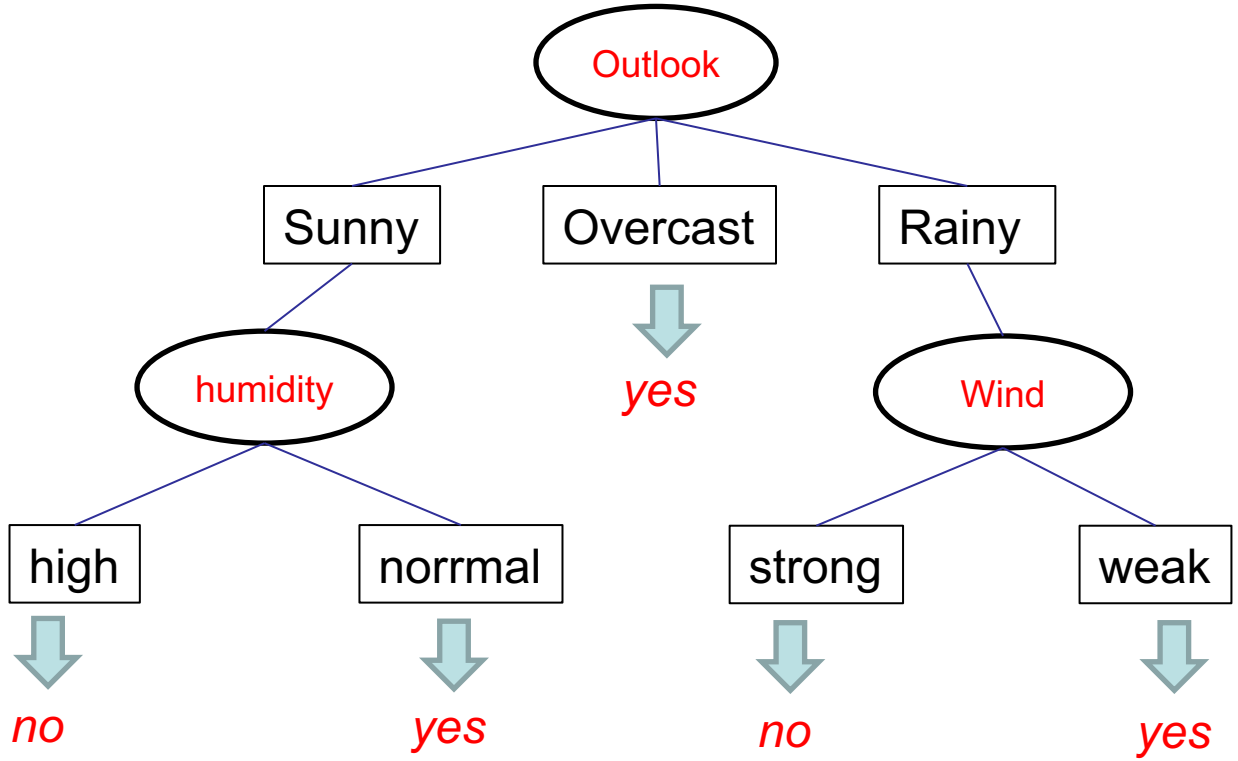
day	outlook	temperature	humidity	wind	play
~~1~~	~~sunny~~	~~hot~~	~~high~~	~~weak~~	~~no~~
~~2~~	~~sunny~~	~~hot~~	~~high~~	~~strong~~	~~no~~
~~3~~	~~overcast~~	~~hot~~	~~high~~	~~weak~~	~~yes~~
4	rainy	mild	high	weak	yes
5	rainy	cool	normal	weak	yes
6	rainy	cool	normal	strong	no
~~7~~	~~overcast~~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

- Temp: L(Temp) = 0.466
- Humidity: L(humidity) = 466
- Wind: L(Wind) = 0

➔ Chọn Wind làm node (loss nhỏ nhất).

Tổng hợp

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



Dừng !!!

Bài tập

Class-Labeled Training Tuples from the *AllElectronics* Customer Database

RID	age	income	student	credit_rating	Class: buys_computer
1	youth	high	no	fair	no
2	youth	high	no	excellent	no
3	middle_aged	high	no	fair	yes
4	senior	medium	no	fair	yes
5	senior	low	yes	fair	yes
6	senior	low	yes	excellent	no
7	middle_aged	low	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	medium	yes	fair	yes
11	youth	medium	yes	excellent	yes
12	middle_aged	medium	no	excellent	yes
13	middle_aged	high	yes	fair	yes
14	senior	medium	no	excellent	no

Bài tập

ID	Độ tuổi	Hôn nhân	Sở hữu BĐS	Thu nhập	Rủi ro tín dụng
1	25	Độc thân	Ở cùng bố mẹ	7000000	0
2	40	Đã kết hôn	Nhà sở hữu	18000000	0
3	35	Từng ly hôn	Nhà thuê	12000000	1
4	27	Đã kết hôn	Ở cùng bố mẹ	9000000	1
5	31	Độc thân	Nhà thuê	6000000	1
6	36	Đã kết hôn	Nhà sở hữu	8000000	1
7	48	Độc thân	Nhà thuê	7000000	0
8	26	Đã kết hôn	Nhà thuê	8000000	1
9	33	Từng ly hôn	Ở cùng bố mẹ	5000000	1
10	29	Độc thân	Nhà thuê	10000000	0
11	38	Đã kết hôn	Nhà sở hữu	15000000	0
12	44	Độc thân	Nhà sở hữu	14000000	1
13	42	Đã kết hôn	Nhà sở hữu	10000000	0
14	28	Độc thân	Nhà thuê	7000000	1
15	30	Đã kết hôn	Ở cùng bố mẹ	6000000	1

Hiện thực bằng thư viện sklearn

- Tạo file csv từ dữ liệu (file golf.csv).

- Đọc dữ liệu bằng thư viện pandas:

```
1. import pandas as pd
2. dataset = pd.read_csv(golf.csv', index_col=False)
```

Mã hoá dữ liệu

- Mã hoá các giá trị ứng với từng thuộc tính thành dạng số (numeric).
- Sử dụng thư viện LabelEncoder trong sklearn.

```
1. from sklearn.preprocessing import LabelEncoder
2. from copy import deepcopy
3. le = LabelEncoder()
4. cols = dataset.columns
5. dataset_encoded = deepcopy(dataset)
6. for c in cols:
7.    dataset_encoded[c] = le.fit_transform(dataset_encoded[c])
```

Huấn luyện mô hình

- Chuẩn bị dữ liệu:

```
1. X = dataset_encoded.iloc[:, 1:5]
2. y = dataset_encoded['play']
```

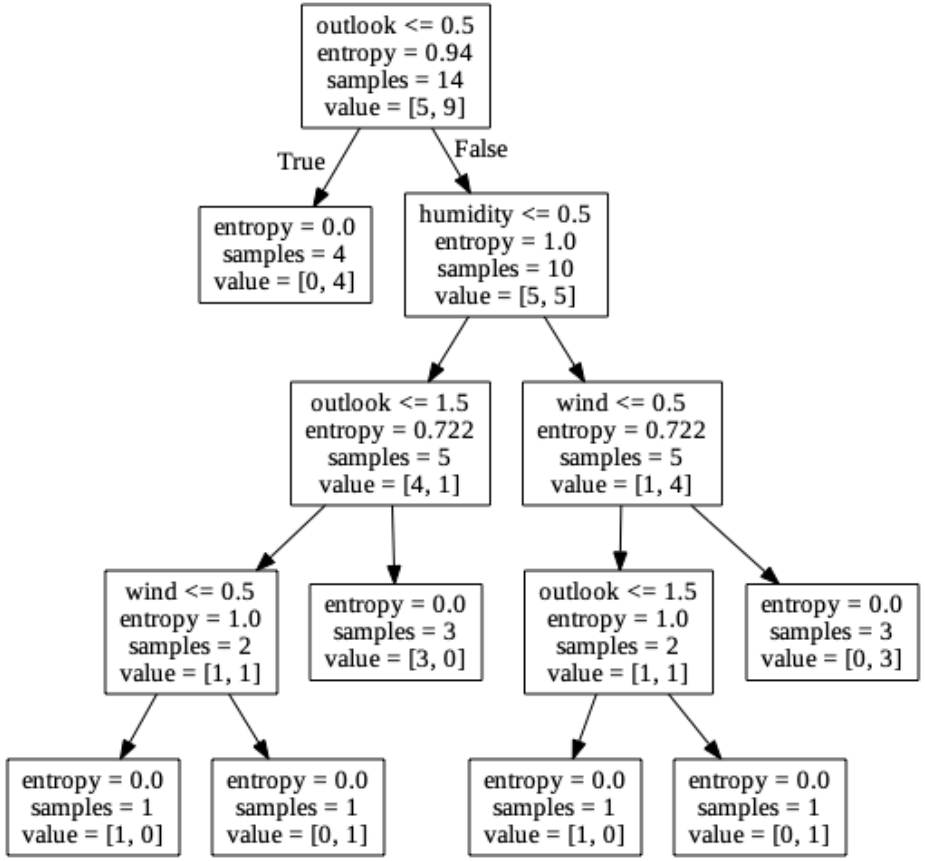
- Huấn luyện cây quyết định: sử dụng thư viện *DecisionTreeClassifier* với độ đo Entropy.

```
1. from sklearn.tree import DecisionTreeClassifier


2. model = DecisionTreeClassifier(criterion='entropy')
3. model.fit(X,y)
```

Visualize mô hình

```
1. from sklearn.tree import export_graphviz
2. dot_data = export_graphviz(
3.                 model,
4.                 feature_names=dataset.columns[1:5]
5.             )
6. graph = graphviz.Source(dot_data)
7. graph = graphviz.Source(dot_data)
8. graph.render(golf')
```



Bài tập áp dụng

- Bộ dữ liệu Titanic dataset dùng để dự đoán khả năng một người sống sót sau thảm hoạ Titanic dựa vào các thuộc tính khác nhau.
- Link: https://www.kaggle.com/c/titanic/data?select=train.csv
- Yêu cầu:
  1. Xây dựng mô hình Decision Tree từ bộ dữ liệu trên.
  2. Đánh giá khả năng phân lớp của mô hình (dựa vào tập test và các độ đo đã học).

TÀI LIỆU THAM KHẢO

1. Chương 5 của sách: *Hands-on Machine Learning with ScikitLearn, Keras & TensorFlow.*
2. Chương 6 của sách: *Hands-on Machine Learning with ScikitLearn, Keras & TensorFlow.*