

DATA MINING

GROUP 8

NỘI DUNG

1. DATA MINING
TECHNIQUES

2. DECISION
TREE (ID3
ALGORITHM)

3. MODEL
EVALUATION

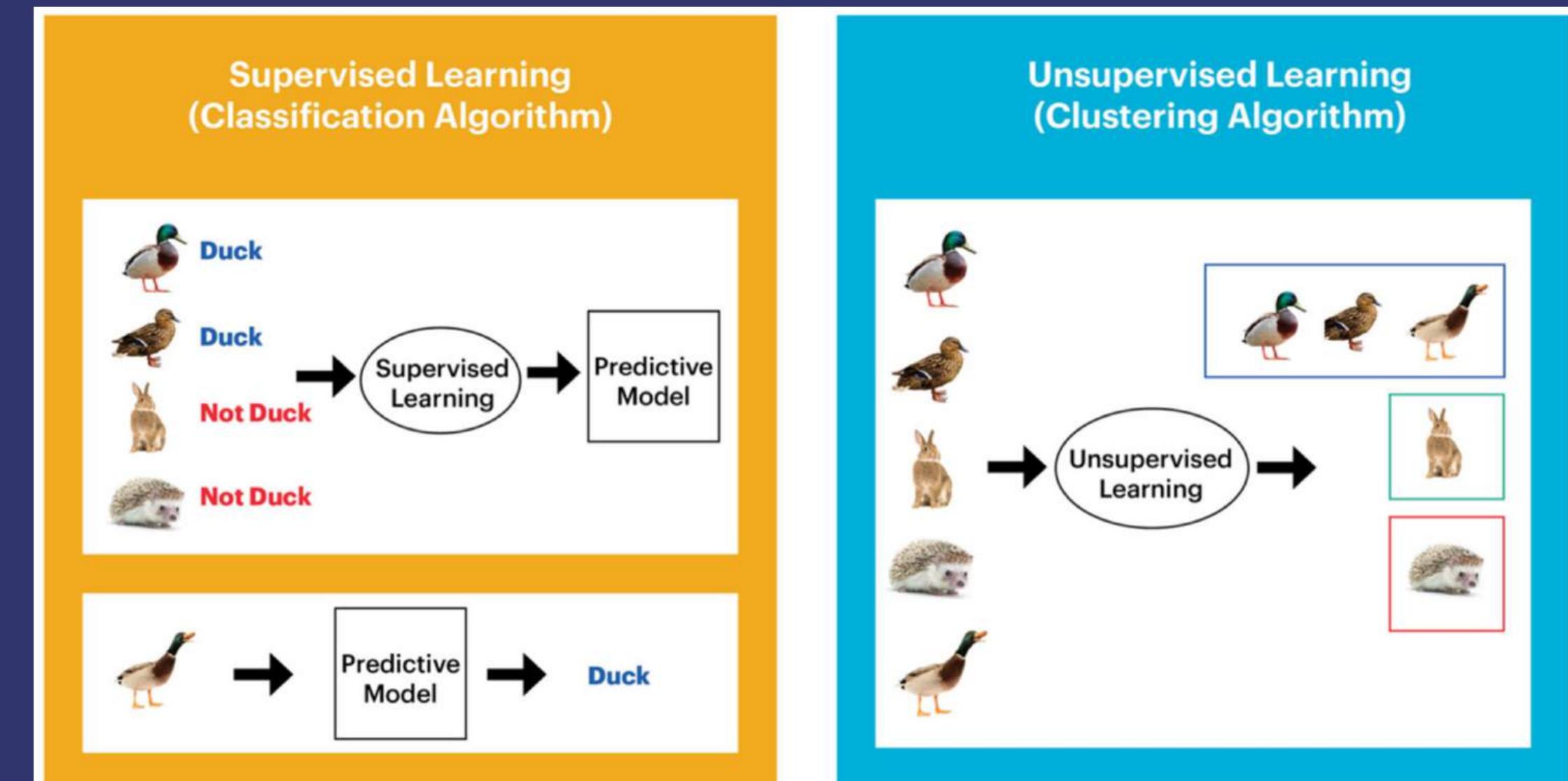
4. DEMO

DATA MINING TECHNIQUES

Kỹ thuật khai phá dữ liệu là các phương pháp và công cụ được sử dụng để khám phá thông tin ẩn trong các cơ sở dữ liệu lớn và phức tạp.

Có nhiều loại kỹ thuật khai phá dữ liệu, trong đó một số phổ biến:

- Phân lớp dữ liệu
- Phân cụm dữ liệu
- Phát hiện luật kết hợp



DATA MINING TECHNIQUES

1. PHÂN LỚP DỮ LIỆU (CLASSIFICATION)

Cho phép phân loại một đối tượng vào một hoặc một số lớp cho trước.

2. PHÂN CỤM DỮ LIỆU (CLUSTERING)

Tìm ra những mẫu dữ liệu có đặc tính tương tự nhau và phân chung vào các nhóm khác nhau.

3. PHÁT HIỆN LUẬT KẾT HỢP (ASSOCIATION RULE DISCOVERY)

Tìm ra các quy tắc (rules) xuất hiện thường xuyên trong tập dữ liệu.

DECISION TREE (ID3 ALGORITHM)

1.TỔNG QUAN

2.XÂY DỰNG
CÂY

3.THUẬT TOÁN
ID3

TỔNG QUAN VỀ ID3

Cây quyết định là cây mà mỗi nút biểu diễn một đặc trưng (tính chất), mỗi nhánh biểu diễn một quy luật và mỗi lá biểu diễn một kết quả (giá trị cụ thể hay một nhánh tiếp tục).

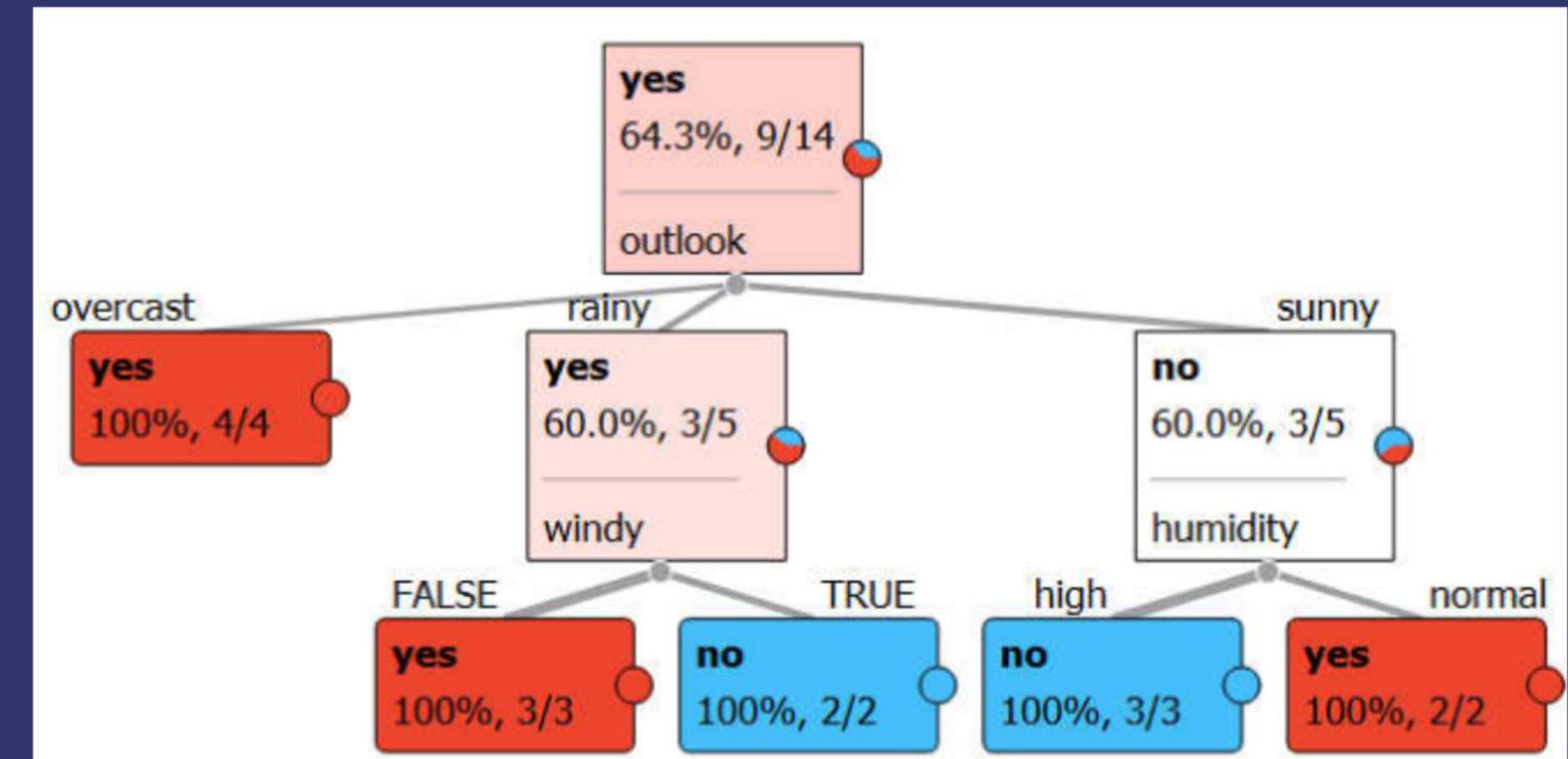
Cây quyết định có cấu trúc cây sao cho:

- Mỗi nút trong ứng với một phép kiểm tra trên một thuộc tính
- Mỗi nhánh biểu diễn kết quả phép kiểm tra
- Các nút lá biểu diễn các lớp hay các phân bối lớp
- Nút cao nhất trong cây là nút gốc

XÂY DỰNG CÂY

- Chọn thuộc tính “tốt nhất” theo một độ đo chọn lựa cho trước
- Mở rộng cây bằng cách thêm các nhánh mới cho từng giá trị thuộc tính
- Sắp xếp các mẫu nút lá
- Nếu các mẫu được phân lớp rõ thì Stop, nếu không lặp lại các bước 1-4 cho các nút lá
- Tỉa các nút lá không ổn định

KIỀU DOAN

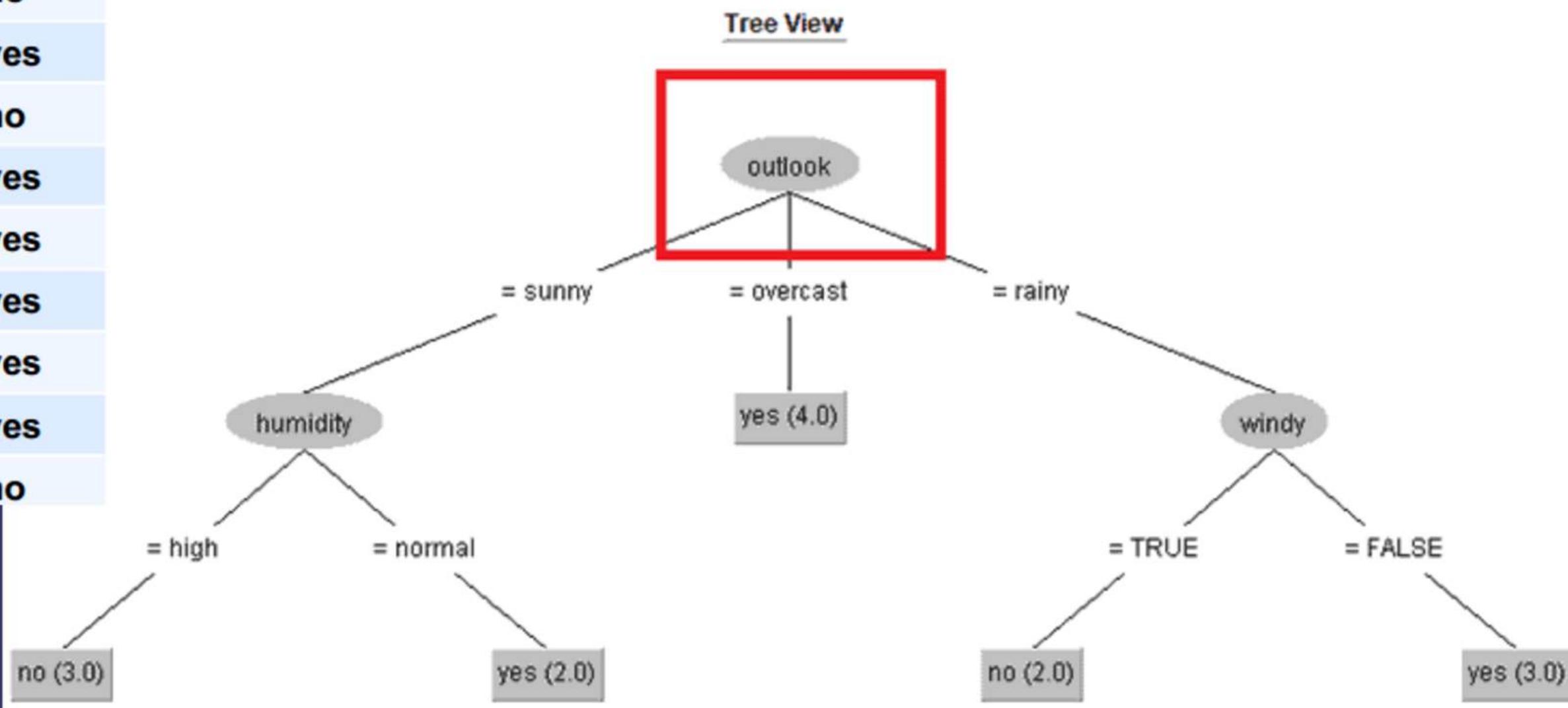


THUẬT TOÁN ID3

No.	Outlook	Temperature	Humidity	Windy	Play
1	sunny	hot	high	FALSE	no
2	sunny	hot	high	TRUE	no
3	overcast	hot	high	FALSE	yes
4	rainy	mild	high	FALSE	yes
5	rainy	cool	normal	FALSE	yes
6	rainy	cool	normal	TRUE	no
7	overcast	cool	normal	TRUE	yes
8	sunny	mild	high	FALSE	no
9	sunny	cool	normal	FALSE	yes
10	rainy	mild	normal	FALSE	yes
11	sunny	mild	normal	TRUE	yes
12	overcast	mild	high	TRUE	yes
13	overcast	hot	normal	FALSE	yes
14	rainy	mild	high	TRUE	no

KIỀU DOAN

ID3, J48 và C4.5 là các thuật toán phổ biến để xây dựng cây quyết định



THUẬT TOÁN ID3

- Thuật toán ID3 sử dụng độ lợi thông tin (IG - Information Gain) làm độ đo để chọn nút gốc phân tách cây.
- Gain được tính toán dựa trên Entropy trong lý thuyết thông tin.
- Node được chọn làm gốc là node có Gain lớn nhất.

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

S: tập dữ liệu huấn luyện

p_i: xác suất các mẫu thuộc lớp i

c: tập các giá trị của thuộc tính phân loại

$$\text{Gain}(S, A) = \text{Entropy}(S) - \sum_{v \in A} \frac{|S_v|}{|S|} * \text{Entropy}(S_v)$$

v: các giá trị của thuộc tính A

S_v: tập con của tập S với các mẫu thuộc tính A có giá trị v

|S_v|: số các mẫu thuộc tính S,

|S|: số mẫu của tập S

	outloo	temperatur	humidit	wind	pla
2	overcast	hot	high	FALSE	yes
3	overcast	cool	normal	TRUE	yes
4	overcast	mild	high	TRUE	yes
5	overcast	hot	normal	FALSE	yes
6	rainy	mild	high	FALSE	yes
7	rainy	cool	normal	FALSE	yes
8	rainy	cool	normal	TRUE	no
9	rainy	mild	normal	FALSE	yes
10	rainy	mild	high	TRUE	no
11	sunny	hot	high	FALSE	no
12	sunny	hot	high	TRUE	no
13	sunny	mild	high	FALSE	no
14	sunny	cool	normal	FALSE	yes
15	sunny	mild	normal	TRUE	yes

THUẬT TOÁN ID3

S có 9 yes (+), 5 no (-)



Entropy(S)=

$$-(9/14)\log_2(9/14) - (5/14)\log_2(5/14)$$

$$=0.940$$

Chú ý:

1. Entropy bằng 0 nếu tất cả các mẫu của S đều thuộc về cùng một lớp.

$$\text{Entropy}(S) = -1 \cdot \log_2(1) - 0 \cdot \log_2(0) = -1 \cdot 0 - 0 \cdot \log_2(0) = 0.$$

2. Entropy bằng 1 nếu số các mẫu bằng nhau ở các lớp

3. Nếu số mẫu khác nhau ở các lớp, entropy sẽ là (0, 1)

1	outloo	temperatur	humidit	wind	pla
2	overcast	hot	high	FALSE	yes
3	overcast	cool	normal	TRUE	yes
4	overcast	mild	high	TRUE	yes
5	overcast	hot	normal	FALSE	yes
6	rainy	mild	high	FALSE	yes
7	rainy	cool	normal	FALSE	yes
8	rainy	cool	normal	TRUE	no
9	rainy	mild	normal	FALSE	yes
10	rainy	mild	high	TRUE	no
11	sunny	hot	high	FALSE	no
12	sunny	hot	high	TRUE	no
13	sunny	mild	high	FALSE	no
14	sunny	cool	normal	FALSE	yes
15	sunny	mild	normal	TRUE	yes



THUẬT TOÁN ID3

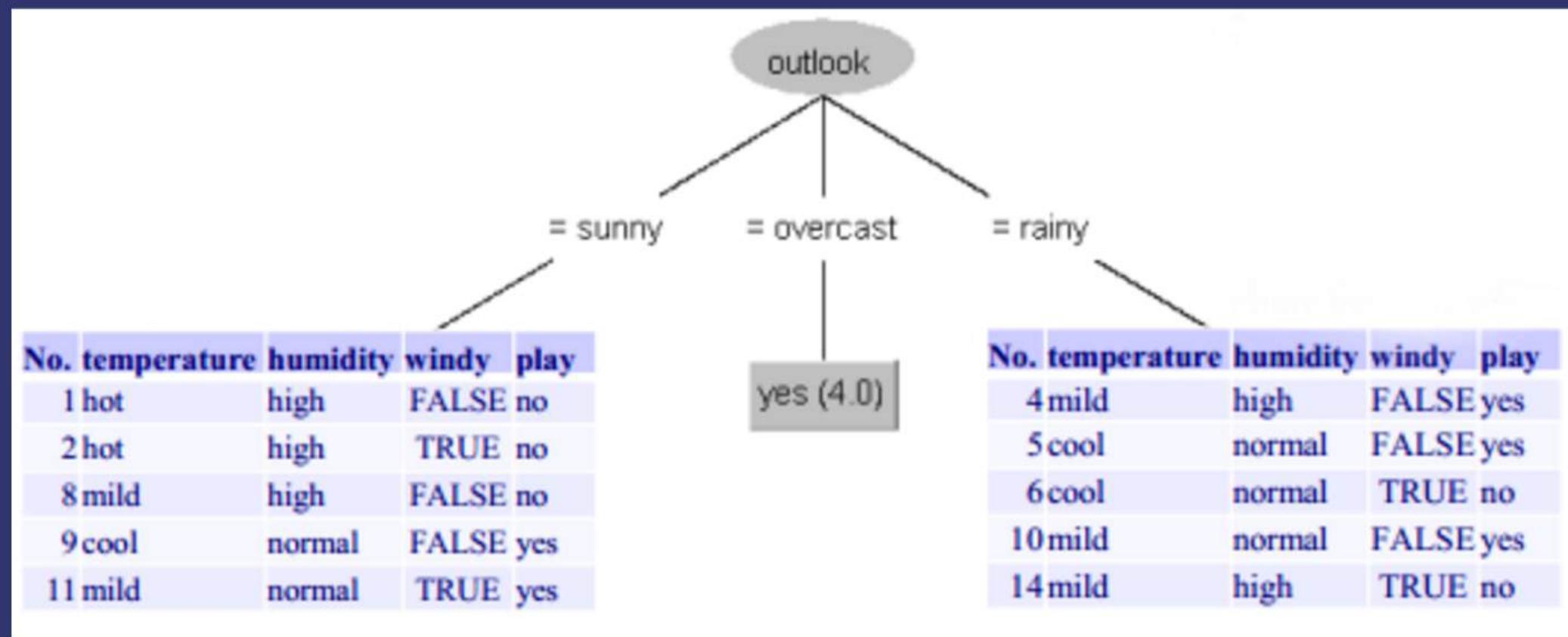
$$\text{Entropy}(S) = 0.940$$

Outlook = {sunny, overcast, rainy}
 $\sim S\{[2+/3-], [4+/0-], [3+/2-]\}$

$$\begin{aligned} \text{Entropy}(S_{\text{Outlook}}) &= \\ (\text{sunny}): & -2/5\log_2(2/5) - 3/5\log_2(3/5) \rightarrow 0.97 \\ (\text{overcast}): & -4/4\log_2(4/4) - 0/4\log_2(0/4) \rightarrow 0 \\ (\text{rainy}): & -3/5\log_2(3/5) - 2/5\log_2(2/5) \rightarrow 0.97 \end{aligned}$$

$$\begin{aligned} \text{Gain}(S, S_{\text{Outlook}}) &= 0.940 - [5/14 * (-2/5\log_2(2/5) - 3/5\log_2(3/5)) + 4/14 * 0 \\ &\quad + 5/14 * (-3/5\log_2(3/5) - 2/5\log_2(2/5))] = 0.246 \end{aligned}$$

THUẬT TOÁN ID3

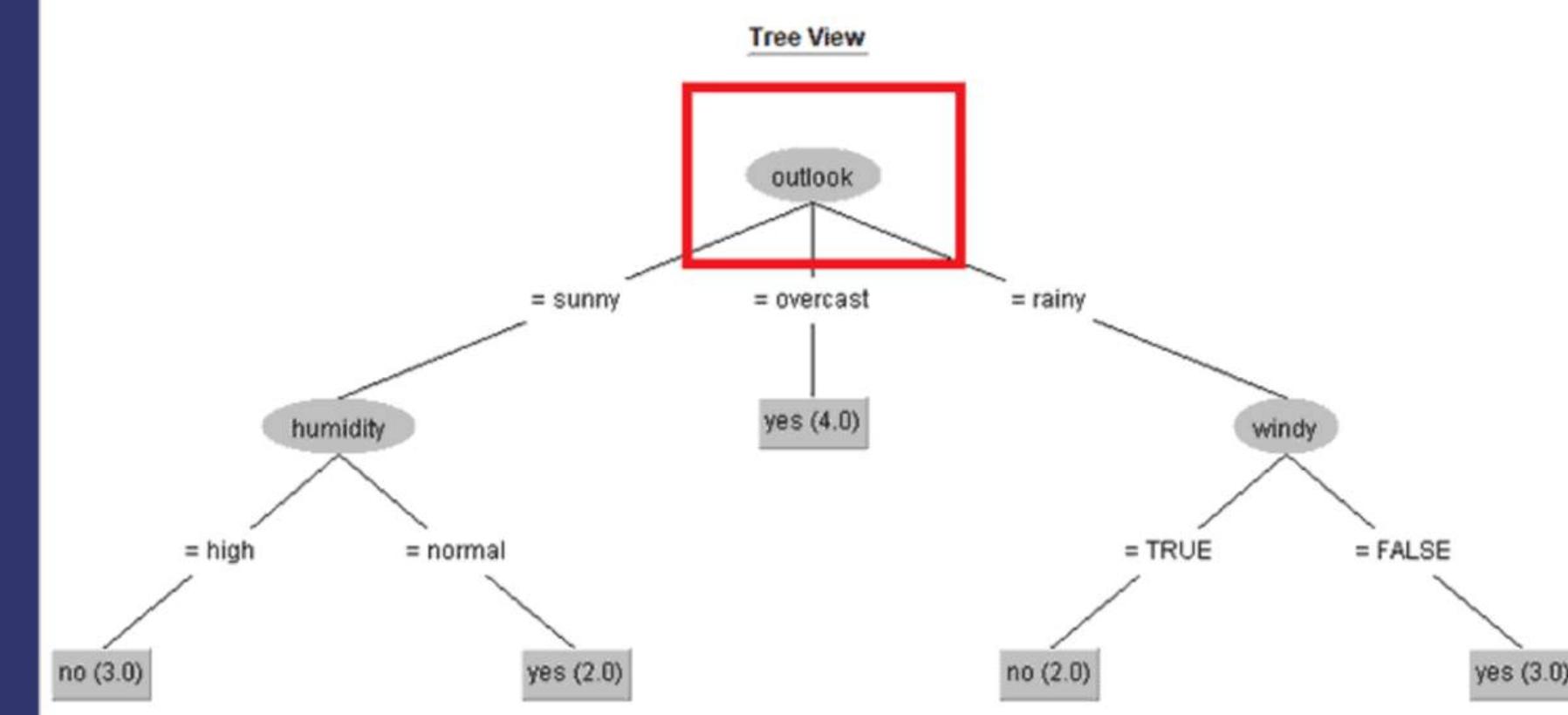


1	outloo	temperatur	humidit	wind	pla
2	overcast	hot	high	FALSE	yes
3	overcast	cool	normal	TRUE	yes
4	overcast	mild	high	TRUE	yes
5	overcast	hot	normal	FALSE	yes
6	rainy	mild	high	FALSE	yes
7	rainy	cool	normal	FALSE	yes
8	rainy	cool	normal	TRUE	no
9	rainy	mild	normal	FALSE	yes
10	rainy	mild	high	TRUE	no
11	sunny	hot	high	FALSE	no
12	sunny	hot	high	TRUE	no
13	sunny	mild	high	FALSE	no
14	sunny	cool	normal	FALSE	yes
15	sunny	mild	normal	TRUE	yes

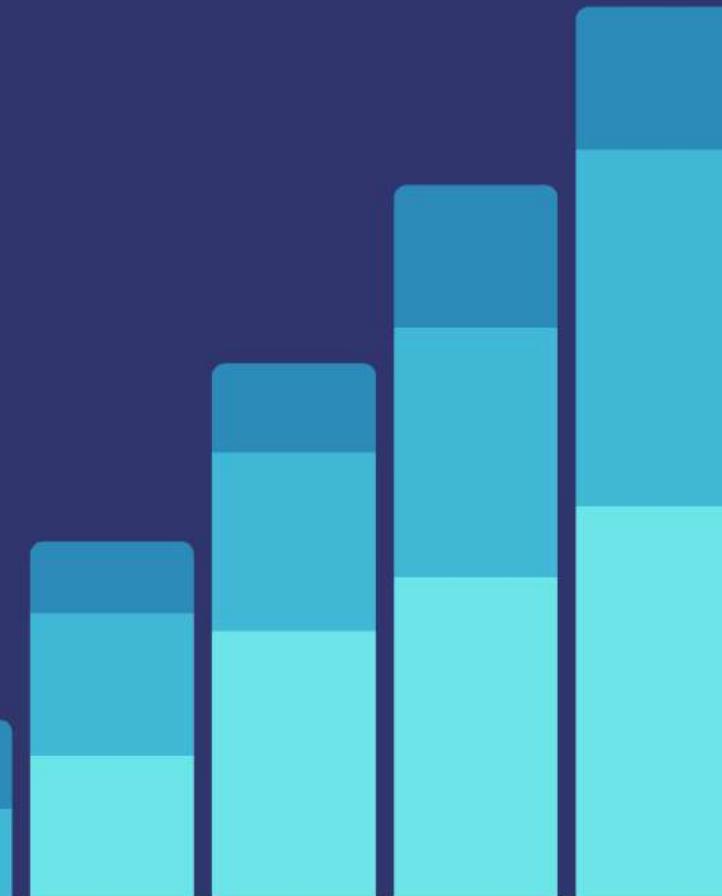
$$\begin{aligned} \text{Gain}(S, S_{\text{Outlook}}) &= 0,246 \\ \text{Gain}(S, S_{\text{Humidity}}) &= 0,151 \\ \text{Gain}(S, S_{\text{Windy}}) &= 0,048 \\ \text{Gain}(S, S_{\text{Temperature}}) &= 0,029 \end{aligned}$$



KIỀU DOAN



3. MODEL EVALUATION



Model evaluation : là quá trình **đánh giá tính hiệu quả** của một mô hình dữ liệu được xây dựng từ quá trình khai phá dữ liệu

Mục đích: đo lường độ chính xác, tính khả thi, độ tin cậy của mô hình trên dữ liệu mới để đưa ra những đánh giá chính xác về khả năng dự đoán của mô hình trên dữ liệu thực tế.

PHÂN CHIA TẬP DỮ LIỆU (SPLIT DATASET)

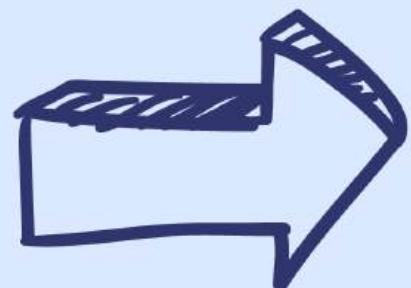


Là quá trình chia tập dữ liệu ban đầu thành hai hay nhiều tập con riêng biệt. Thông thường là chia 2 phần để sử dụng trong quá trình huấn luyện và kiểm tra mô hình dữ liệu.

Training set: đây thường là một tập dữ liệu có kích thước lớn, được dùng để xây dựng mô hình

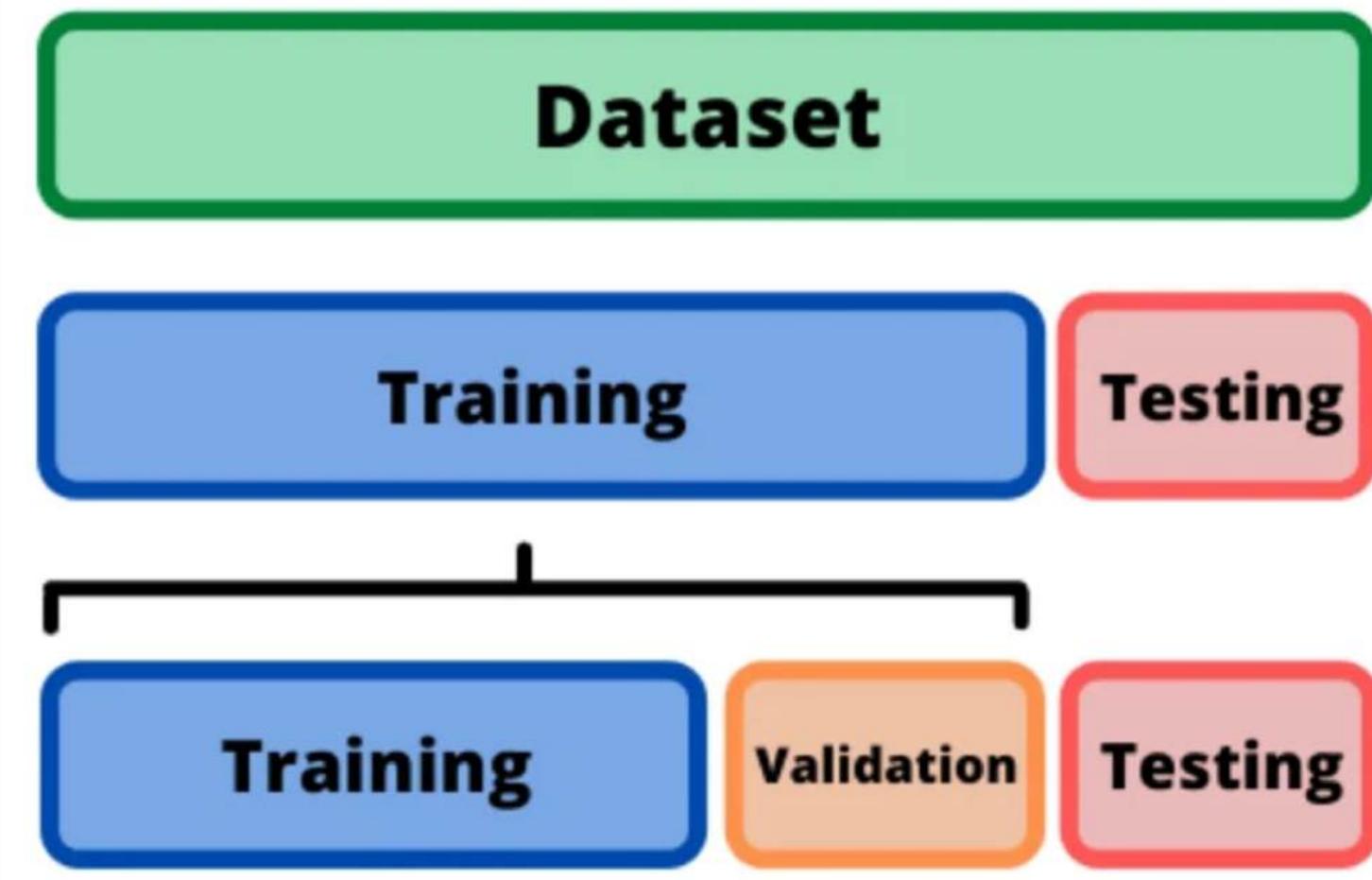
Gồm 2 phần:

- Input: sẽ là những dữ liệu đầu vào.
- Output: sẽ là những kết quả tương ứng với tập input.



con chó

VŨ THỊ HÀ



VALIDATION SET

- Để mô hình được đào tạo một cách bài bản và khoa học, nó cần phải được đánh giá định kỳ, thường xuyên và đó cũng chính là mục đích cụ thể nhất của bộ xác thực.
- Tập các giá trị input đi kèm với giá trị output và được dùng để kiểm thử độ chính xác của mô hình trong quá trình huấn luyện.

TESTING SET

- Tập dữ liệu dùng để test sau khi xây dựng mô hình để đánh giá mức độ chính xác của mô hình.
- Nếu đáp ứng được yêu cầu thì là đạt, không thì cần phải xem xét lại.
- Testing set chỉ gồm các giá trị input mà không có các giá trị output.

CÁCH PHÂN CHIA DỮ LIỆU

- **Phân chia ngẫu nhiên:** Tập dữ liệu ban đầu được chia ngẫu nhiên thành các tập con theo các tỉ lệ nhất định đảm bảo rằng mỗi mẫu dữ liệu có cơ hội ngang nhau để được chọn vào một trong hai tập.
- **Phân chia theo tỉ lệ:** Dữ liệu được chia sao cho tỷ lệ các lớp trong dữ liệu gốc được giữ nguyên trong mỗi phần chia
- **Phân chia theo thời gian:** Đối với dữ liệu có yếu tố thời gian, chẳng hạn như chuỗi thời gian hoặc dữ liệu giao dịch, phương pháp này phân chia dữ liệu dựa trên thời gian.

CÁCH PHÂN CHIA DỮ LIỆU

Kiểm định chéo (Cross-Validation)

Dữ liệu được chia thành nhiều “fold” và mỗi fold lần lượt được sử dụng làm Testing Set, trong khi các fold còn lại được sử dụng làm Training Set.



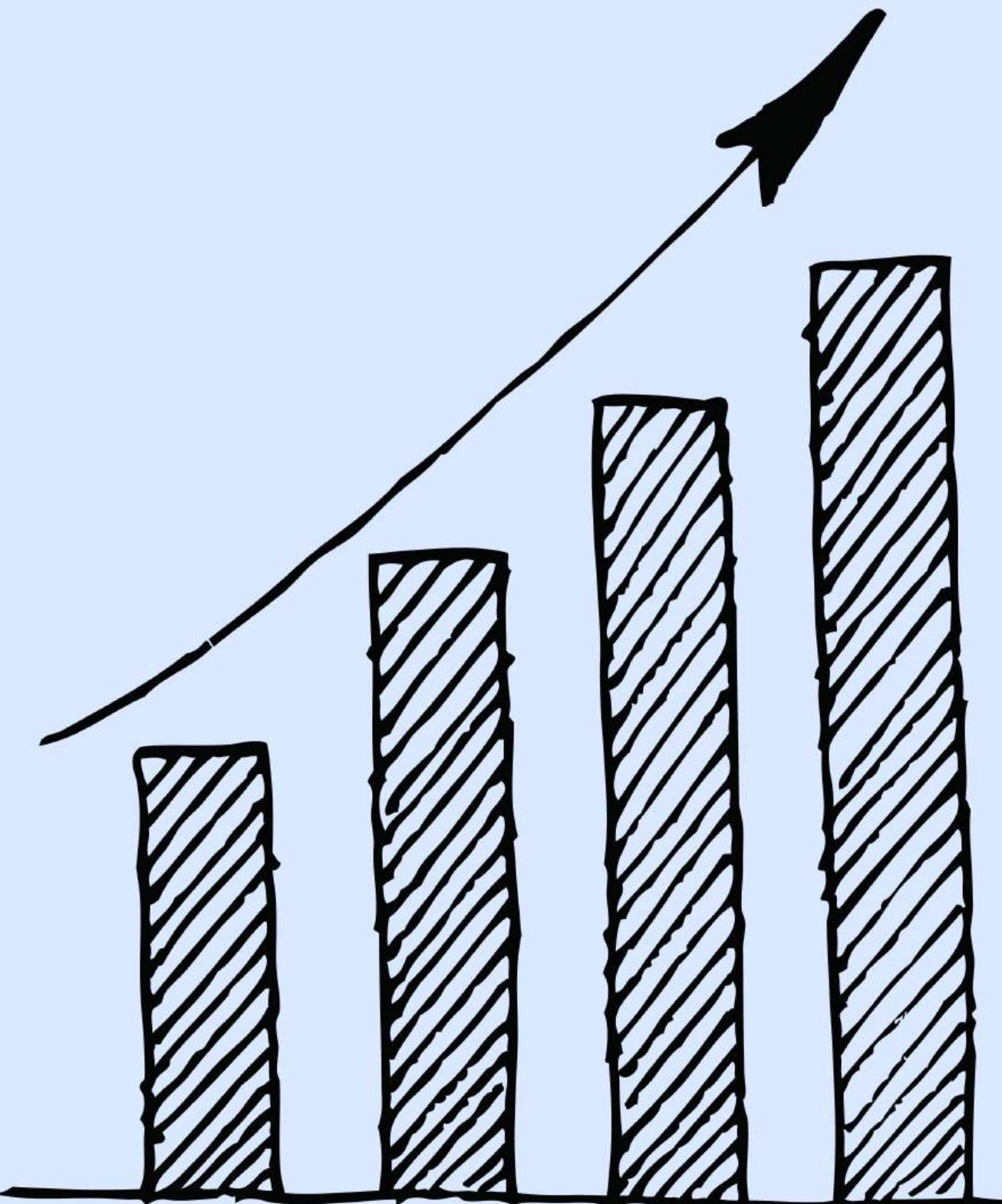
CÁC CHỈ SỐ ĐỂ ĐÁNH GIÁ HIỆU SUẤT MODEL

Accuracy (Độ chính xác)

Precision(Độ tập trung)

Specificity (Độ đặc hiệu)

Sensitivity (Độ nhạy)



CONFUSION MATRIX

		THỰC TẾ	
		POSITIVE	NEGATIVE
DỰ ĐOÁN	POSITIVE	True Positive (TP)	False Positive (FP)
	NEGATIVE	False Negative (FN)	True Negative (TN)

Thực tế: Email là spam.

Dự đoán:

- Email là spam -> TP
- Email KHÔNG là spam -> FN

Thực tế: Email KHÔNG là spam.

Dự đoán:

- Email là spam -> FP
- Email KHÔNG là spam -> TN

$$\text{Accuracy (Độ chính xác)} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}}$$

- Cho thấy tỷ lệ của dự đoán đúng (TP + TN) trên tổng số dữ liệu (TP+TF+FP+FN).
- Đo lường khả năng của mô hình trong việc dự đoán đúng cả Positive và Negative.
- Accuracy cao cho thấy mô hình hoạt động tốt trong việc phân loại các mẫu dữ liệu.
- Không phản ánh đúng về hiệu suất của mô hình, đặc biệt khi dữ liệu bị mất cân bằng (imbalanced data).

		THỰC TẾ	
		POSITIVE	NEGATIVE
DỰ ĐOÁN	POSITIVE	TP	FP
	NEGATIVE	FN	TN

$$\text{Specificity (Độ đặc hiệu)} = \frac{\text{TN}}{\text{TN} + \text{FP}}$$

- Còn được gọi là True Negative Rate (TNR).
- Cho thấy tỷ lệ của dự đoán đúng Negative trên tổng số dữ liệu thực tế Negative (TN + FP).
- Đo lường khả năng phát hiện các trường hợp Negative một cách chính xác.
- Specificity càng cao thì mô hình càng tốt trong việc phát hiện các trường hợp Negative chính xác.
- Không nói lên khả năng của mô hình trong việc tránh sai Negative (FN).

		THỰC TẾ	
		POSITIVE	NEGATIVE
DỰ ĐOÁN	POSITIVE	TP	FP
	NEGATIVE	FN	TN

$$\text{Sensitivity (Độ nhạy)} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

- Còn được gọi là Recall .
- Cho thấy tỷ lệ của dự đoán đúng Positive (TP) trên tổng số dữ liệu thực tế Positive (TP + FN).
- Đo lường khả năng phát hiện chính xác các Positive.
- Sensitivity càng cao thì mô hình càng tốt trong việc phát hiện các Positive.
- **VD: phát hiện bệnh COVID**

		THỰC TẾ	
		POSITIVE	NEGATIVE
DỰ ĐOÁN	POSITIVE	TP	FP
	NEGATIVE	FN	TN

- Không nói lên gì về khả năng của mô hình trong việc tránh sai Positive (FP).
- Thường được dùng với Precision để đánh giá toàn diện hiệu suất của mô hình.

PHƯƠNG UYÊN

$$\text{Precision (Độ tập trung)} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

- Còn được gọi là Positive Predictive Value(PPV)

- Cho thấy tỷ lệ của dự đoán đúng Positive (TP) trên tổng số dự đoán của Positive (TP + FP).

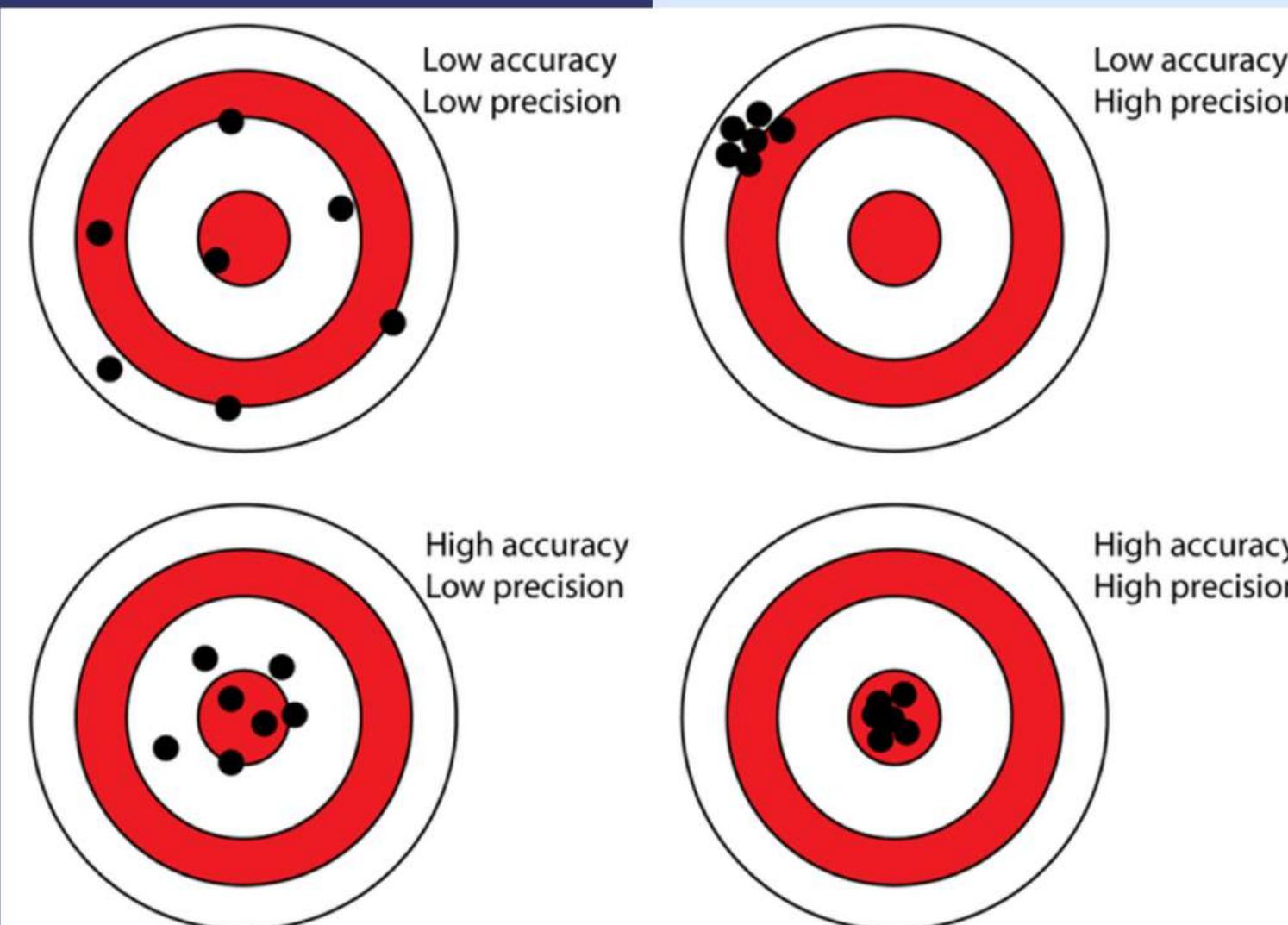
- Tập trung vào khả năng tránh dự đoán sai Positive (FP).

- Precision cao cho thấy mô hình có xu hướng dự đoán chính xác các trường hợp Positive.
- Cần được đánh giá cùng với các phép đo khác như Sensitivity để hiểu rõ hơn về hiệu suất tổng thể của mô hình.

		THỰC TẾ	
		POSITIVE	NEGATIVE
DỰ ĐOÁN	POSITIVE	TP	FP
	NEGATIVE	FN	TN

ACCURACY

Đánh giá khả năng dự đoán chính xác cả Positive và Negative.
Dữ liệu cân bằng cả hai lớp Positive và Negative.



PRECISION

Đánh giá khả năng dự báo chính xác các trường hợp Positive.

SENSITIVITY

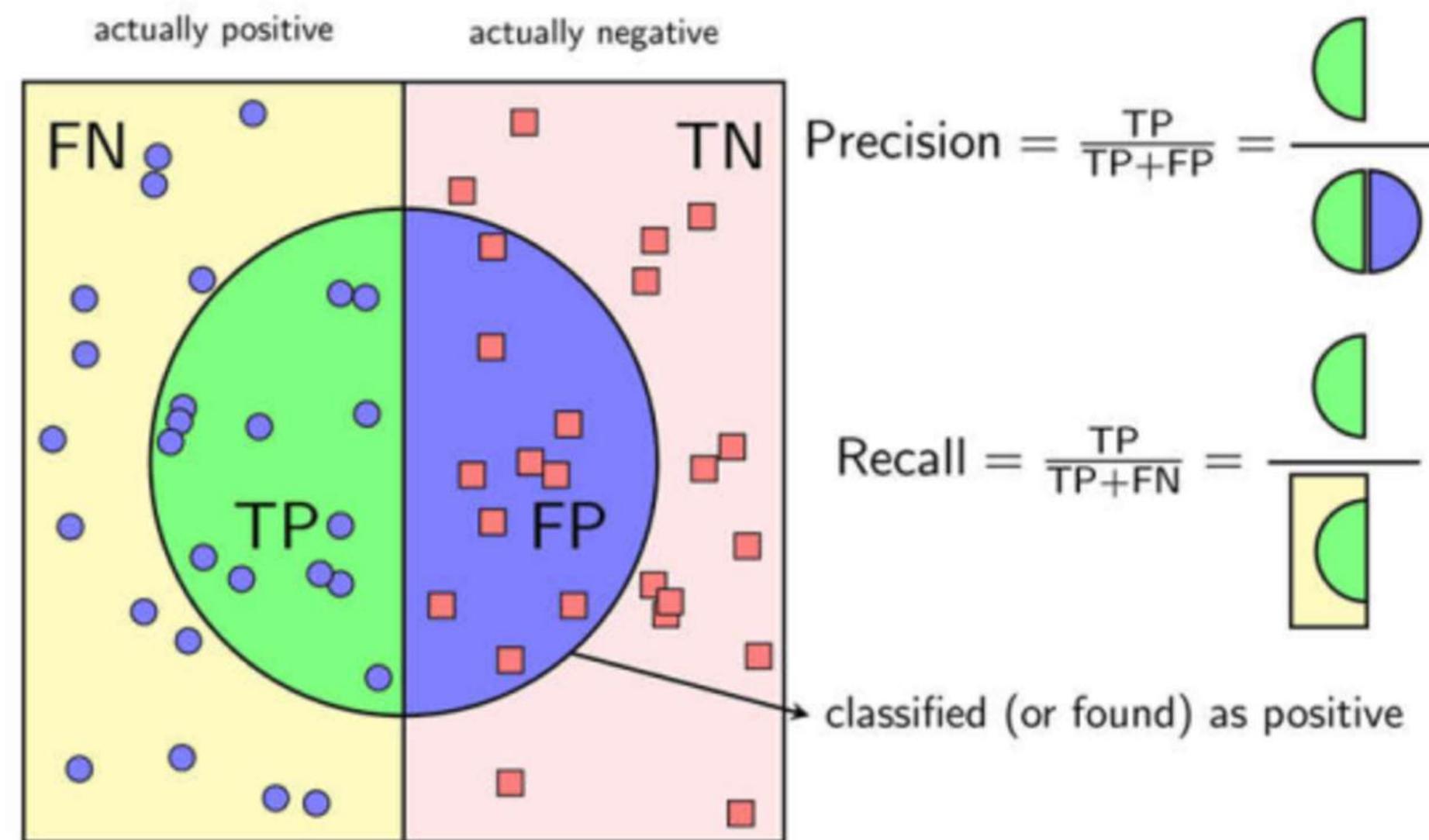
Đánh giá khả năng dự báo đúng các trường hợp Positive gồm cả TP và FN

Để đảm bảo không bỏ sót các trường hợp Positive, thì Sensitivity là độ đo quan trọng hơn để đánh giá hiệu suất của mô hình.

PHƯƠNG UYÊN

PRECISION

Đánh giá khả năng dự báo đúng các trường hợp Positive dựa trên tổng số dự báo Positive (TP+FP)

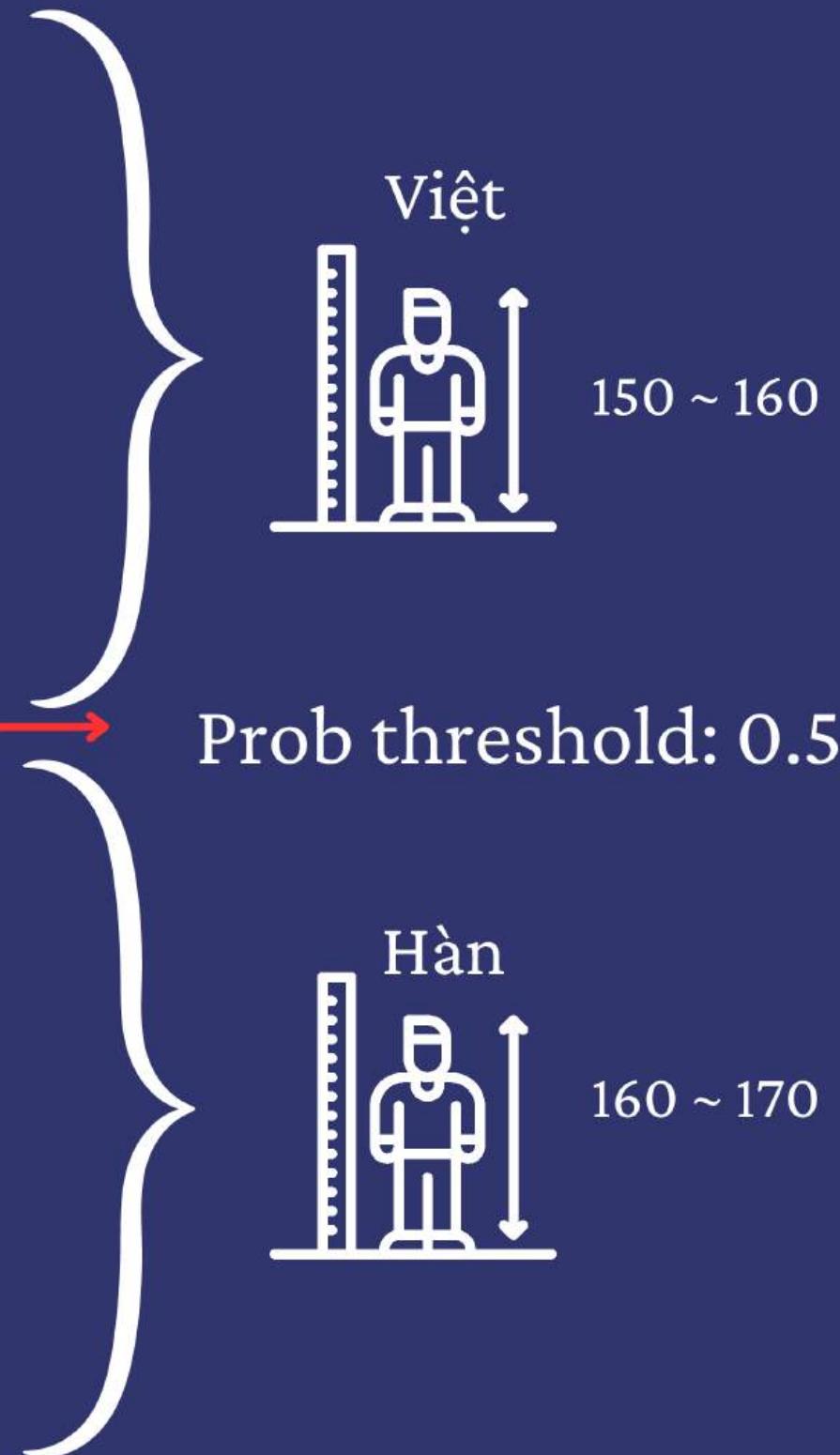


$$\text{F1 - Score} = 2 \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

- F1-Score là một độ đo kết hợp giữa Sensitivity (Recall) và Precision.
 - Giá trị của F1-Score nằm trong khoảng từ 0 -> 1, với 1 là giá trị TỐT nhất và 0 là giá trị THẤP nhất.
 - Thích hợp dùng khi:
 - DL cân bằng và không cân bằng.
 - Cả Precision và Recall đều quan trọng.
 - Muốn sử dụng chỉ 1 độ đo.
 - Lựa chọn và điều chỉnh mô hình.

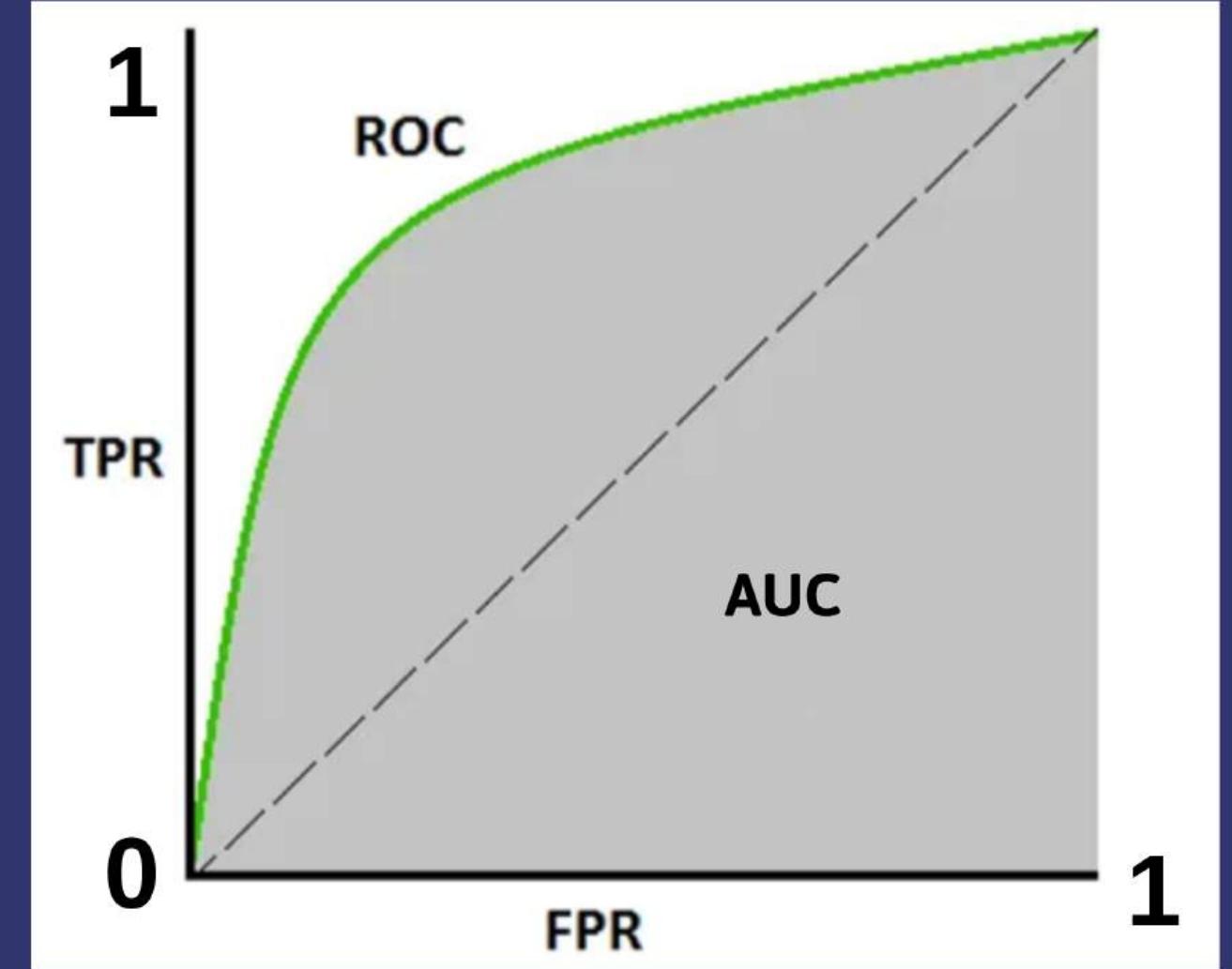
ROC Curve

	Chiều cao	Class (Quốc tịch)
1	151	Việt
2	153.5	Việt
3	156.5	Việt
4	158	Hàn
5	159.5	Việt
6	162	Hàn
7	163.5	Việt
8	165.5	Hàn
9	167	Hàn
10	168	Hàn



ROC Curve

- Mỗi điểm trên ROC tương ứng với một ngưỡng (threshold) quyết định khác nhau của mô hình.
- Đường ROC cho chúng ta cái nhìn tổng quan về hiệu suất phân loại của mô hình ở các threshold quyết định khác nhau.



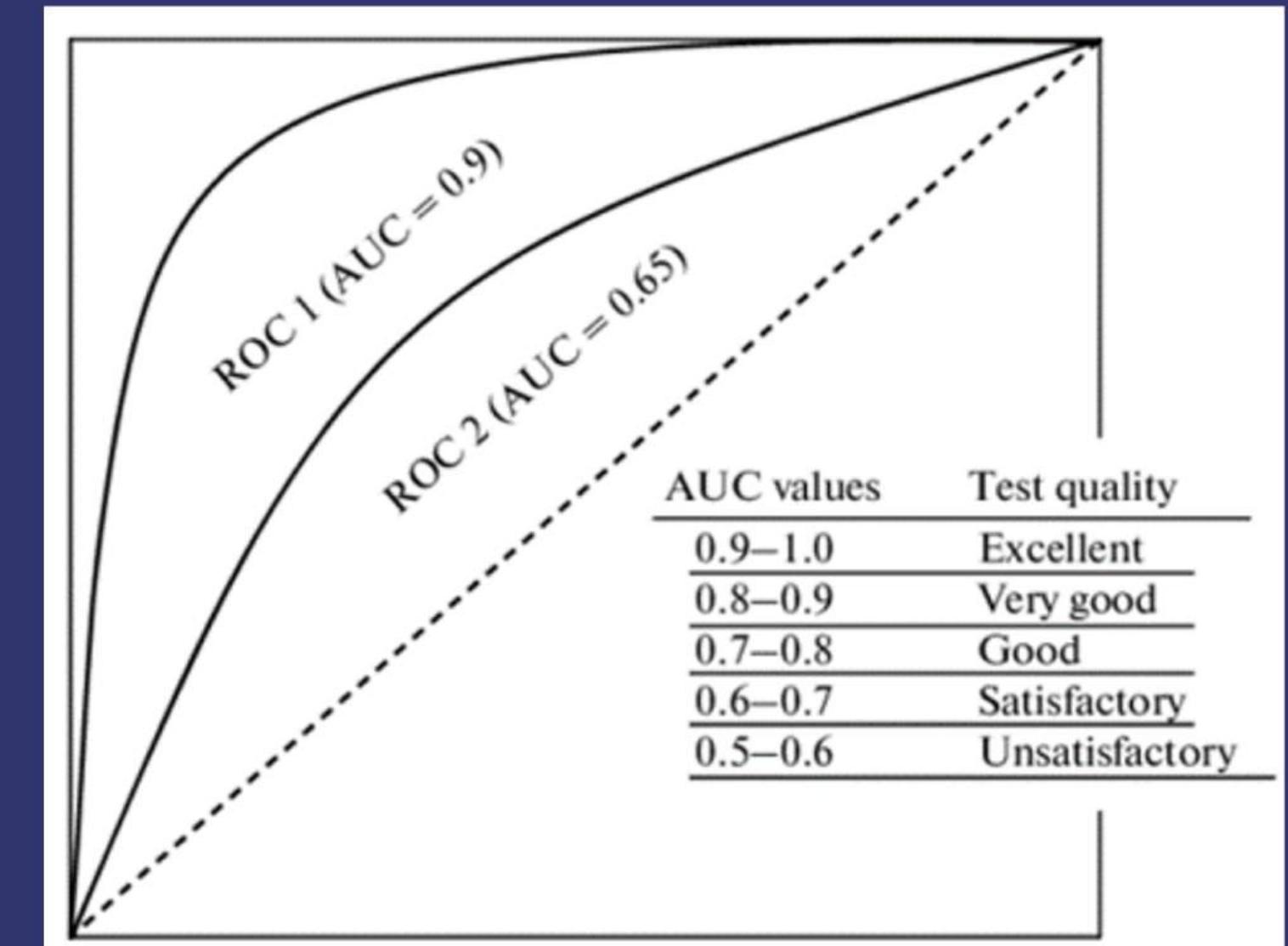
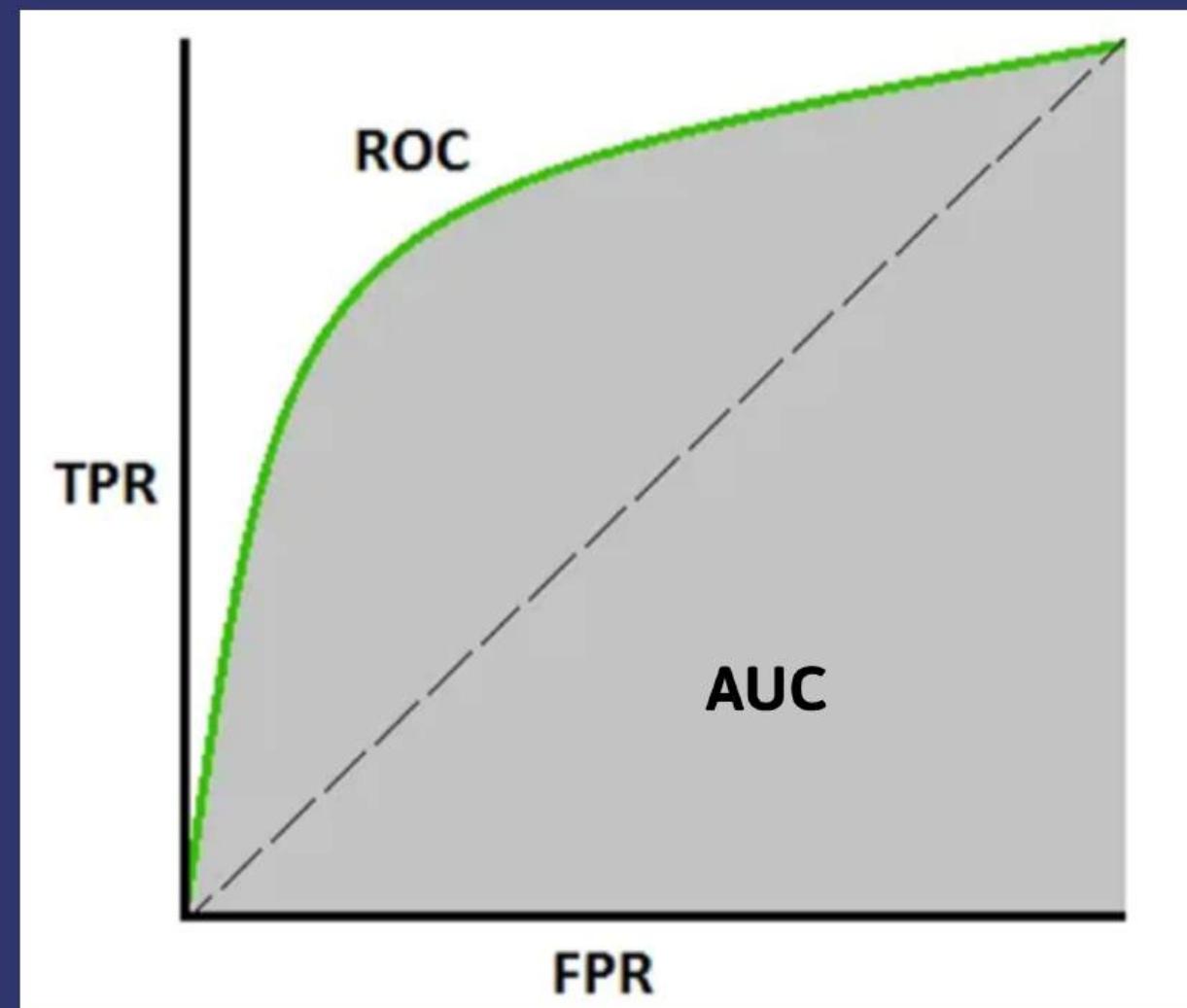
$$TPR = \frac{TP}{TP + FN}$$

$$\text{Specificity} = 1 - FPR$$

- TPR tăng \rightarrow FPR tăng \rightarrow Specificity giảm

AUC

- AUC là phần diện tích bên dưới ROC curve.
- Để đánh giá hiệu năng phân loại của mô hình.
- AUC càng lớn thì mô hình càng ý nghĩa.



Thank You