ĐẠI HỌC QUỐC GIA TP. HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN
KHOA KHOA HỌC MÁY TÍNH

BÁO CÁO ĐỒ ÁN MÔN HỌC
CẤU TRÚC DỮ LIỆU VÀ GIẢI THUẬT NÂNG CAO
CS523.M21.KHCL

ĐỀ TÀI:
DECISION TREE

GIẢN VIÊN HƯỚNG DẪN: **NGUYỄN THANH SƠN**
SINH VIÊN THỰC HIỆN: **LÊ THẾ TUẤN: 20522113**
NGUYỄN THANH PHÚC: 20521769
PHẠM KIÊN: 20521490

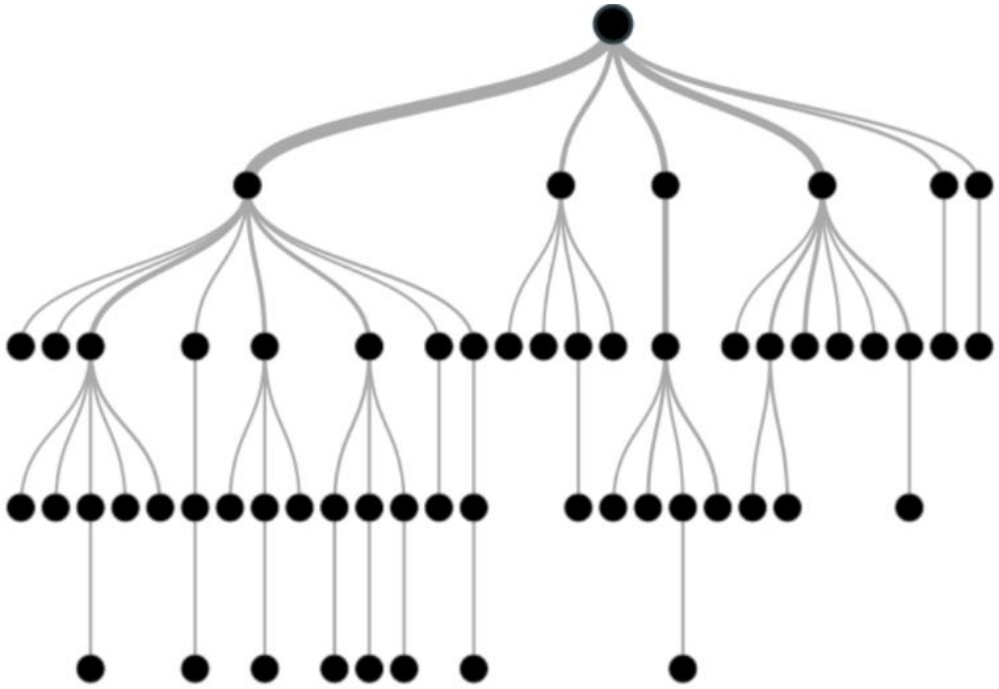
TP. HỒ CHÍ MINH, 04/2022

Mục lục

1) Giới thiệu.

a) Giới thiệu chung.

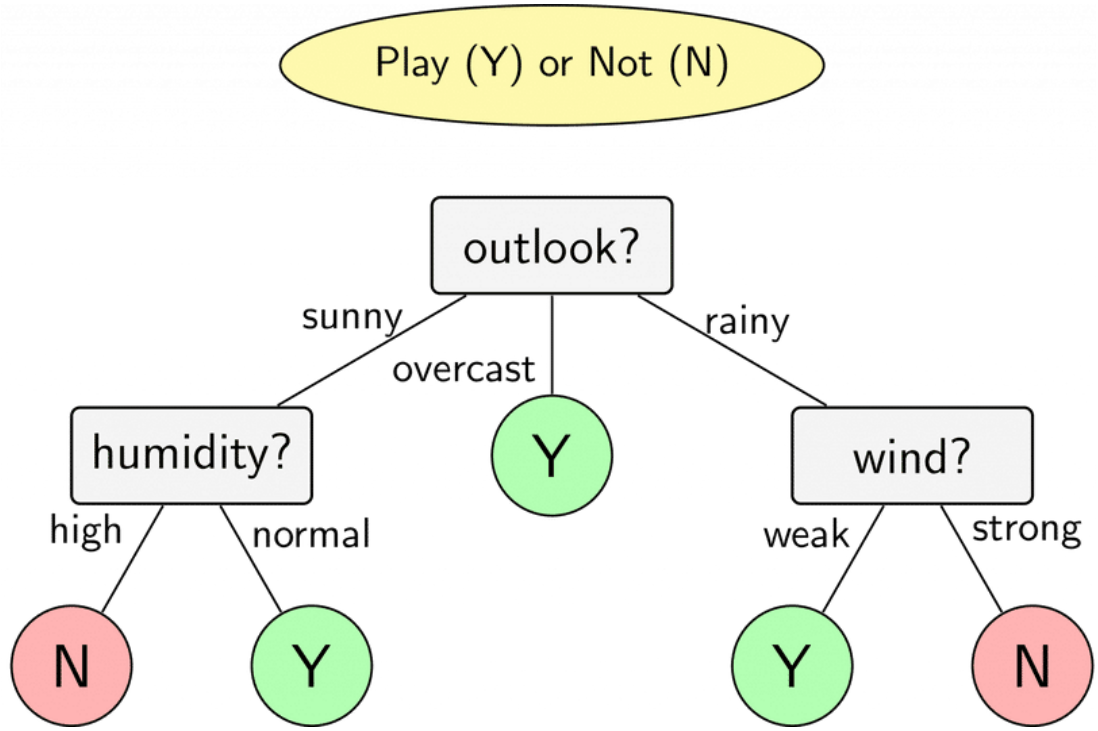
Cây quyết định là một thuật toán dựa trên cấu trúc cây được sử dụng để giải quyết vấn đề hồi quy (Regression) hoặc phân loại (Classification), và rất hữu ích cho các bộ dữ liệu phức tạp. Cây quyết định thuộc mô hình thuật toán học có giảm sát (Supervised learning), nó hoạt động bằng cách quan sát các đặc điểm, tính năng của một đối tượng và đào tạo một mô hình trong cấu trúc của cây, chia nhỏ tập dữ liệu thành các tập con ngày càng nhỏ hơn và sau đó đưa ra các dự đoán dữ liệu trong tương lai dựa trên tập con.



b) Cây quyết định.

Cây quyết định là một cây phân cấp có cấu trúc được dùng để phân lớp các đối tượng dựa vào dãy các luật. Các thuộc tính của đối tượng có thể thuộc các kiểu dữ liệu khác nhau như Nhị phân (Binary) Định danh (Nominal), Thứ tự (Ordinal), Số lượng (Quantitative) trong khi đó thuộc tính phân lớp phải có kiểu dữ liệu là Binary hoặc Ordinal. Nói chung, cho dữ liệu về các đối tượng gồm các thuộc tính cùng với lớp (classes) của nó, cây quyết định sẽ sinh ra các luật để dự đoán lớp của các dữ liệu chưa biết.

Nhìn chung Decision Tree thường áp dụng vào cả 2 bài toán: Phân loại (Classification) và Hồi quy (Regression). Dạng phân loại kết quả thường là rời rạc và không có thứ tự (Ví dụ: mô hình dự đoán thời tiết dự đoán có hay không mưa vào một ngày cụ thể), dạng hồi quy thì dữ liệu có vector đặc trưng bao gồm cả thuộc tính dạng categorical và liên tục, có ý nghĩa (Ví dụ: một mô hình dự đoán lợi nhuận cho biết lợi nhuận có thể được tạo ra từ việc bán một sản phẩm).

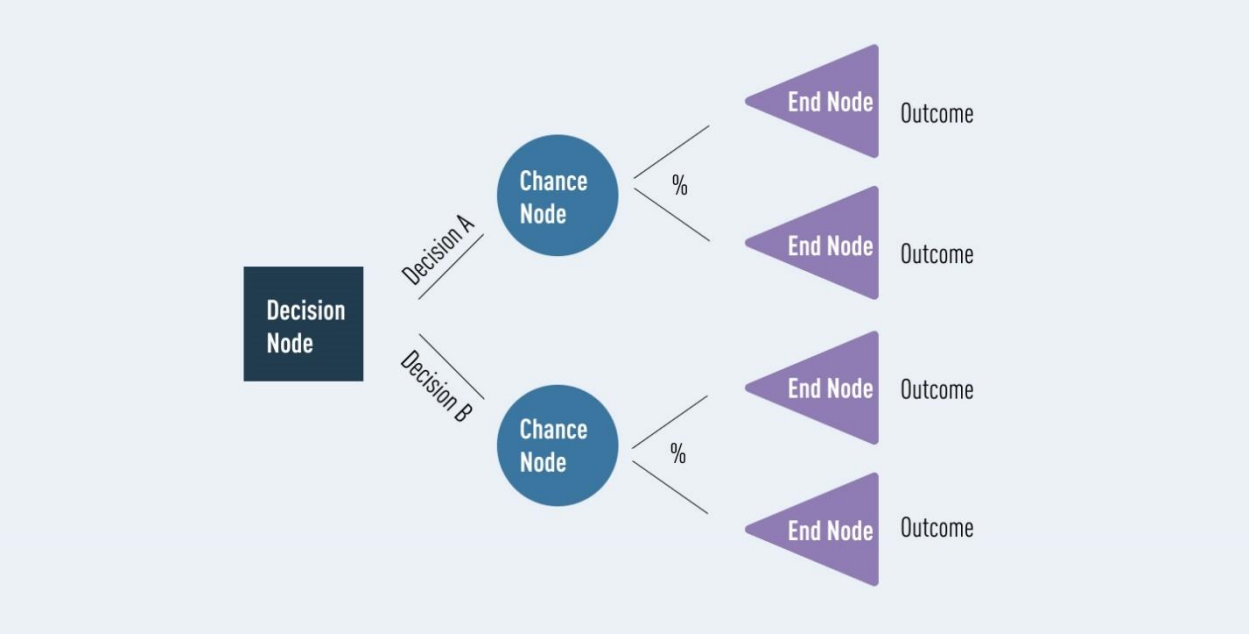


c) Thành phần của cây quyết định.

Thành phần chính:

- Nút quyết định: Đại diện cho một quyết định.

- Nút cơ hội: Đại diện cho xác suất hoặc sự không chắc chắn.
- Nút kết thúc: Đại diện cho một kết quả.



Cây phân loại hay cây hồi quy đều hoạt động dựa theo mô hình trên. Điểm khác nhau giữa chúng là ở đầu ra. Nếu mục tiêu dự đoán của cây là phân loại thì sẽ là cây phân loại, còn nếu là số thực thì sẽ là cây hồi quy.

d) Ưu điểm, nhược điểm.

Ưu điểm:

- Cây quyết định là một thuật toán đơn giản và phổ biến. Mô hình sinh ra các quy tắc dễ hiểu cho người đọc, tạo ra bộ luật với mỗi nhánh lá là một luật của cây. Cây có thể được trực quan hóa.
- Dữ liệu cần chuẩn bị nhỏ. Dữ liệu đầu vào có thể là là dữ liệu trống, không cần chuẩn hóa hoặc tạo biến giả.
- Có khả năng xử lý các vấn đề yêu cầu đa đầu ra. Sử dụng mô hình hộp màu trắng. Nếu một tình huống
- nhất định có thể quan sát được trong 1 mô hình, thì lời giải thích cho điều kiện đó dễ dàng được giải thích bằng logic boolean.
- Có thể xác thực mô hình bằng cách sử dụng các kiểm tra thống kê
- Có khả năng là việc với dữ liệu lớn.

Nhược điểm:

- Mô hình cây quyết định phụ thuộc rất lớn vào dữ liệu của bạn. Thậm chí, với một sự thay đổi nhỏ trong bộ dữ liệu, cấu trúc mô hình cây quyết định có thể thay đổi hoàn toàn.
- Mô hình cây quyết định có thể trở thành cây thiên vị nếu một số lớp chiếm ưu thế. Do đó, nên cân bằng dữ liệu trước khi fitting.
- Mô hình cây quyết định có thể tạo ra cây quá phức tạp, khiến cho việc tổng quan hóa dữ liệu không được tốt. Vấn đề này gọi là overfitting. Để tránh vấn đề này, các cơ chế như pruning, thiết lập số lượng mẫu tối thiểu tại một nút lá hoặc thiết lập độ sâu tối đa của cây là cần thiết.

e) Ứng dụng.

Cây quyết định là một công cụ mạnh mẽ và phổ biến, thường được dụng trong phân tích và dự đoán và là công cụ mạnh mẽ cho máy học và trí tuệ nhân tạo, được sử dụng làm thuật toán đào tọa cho học có giám sát. Và được ứng dụng trong nhiều nghành công nghiệp khác nhau (công nghệ, tài chính, y tế…)

- Doanh nghiệp đánh giá cơ hội mở rộng, phát triển dựa trên việc phân tích dữ liệu bán hàng từ quá khứ.
- Công ty quyết định chi tiêu cho ngân sách quảng cáo dựa trên sức mua của khách hàng.
- Ngân hàng sử dụng dữ liệu khách hàng cung cấp để dự đoán xem khách hàng đó có thể thanh toán khoảng vay hay không.

2) Thuật toán.

a) Giới thiệu chung.

Thuật toán cây quyết định là một cây trong đó các nút đại diện cho các tính năng (thuộc tính), nhánh đại diện cho quyết định (quy tắc) và các nút lá đại diện cho các kết quả (rời rạc và liên tục).

Có nhiều thuật toán khác nhau được sử dụng để tạo cây quyết định từ dữ liệu, một số thuật toán như sau:

- Classification and regression tree CART
- ID 3(Information Gain)
- CHAID
- ID 4.5

Trong phần trình bày này đề cập đến 2 thuật toán:

- Thuật toán CART

    - GINI Index
- Thuật toán ID3
    - Hàm Entropy
    - Information Gain

Câu hỏi đặt ra: Làm thế nào để chọn ra được node root?

Thuộc tính phân loại tốt nhất dữ liệu đào tạo, sử dụng thuộc tính này thuộc tính ở gốc của cây.

Vậy làm thế nào để chọn ra được thuộc tính tốt nhất để làm node root, đó là lý do mà thuật toán CART và ID3 được tạo ra.

Thuật toán tìm ra cây quyết định sẽ theo quy tắc tại các node root thì độ hỗn loạn của dữ liệu sẽ là cao nhất và không biết nó thuộc về lớp nào, còn tại các node lá thì độ hỗn loạn của dữ liệu thấp nhất và biết nó thuộc về lớp nào (có nghĩa là giảm Entropy về 0 nhanh nhất có thể, độ sâu của cây quyết định thấp nhất).

Thuật toán dừng khi tất cả các thuộc tính thuộc cùng 1 lớp (hay nói các cách là đến node lá).

b) Thuật toán CART.

- Được sử dụng để tạo ra cả cây phân loại và cây hồi quy.

- Nó sử dụng bình phương nhỏ nhất làm số liệu để chọn các tính năng trong trường hợp Cây hồi quy.

- Gini là chỉ số thể hiện mức độ phân loại sai khi ta chọn ngẫu nhiên một phần tử từ tập data, được sử dụng để đo lường mức độ bất bình đẳng trong phân phối của các lớp, được tính bằng cách lấy 1 trừ đi tổng bình phương tỷ lệ phần trăm ở mỗi lớp.

Công thức tính giá trị GINI

$$1 - \sum_{i=1}^{n} p_i^2$$

Dùng thuật toán CART để tạo ra classification tree trong bộ dữ liệu thời tiết với mục tiêu là thời tiết có đảm bảo cho việc đi chơi hay không (có hay không).

Day	outlook	temperature	humidity	wind	Decision
1	sunny	hot	high	weak	No
2	sunny	hot	high	strong	No
3	overcast	hot	high	weak	Yes
4	rainfall	mild	high	weak	Yes
5	rainfall	cool	normal	weak	Yes
6	rainfall	cool	normal	strong	No
7	overcast	cool	normal	wtrong	Yes
8	sunny	mild	high	weak	No
9	sunny	cool	normal	weak	Yes
10	rainfall	mild	normal	weak	Yes
11	sunny	mild	normal	strong	Yes
12	overcast	mild	high	strong	Yes
13	overcast	hot	normal	weak	Yes
14	rainfall	mild	high	strong	No

Theo bộ dữ liệu thời tiết thì có thể dễ dàng thấy được **outlook** sẽ bao gồm các feature: **sunny, overcast, rainfall.**

- Tính giá trị Gini cho Outlook

Outlook	Yes	No	# Instances
sunny	2	3	5
overcast	4	0	4
rainfall	3	2	5

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

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

$$Gini(outlook = rainfall) = 1 - \left(\frac{3}{5}\right)^2 - \left(\frac{2}{5}\right)^2 = 0.48$$

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

- Tính giá trị Gini cho Temperature

Temperature	Yes	No	# Instances
hot	2	2	4
cool	3	1	4
mild	4	2	6

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

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

$$Gini(temperature = mild) = 1 - \left(\frac{4}{6}\right)^2 - \left(\frac{2}{6}\right)^2 = 0.445$$

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

- Tính giá trị Gini cho Humidity

Humidity	Yes	No	# Instances
high	3	4	7
Normal	6	1	7

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

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

$$Gini(humidity) = \frac{7}{14} * 0.489 + \frac{7}{14} * 0.244 = 0.369$$

- Tính giá trị Gini cho Wind

wind	Yes	No	# Instances
weak	6	2	8
strong	3	3	6

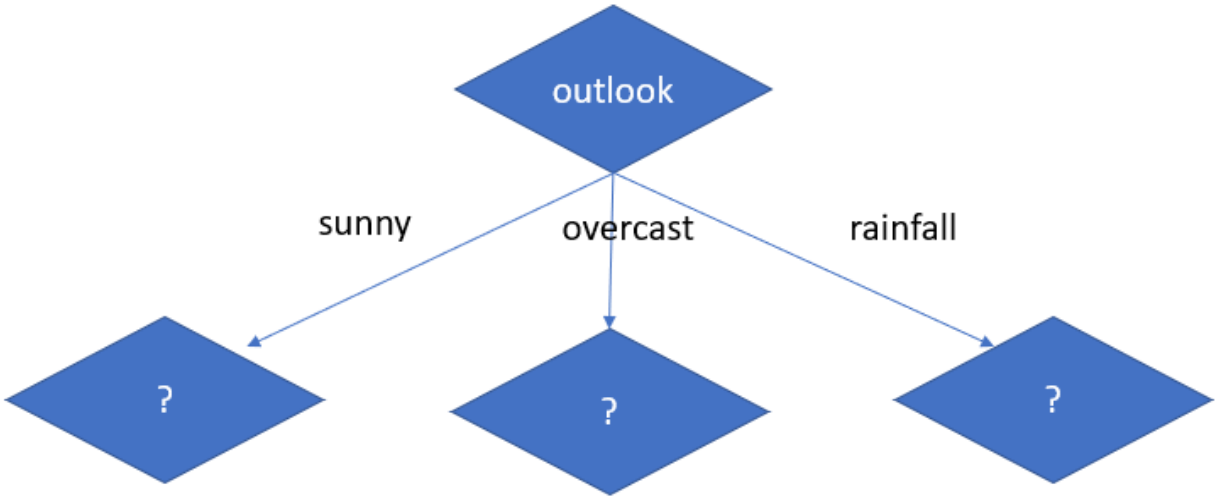
$$Gini(wind = weak) = 1 - \left(\frac{6}{8}\right)^2 - \left(\frac{2}{8}\right)^2 = 0.375$$

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

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

Features	Gini Index
outlook	0.342
temperature	0.439
humidity	0.367
wind	0.428

Dựa vào bảng trên ta thấy Outlook có giá trị GINI là thấp nhất nên chọn outlook là root của cây



Sau khi tìm được root là Outlook thì chúng ta sẽ tính Giá trị GINI cho từng dữ liệu phụ của outlook bao gồm: **sunny, overcast và rainfall.**

Day	outlook	temperature	humidity	wind	decision
1	sunny	hot	high	weak	No
2	sunny	hot	high	strong	No
8	sunny	mild	high	weak	No
9	sunny	cool	normal	weak	Yes
11	sunny	mild	normal	strong	Yes

Trường hợp sunny:

- **Chỉ số Gini cho Temperature khi outlook = sunny**

Temperature	Yes	No	# Instances
hot	0	2	2
cool	1	1	1
mild	1	1	2

$$Gini(\text{outlook} = \text{sunny} \,\&\, \text{temperature} = \text{hot}) = 1 - \left(\frac{0}{2}\right)^2 - \left(\frac{2}{2}\right)^2 = 0$$

$$Gini(\text{outlook} = \text{sunny} \,\&\text{temperature} = \text{cool}) = 1 - \left(\frac{1}{1}\right)^2 - \left(\frac{0}{1}\right)^2 = 0$$

$$Gini(\text{outlook} = \text{sunny} \,\&\, \text{temperature} = \text{mild}) = 1 - \left(\frac{1}{2}\right)^2 - \left(\frac{1}{2}\right)^2 = 0.5$$

$$Gini(\text{outlook} = \text{sunny} \,\&\, \text{temperature}) = \frac{2}{5} * 0 + \frac{1}{5} * 0 + \frac{2}{5} * 0.5 = 0.2$$

- **Chỉ số Gini cho Humidity khi outlook = sunny**

Humidity	Yes	No	# Instances
high	0	3	3
Normal	2	0	2

$$Gini(\text{outlook} = \text{sunny} \,\&\, \text{humidity} = \text{high}) = 1 - \left(\frac{0}{3}\right)^2 - \left(\frac{3}{3}\right)^2 = 0$$

$$Gini(\text{outlook} = \text{sunny} \,\&\text{humidity} = \text{normal}) = 1 - \left(\frac{2}{2}\right)^2 - \left(\frac{0}{2}\right)^2 = 0$$

$$Gini(\text{outlook} = \text{sunny} \,\&\, \text{humidity}) = \frac{3}{5} * 0 + \frac{2}{5} * 0 = 0$$

- **Chỉ số Gini cho Wind khi outlook = sunny**

wind	Yes	No	# Instances
weak	1	2	3
strong	1	1	2

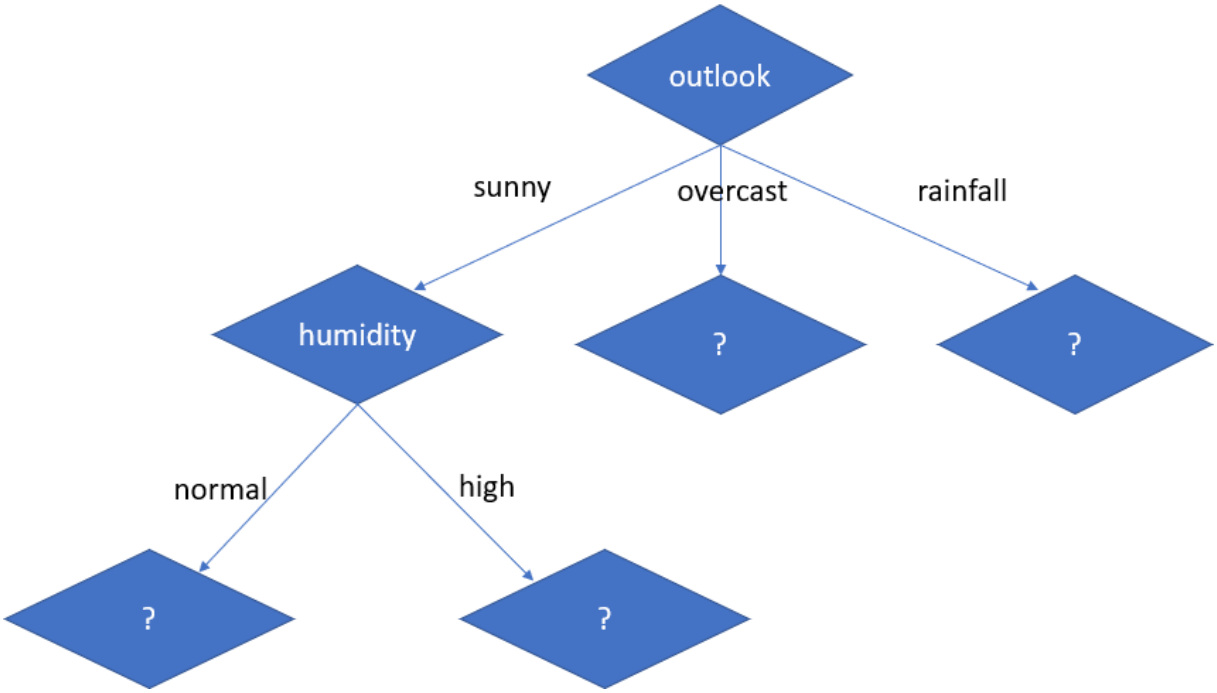
$$Gini(\text{outlook} = \text{sunny} \,\&\, \text{wind} = \text{weak}) = 1 - \left(\frac{1}{3}\right)^2 - \left(\frac{2}{3}\right)^2 = 0.44$$

$$Gini(\text{outlook} = \text{sunny} \,\&\text{wind} = \text{strong}) = 1 - \left(\frac{1}{2}\right)^2 - \left(\frac{1}{2}\right)^2 = 0.5$$

$$Gini(\text{outlook} = \text{sunny} \,\&\, \text{wind}) = \frac{3}{5} * 0.44 + \frac{2}{5} * 0.5 = 0.466$$

Features	Gini Index
temperature	0.2
humidity	0
wind	0.466

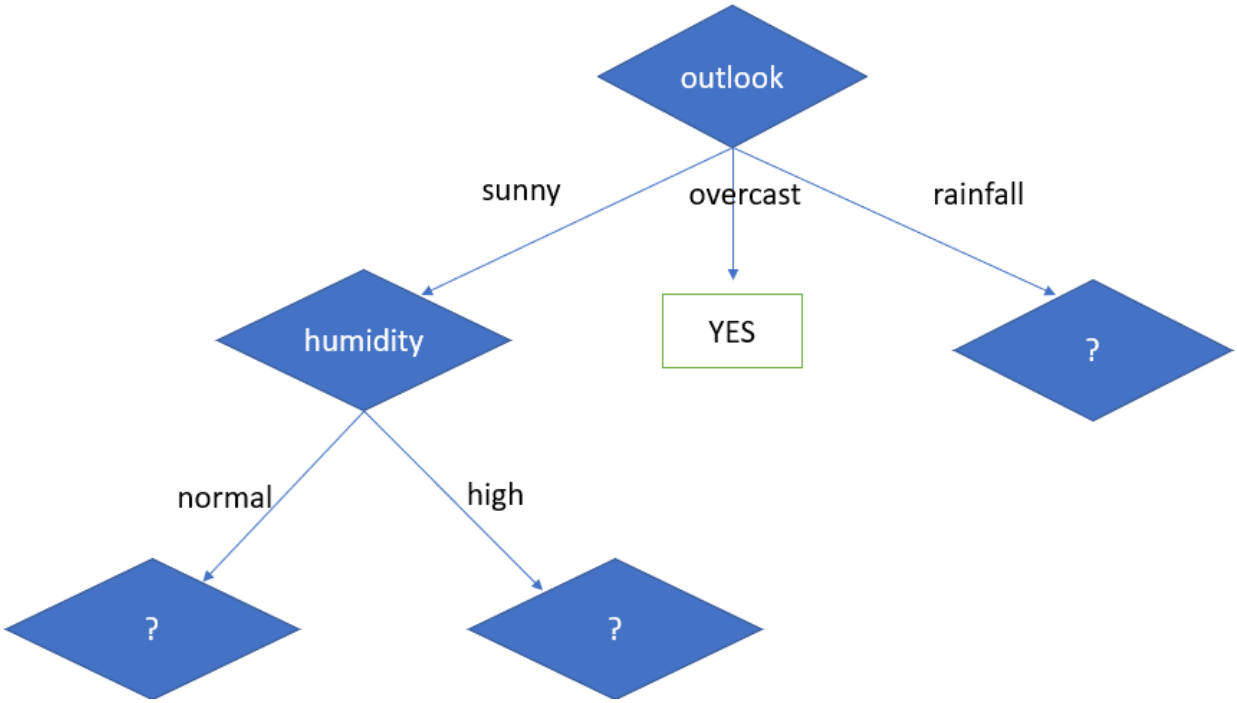
Ta thấy **Humidity** có giá trị nhỏ nhất vì thế node tiếp theo là **Humidity**



Trường hợp overcast:

Day	outlook	temperature	humidity	wind	decision
3	overcast	hot	high	weak	Yes
7	overcast	cool	normal	strong	Yes
12	overcast	mild	high	strong	Yes
13	overcast	hot	normal	weak	Yes

Ta thấy trong trường hợp **overcast** thì quyết định luôn là **YES** (tất cả các thuộc tính thuộc cùng 1 lớp) nên đến đã đến node lá.



Trường hợp Humidity

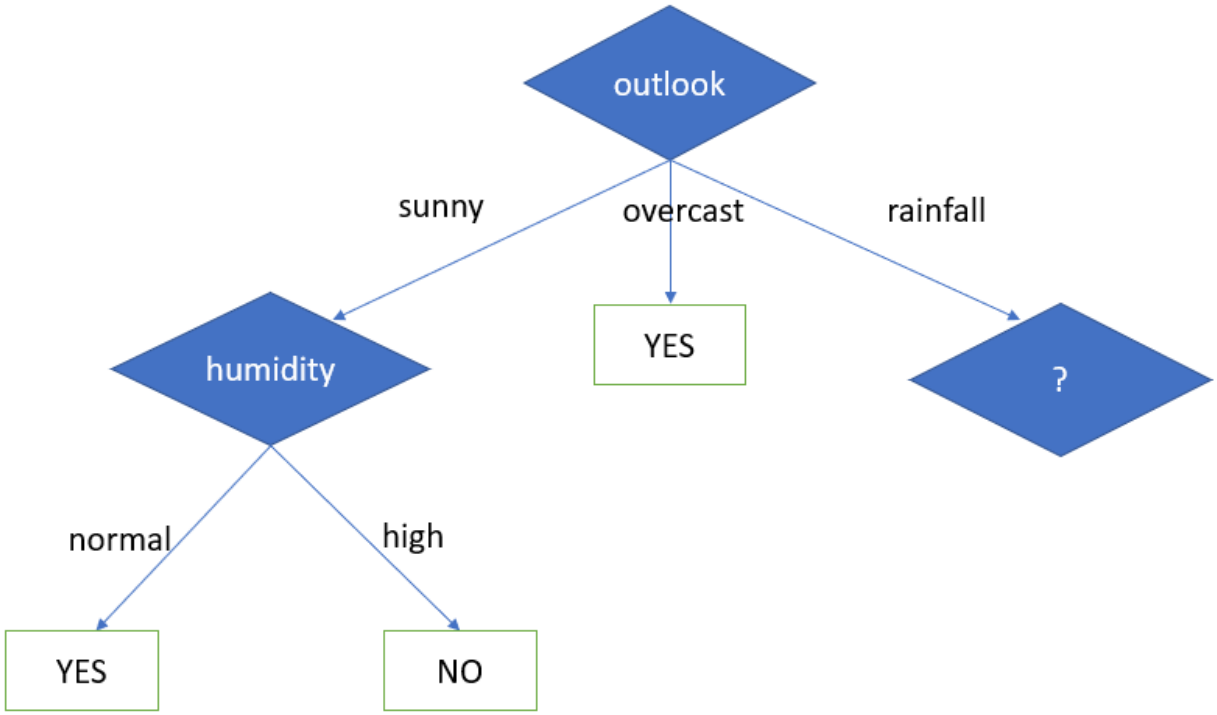
- **Humidity = high**

Day	outlook	temperature	humidity	wind	decision
1	sunny	hot	high	weak	No
2	sunny	hot	high	strong	No
8	sunny	mild	high	weak	No

- **Humidity = Normal**

Day	outlook	temperature	humidity	wind	decision
9	sunny	cool	normal	weak	Yes
11	sunny	mild	normal	strong	Yes

Từ 2 bảng trên ta luôn thấy **High Humidity** thì quyết định luôn là **NO**, còn **Normal Humidity** thì quyết định luôn là **YES**, vậy nên có 2 node lá



Trường hợp outlook = rainfall

Day	outlook	temperature	humidity	wind	Decision
4	rain	mild	high	weak	Yes
5	rain	cool	normal	weak	Yes
6	rain	cool	normal	strong	No
10	rain	mild	normal	weak	Yes
14	rain	mild	high	strong	No

- **Chỉ số Gini cho Temperature khi outlook = rainfall**

temperature	Yes	No	# Instances
cool	1	1	2
mild	2	1	3

$$Gini(\text{outlook} = \text{rainfall} \,\&\, \text{temp} = \text{Cool}) = 1 - \left(\frac{1}{2}\right)^2 - \left(\frac{1}{2}\right)^2 = 0.5$$

$$Gini(\text{outlook} = \text{rainfall} \,\&\, \text{temp} = \text{Mild}) = 1 - \left(\frac{2}{3}\right)^2 - \left(\frac{1}{3}\right)^2 = 0.444$$

$$Gini(\text{outlook} = \text{rainfall} \,\&\, \text{temp}) = \frac{2}{5} * 0.5 + \frac{3}{5} * 0.444 = 0.466$$

- **Chỉ số Gini cho Humidity khi outlook = rainfall**

humidity	Yes	No	# Instances
high	1	1	2
normal	2	1	3

$$Gini(\text{outlook} = \text{rainfall and humidity} = \text{high}) = 1 - \left(\frac{1}{2}\right)^2 - \left(\frac{1}{2}\right)^2 = 0.5$$

$$Gini(\text{outlook} = \text{rainfall and humidity} = \text{normal}) = 1 - \left(\frac{2}{3}\right)^2 - \left(\frac{1}{3}\right)^2 = 0.444$$

$$Gini(\text{outlook} = \text{rainfall \& temp}) = \frac{2}{5} * 0.5 + \frac{3}{5} * 0.444 = 0.466$$

- **Chỉ số Gini cho Wind khi outlook = rainfall**

wind	Yes	No	# Instances
weak	3	0	3
strong	0	2	2

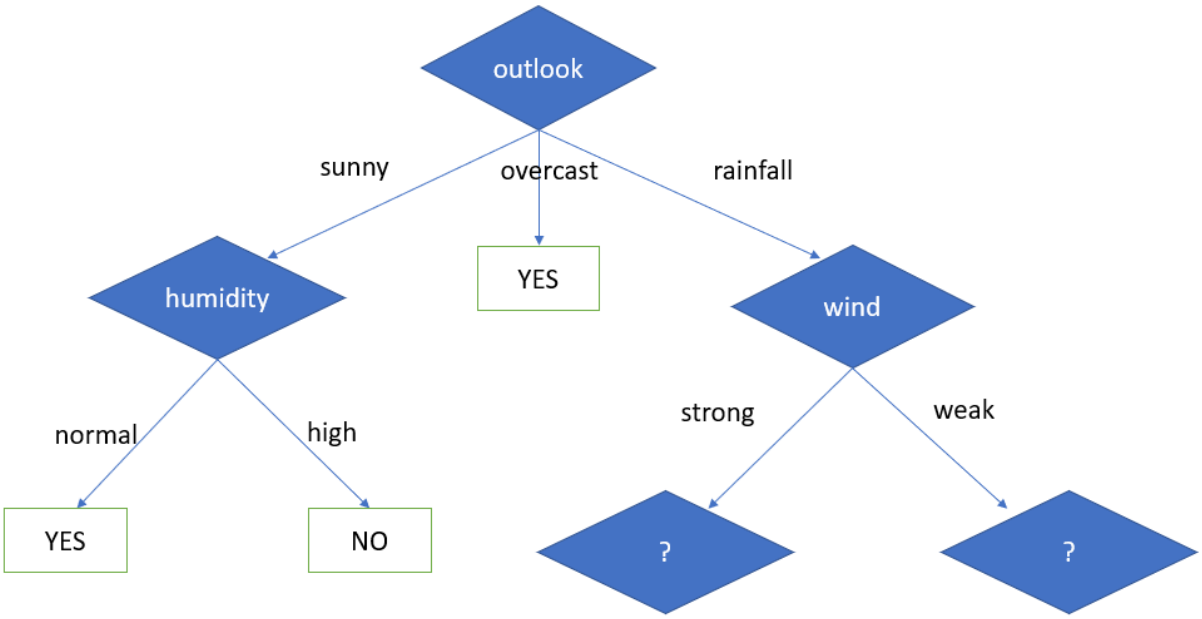
$$Gini(\text{outlook} = \text{rainfall \& wind} = \text{weak}) = 1 - \left(\frac{3}{3}\right)^2 - \left(\frac{0}{3}\right)^2 = 0$$

$$Gini(\text{outlook} = \text{rainfall \& wind} = \text{strong}) = 1 - \left(\frac{0}{2}\right)^2 - \left(\frac{2}{2}\right)^2 = 0$$

$$Gini(\text{outlook} = \text{rainfall \& and wind}) = \frac{3}{5} * 0 + \frac{2}{5} * 0 = 0$$

Features	Gini Index
temperature	0.466
humidity	0.466
wind	0

Giá trị GINI của **Wind** là nhỏ nhất nên node tiếp theo là **Wind**



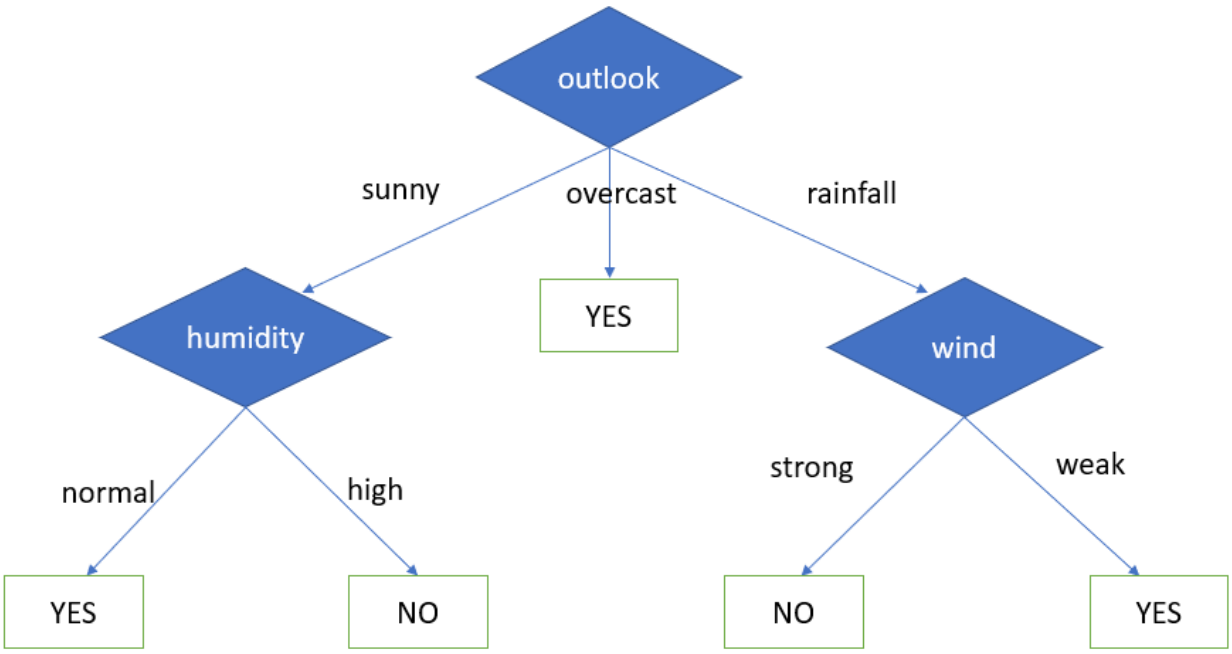
Trường hợp wind = strong

Day	outlook	temperature	humidity	wind	decision
6	rainfall	cool	normal	strong	No
14	rainfall	mild	high	strong	No

Trường hợp wind = weak

Day	outlook	temperature	humidity	wind	decision
4	rainfall	mild	high	weak	Yes
5	rainfall	cool	normal	weak	Yes
10	rainfall	mild	normal	weak	Yes

Dựa vào 2 bảng trên thì ta thấy nếu **wind: strong** thì quyết định là **NO** ngược lại nếu **wind: weak** thì quyết đinh là **YES** => có 2 node lá tiếp theo.



c) Thuật toán ID3

Thuật toán ID3 (Iterative Dichotomiser 3) là giải thuật lâu đời được tạo ra bởi Ross Quinlan nhằm xây dựng cây quyết định phù hợp từ một bộ dữ liệu, áp dụng cho bài toán Phân loại (Classification) mà tất các các thuộc tính để ở dạng category. Thuật toán ID3 dùng Entropy function và Information gain để kiểm tra kết quả, có thể tóm tắt các bước làm của ID3 như sau:

- Tính Entropy cho tập dữ liệu.
- Đối với mỗi feature. Tính toán entropy cho tất cả các giá trị phân loại của nó. Tính information gain cho từng feature.
- Tìm feature có information gain lớn nhất.
- Lặp lại cho đến khi tạo ra thành công cây quyết định.

Entropy: Là thước đo về độ hỗn loạn của thông tin. *(là một khái niệm được lấy trong môn vật lý cụ thể là môn nhiệt động lực được đo bằng tổng động năng của các hạt bên trong một khối vật chất, Entropy càng cao thì độ chuyển động của các hạt càng cao hay nói cách khác là các hạt chuyển động càng hỗn loạn dẫn đến việc rất khó để nhận biết được 1 hạt đang ở vị trí nào).*

Hay Entropy còn được hiểu là khả năng mà đoán đúng được thông tin.

VD: Trong một cuộc bầu cử, ban đầu khi chưa phiếu bầu chưa được kiểm định thì độ hỗn loạn của các phiếu bầu cao dẫn đến khả năng đoán chính xác phiếu

bầu cho ai rất thấp, khi số lượng phiếu được kiểm định xong tăng lên thì Entropy cũng sẽ tăng và Entropy = 0 khi tất cả các phiếu đã được kiểm định.

Entropy

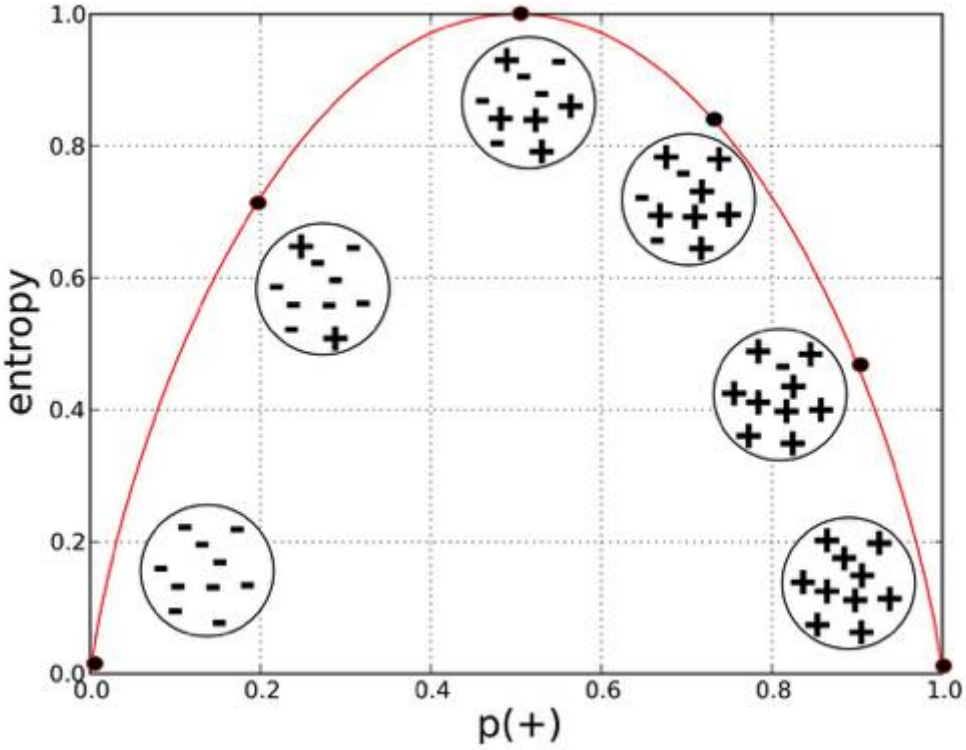
$$E(S) = \sum_{i=1}^{c} -p_i log_2 p_i$$

Average Information

$$I(Attribute)\frac{p_i + n_i}{p + n}E(S)$$

Information Gain (Thước đo độ giảm của Entropy khi chia nhỏ dữ liệu).

$$Gain = E(S) - I(Attribute)$$



Nhận xét Entropy càng cao thì mức độ rối loạn của dữ liệu càng lớn.

Sử dụng thuật toán ID3 để tạo cây quyết đinh trên tập dữ liệu thời tiết

Day	outlook	temperature	humidity	wind	Decision
1	sunny	hot	high	weak	No
2	sunny	hot	high	strong	No
3	overcast	hot	high	weak	Yes
4	rainfall	mild	high	weak	Yes
5	rainfall	cool	normal	weak	Yes
6	rainfall	cool	normal	strong	No
7	overcast	cool	normal	wtrong	Yes
8	sunny	mild	high	weak	No
9	sunny	cool	normal	weak	Yes
10	rainfall	mild	normal	weak	Yes
11	sunny	mild	normal	strong	Yes
12	overcast	mild	high	strong	Yes
13	overcast	hot	normal	weak	Yes
14	rainfall	mild	high	strong	No

Tính Entropy tổng

$$E(S) = -\frac{5}{14}log_2\frac{5}{14} - \frac{9}{14}log_2\frac{9}{14} = 0.94$$

Tính Entropy cho mỗi thuộc tính

- **Outlook**

outlook	Yes	No	Instances
sunny	2	3	5
overcast	4	0	4
rainfall	3	2	5

$$\text{E(Outlook} = \text{sunny)} = -\frac{2}{5}log_2\frac{2}{5} - \frac{3}{5}log_2\frac{3}{5} = 0.971$$

$$\text{E(Outlook} = \text{overcast)} = -1log_2 1 - 0log_2 0 = 0$$

$$\text{E(Outlook} = \text{rainfall)} = -\frac{3}{5}log_2\frac{3}{5} - \frac{2}{5}log_2\frac{2}{5} = 0.971$$

Tính Average Information Entropy

$$I(Outlook) = \frac{3+2}{9+5} * 0971 + \frac{2+3}{9+5} * 0.971 + \frac{4+0}{9+5} * 0 = 0.693$$

Tính Gain:

$$Gain(Outlook) = 0.94 - 0.693 = 0.247$$

- **Temperature**

Temperature	p	n	Entropy
Hot	2	2	1
Mild	4	2	0.918
Cool	3	1	0.811

Tính Average Information Entropy

$$I(Temperature) = \frac{2+2}{9+5} * 1 + \frac{4+2}{9+5} * 0.918 + \frac{3+1}{9+5} * 0.811 = 0.911$$

Tính Gain:

$$Gain(Temperature) = 0.94 - 0.911 = 0.029$$

- **Humidity**

Humidity	p	n	Entropy
High	3	4	0.985
Normal	6	1	0.591

Tính Average Information Entropy

$$I(Humidity) = \frac{3+4}{9+5} * 0.985 + \frac{6+1}{9+5} * 0.591 = 0.788$$

Tính Gain:

$$Gain(Humidity) = 0.94 - 0.788 = 0.152$$

- **Windy**

Windy	p	n	Entropy

Strong	3	3	1
Weak	6	2	0.811

Tính Average Information Entropy

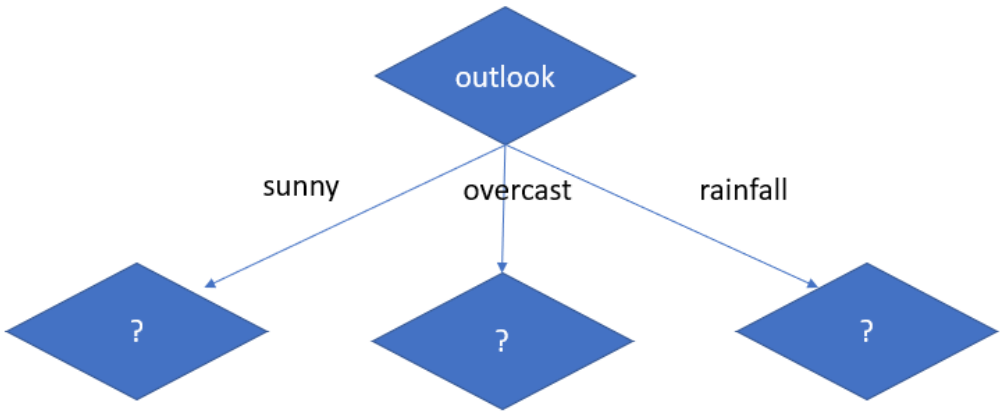
$$I(Windy) = \frac{3+3}{9+5} * 1 + \frac{6+2}{9+5} * 0.811 = 0.892$$

Tính Gain:

$$Gain(Windy) = \ 0.94 - 0.892 = 0.048$$

Attributes	Gain
Outlook	0.247
Temperature	0.029
Humidity	0.152
Windy	0.048

Node root: Outlook (có giá trị Gain nhỏ nhất).



- **Tính Entropy khi Outlook = Sunny**

Outlook	Temperature	Humidity	Windy	Decision
Sunny	Hot	High	Weak	No
Sunny	Hot	High	Strong	No

Sunny	Mild	High	Weak	No
Sunny	Cool	Normal	Weak	Yes
Sunny	Mild	Normal	Strong	Yes

outlook	p	n	Entropy
sunny	2	3	

$$E(S_{Sunny}) = -\frac{2}{5}log_2\frac{2}{5} - \frac{3}{5}log_2\frac{3}{5} = 0.971$$

- **Tính Entropy cho Humidity khi Outlook = Sunny**

Outlook	Humidity	Decision
Sunny	High	No
Sunny	High	No
Sunny	High	No
Sunny	Normal	Yes
Sunny	Normal	Yes

Temperature	p	n	Entropy
High	0	3	0
Normal	2	0	0

$$I(Humidity\ \&outlook = sunny) = 0$$

$$Gain(Humidity\ \&\ outlook = sunny) = 0.971 - 0 = 0.971$$

- **Tính Entropy cho Windy khi Outlook = Sunny**

Outlook	Temperature	Decision
Sunny	Strong	No
Sunny	Strong	Yes
Sunny	Weak	No
Sunny	Weak	No
Sunny	Weak	Yes

Temperature	p	n	Entropy
Strong	1	1	1
Weak	1	2	0.918

$$I(Windy\ \&\ outlook = sunny) = 0.951$$

$$Gain(Windy\ \&\ outlook = sunny) = 0.971 - 0.951 = 0.020$$

- **Tính Entropy cho Temperature khi Outlook = Sunny**

Outlook	Temperature	Decision
Sunny	Cool	Yes
Sunny	Hot	No
Sunny	Hot	No
Sunny	Mild	No
Sunny	Mild	Yes

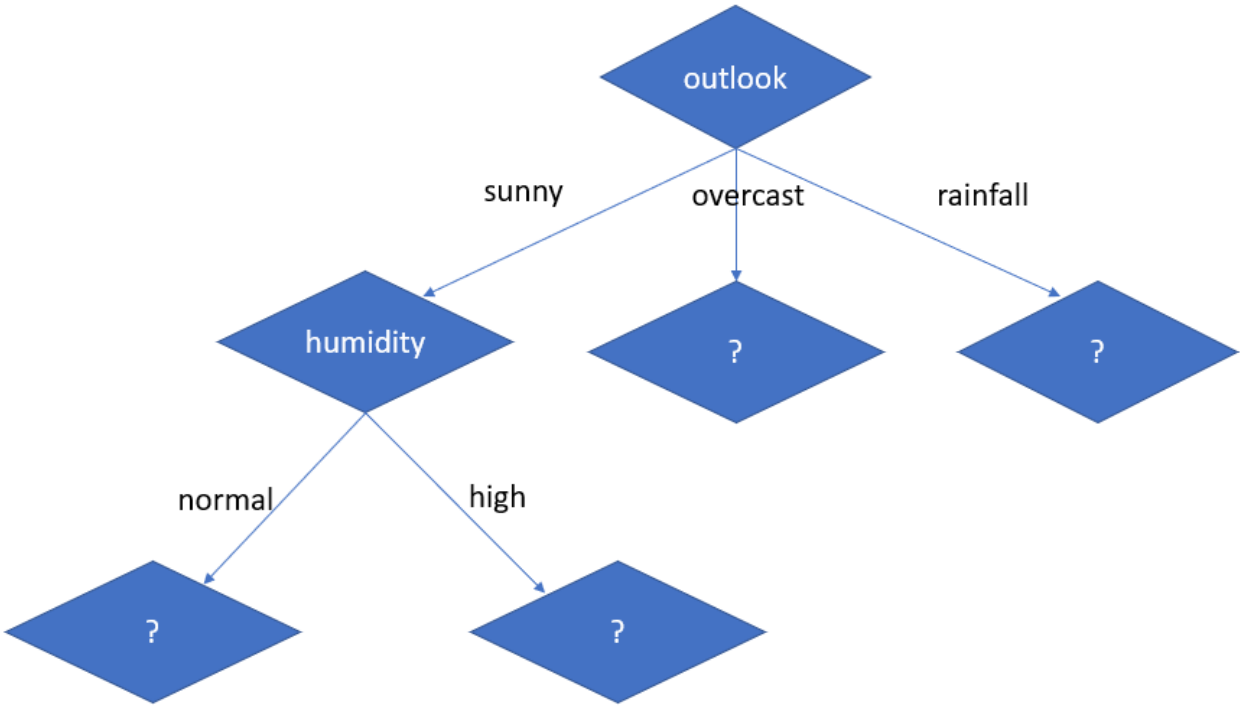
Temperature	p	n	Entropy
Cool	1	0	0
Hot	0	2	0
Mild	1	1	1

$$I(Temperature\ \&\ outlook = sunny) = 0.4$$

$$Gain(Temperature\ \&\ outlook = sunny) = 0.971 - 0.4 = 0.571$$

Attributes	Gain
Temperature	0.571
Humidity	0.971
Windy	0.02

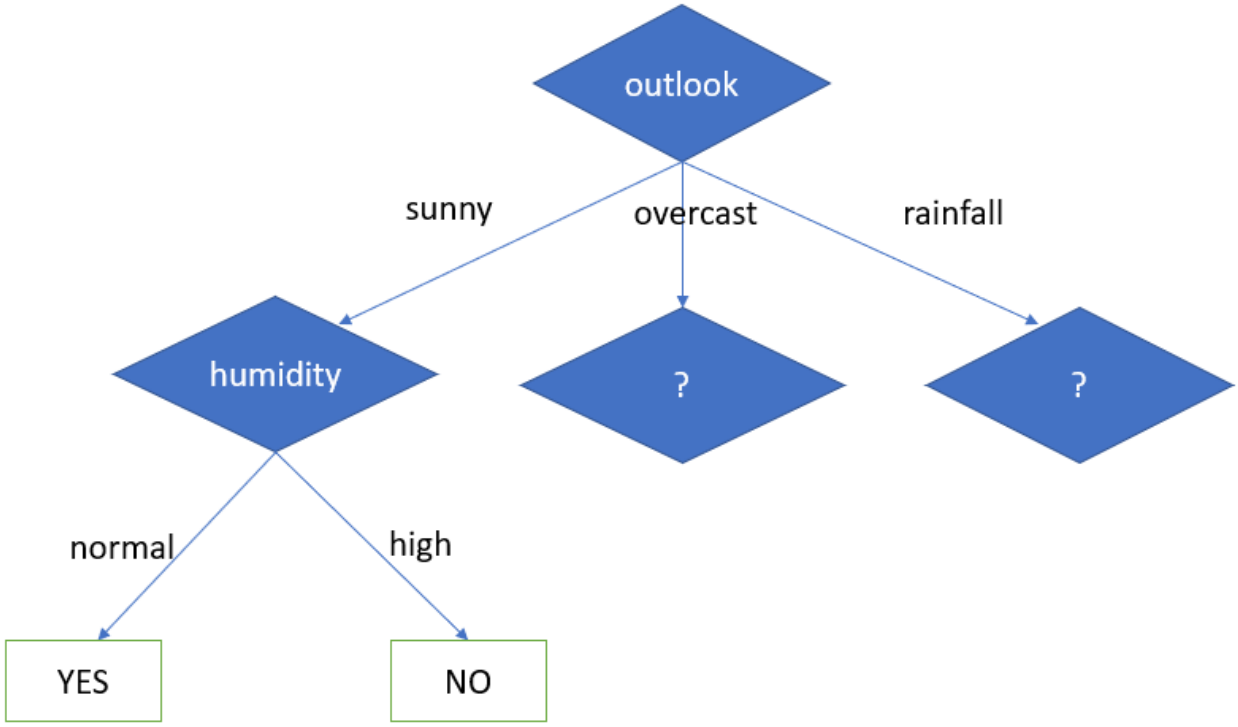
Humidity được chọn làm node tiếp theo (có giá trị Gain lớn nhất).



- Xét trường hợp Humidity khi outlook = sunny

Outlook	Humidity	Decision
Sunny	High	No
Sunny	High	No
Sunny	High	No
Sunny	Normal	Yes
Sunny	Normal	Yes

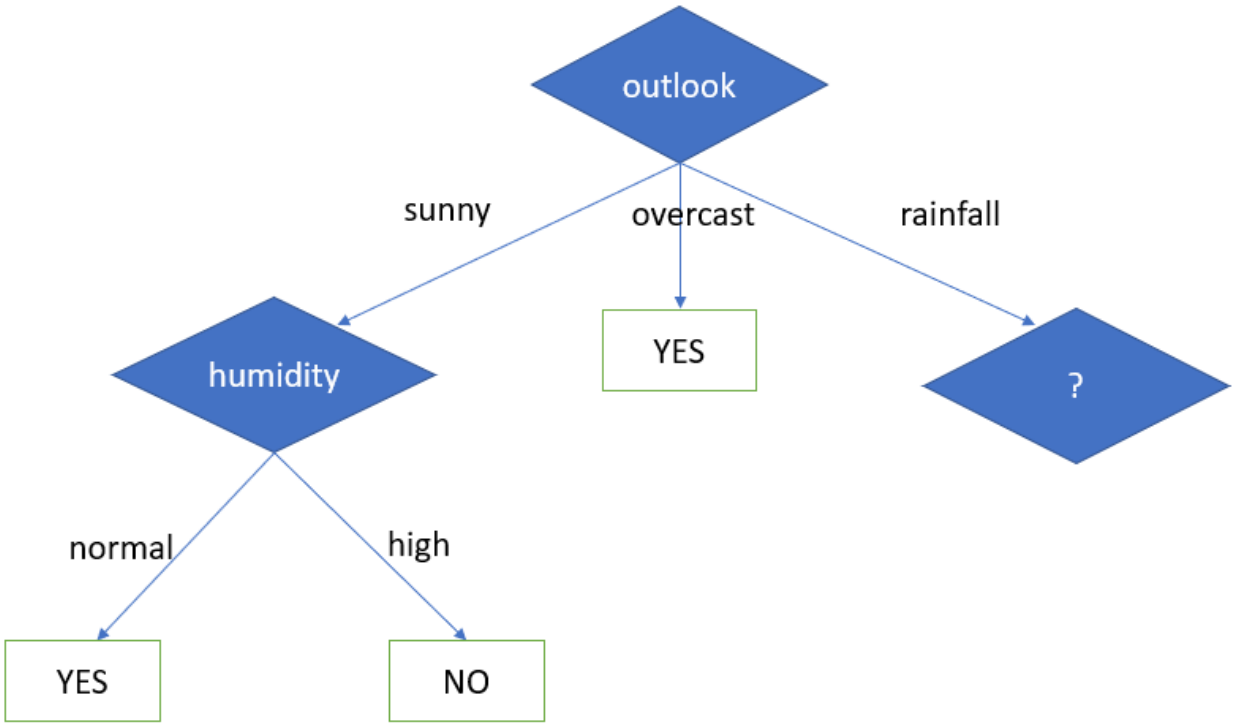
Khi đó ta thấy ứng với Humidity = Hight thì quyết định là No, Humidity = Normal thì quyết định là Yes. Do đó chúng ta có 2 node lá tiếp theo



- Trường hợp outlook = overcast

Outlook	Temperature	Humidity	Windy	Decision
Overcast	Hot	High	Weak	Yes
Overcast	Cool	Normal	Strong	Yes
Overcast	Mild	High	Strong	Yes
Overcast	Hot	Normal	Weak	Yes

Khi outlook = overcast thì quyết đinh luôn là YES nên ta có node lá



- **Trường hợp Outlook = Rainfall**

Outlook	Temperature	Humidity	Windy	Decision
Rainfall	Mild	High	Weak	Yes
Rainfall	Cool	Normal	Weak	Yes
Rainfall	Cool	Normal	Strong	No
Rainfall	Mild	Normal	Weak	Yes
Rainfall	Mild	High	Strong	No

outlook	p	n	Entropy
rainfall	3	2	

$$E(S_{Rainy}) = -\frac{3}{5}log_2\frac{3}{5} - \frac{2}{5}log_2\frac{2}{5} = 0.971$$

- **Tính Entropy cho Humidity khi Outlook = Rainfall**

outlook	Humidity	Decision
Rainfall	High	Yes
Rainfall	High	No
Rainfall	Normal	Yes
Rainfall	Normal	No
Rainfall	Normal	Yes

Attribute	p	n	Entropy
Strong	1	1	1
Weak	2	1	0.918

$I(Humidity \text{ \& Outlook } = \text{ Rainfall}) = 0.951$

$Gain(Humidity \text{ \& Outlook } = \text{ Rainfall}) = 0.971 - 0.951 = 0.020$

- **Tính Entropy cho Windy khi Outlook = Rainfall**

outlook	Windy	Decision
Rainfall	Strong	No
Rainfall	Strong	No
Rainfall	Weak	Yes
Rainfall	Weak	Yes
Rainfall	Weak	Yes

Attribute	p	n	Entropy
Strong	0	2	0
Weak	3	0	0

$I(Windy \text{ \& Outlook } = \text{ Rainfall}) = 0$

$Gain(Windy \text{ \& Outlook } = \text{ Rainfall}) = 0.971 - 0 = 0.971$

- **Tính Entropy cho Temperature khi Outlook = Rainfall**

outlook	Temperature	Decision
Rainfall	Mild	Yes
Rainfall	Cool	Yes
Rainfall	Cool	No

Rainfall	Mild	Yes
Rainfall	Mild	No

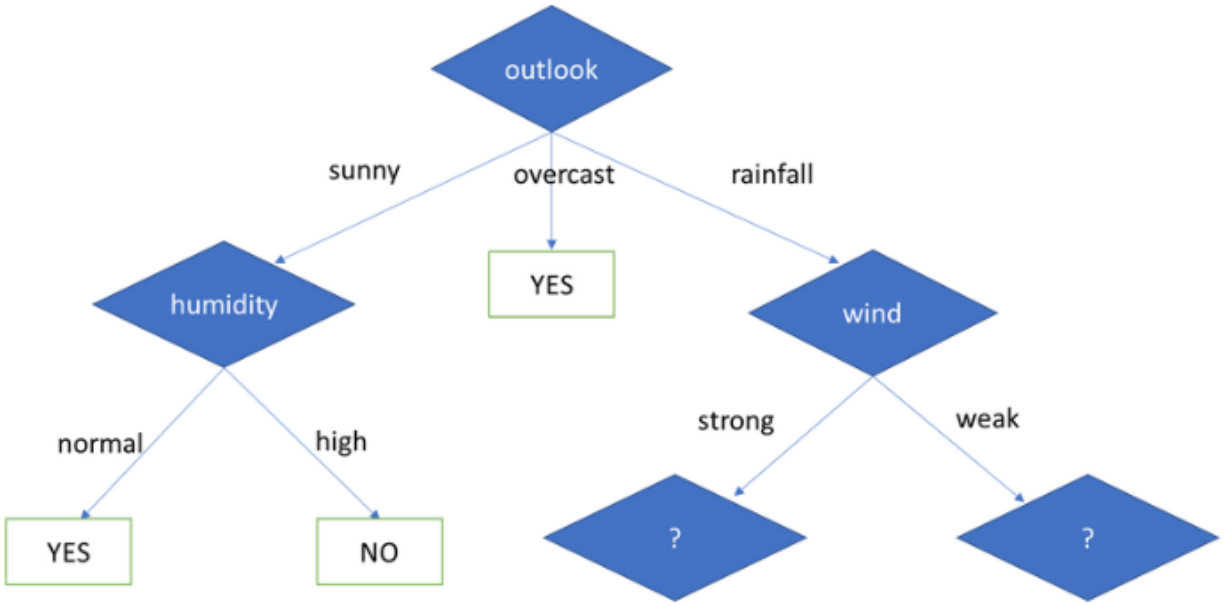
Attributes	p	n	Entropy
Cool	1	1	1
Mild	2	1	0.918

$$I(Temperature \text{ \& Outlook } = \text{ Rainfall }) = 0.951$$

$$Gain(Temperature \text{ \& Outlook } = \text{ Rainfall}) = 0.971 - 0.951 = 0.020$$

Attributes	Gain
Humidity	0.02
Windy	0.971
Temperature	0.02

Windy là node tiếp theo (Giá trị Gain lớn nhất)



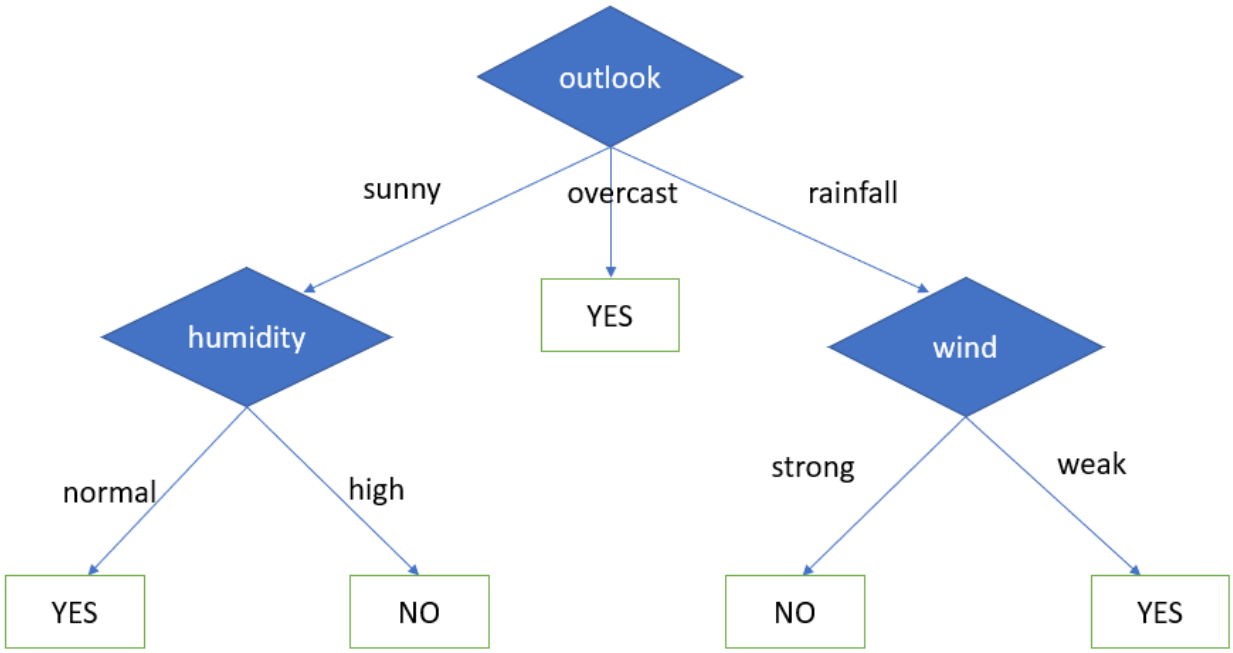
Trường hợp wind = strong & outlook = rainfall

outlook	temperature	humidity	wind	decision
rainfall	cool	normal	strong	No
rainfall	mild	high	strong	No

Trường hợp wind = weak & outlook = rainfall

outlook	temperature	humidity	wind	decision
rainfall	mild	high	weak	yes
rainfall	cool	normal	weak	yes
rainfall	mild	normal	weak	yes

Dựa vào 2 bảng trên thì ta thấy nếu **wind: strong** thì quyết định là **NO** ngược lại nếu **wind: weak** thì quyết đinh là **YES** => có 2 node lá tiếp theo



Tài liệu tham khảo

Lutes, J. (n.d.). *Towards Data Science.* Retrieved from https://towardsdatascience.com/entropy-and-information-gain-in-decision-trees-c7db67a3a293

MLMath.io. (n.d.). Retrieved from Math behind Decision Tree Algorithm: https://ankitnitjsr13.medium.com/math-behind-decision-tree-algorithm-2aa398561d6d

Sakkaf, Y. (n.d.). *Towards Data Science.* Retrieved from https://towardsdatascience.com/decision-trees-for-classification-id3-algorithm-explained-89df76e72df1

Trí tuệ nhân tạo. (n.d.). Retrieved from Cây Quyết Định (Decision Tree): https://trituenhantao.io/kien-thuc/decision-tree/