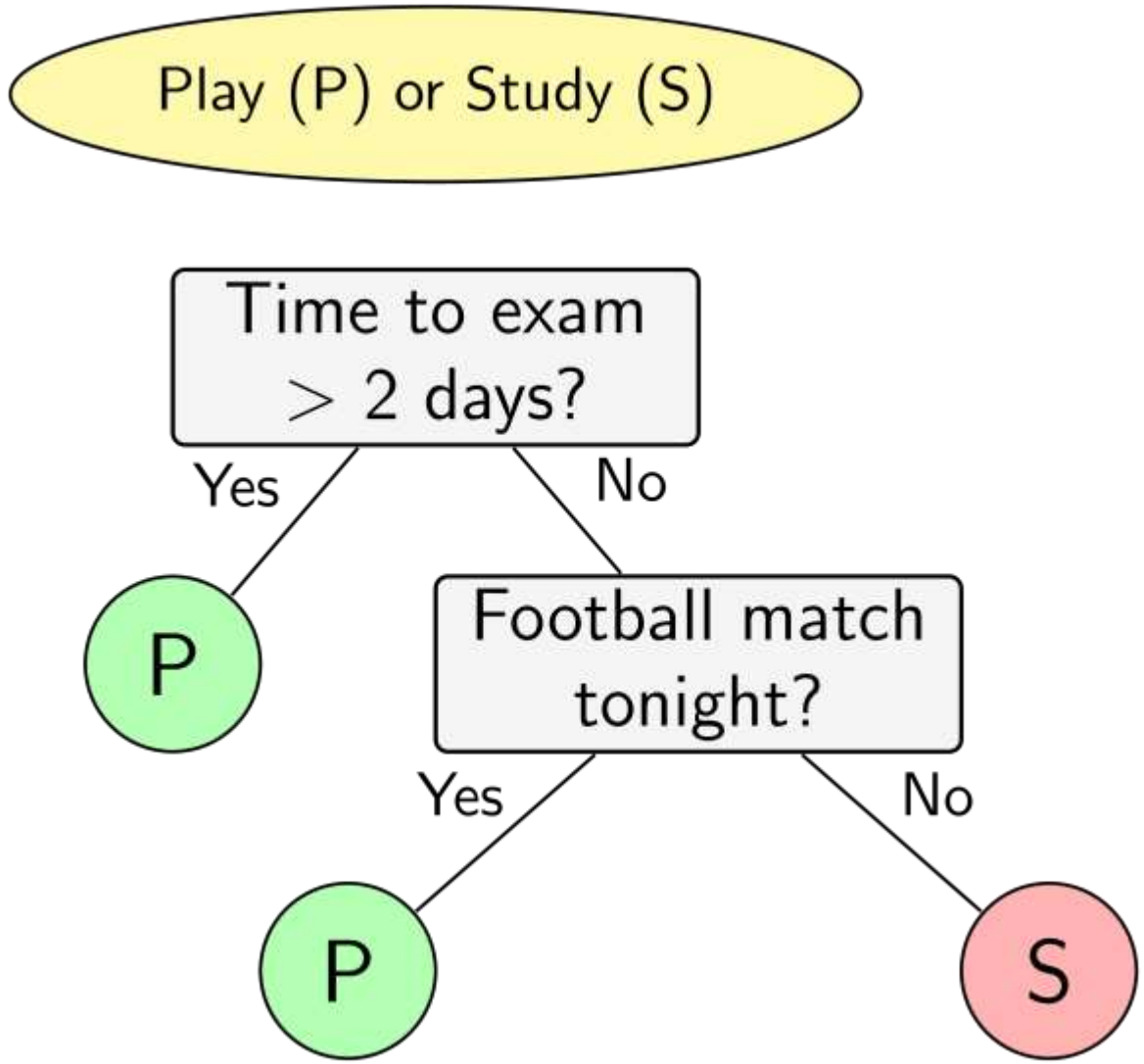
DECISION TREE

PHAM VAN KHANH


Play (P) or Study (S)
Time to exam
> 2 days?
Yes
No
P
Football match
tonight?
Yes
No
P
S

- Decision tree là một mô hình supervised learning, có thể được áp dụng vào cả hai bài toán classification và regression.
- Việc xây dựng một decision tree trên dữ liệu huấn luyện cho trước là việc đi xác định các *câu hỏi* và *thứ tự của chúng*.

- Một điểm đáng lưu ý của decision tree là nó có thể làm việc với các đặc trưng dạng *categorical*, thường là rời rạc và không có thứ tự. Ví dụ, *mưa, nắng* hay *xanh, đỏ*, v.v.
- Decision tree cũng làm việc với dữ liệu có vector đặc trưng bao gồm cả thuộc tính dạng categorical và liên tục (*numeric*). Một điểm đáng lưu ý nữa là decision tree ít yêu cầu việc chuẩn hoá dữ liệu.

ID3

- Trong ID3, chúng ta cần xác định thứ tự của thuộc tính cần được xem xét tại mỗi bước.
- Với các bài toán có nhiều thuộc tính và mỗi thuộc tính có nhiều giá trị khác nhau, việc tìm được nghiệm tối ưu thường là không khả thi.

- Thay vào đó, một phương pháp đơn giản thường được sử dụng là tại mỗi bước, một thuộc tính *tốt nhất* sẽ được chọn ra dựa trên một tiêu chuẩn nào đó (chúng ta sẽ bàn sớm).
- Với mỗi thuộc tính được chọn, ta chia dữ liệu vào các *child node* tương ứng với các giá trị của thuộc tính đó rồi tiếp tục áp dụng phương pháp này cho mỗi *child node*.
- Việc chọn ra thuộc tính *tốt nhất* ở mỗi bước như thế này được gọi là cách chọn *greedy* (*tham lam*).

- Sau mỗi *câu hỏi*, dữ liệu được phân chia vào từng *child node* tương ứng với các câu trả lời cho câu hỏi đó.
- *Câu hỏi* ở đây chính là một thuộc tính, câu trả lời chính là giá trị của thuộc tính đó.
- Để đánh giá *chất lượng* của một cách phân chia, chúng ta cần đi tìm một phép đo.

- Trước hết, thế nào là một phép phân chia tốt? Bằng trực giác, một phép phân chia là tốt nhất nếu dữ liệu trong mỗi *child node* hoàn toàn thuộc vào một class–khi đó *child node* này có thể được coi là một *leaf node*, tức ta không cần phân chia thêm nữa.

- Nếu dữ liệu trong các *child node* vẫn lẫn vào nhau theo tỉ lệ lớn, ta coi rằng phép phân chia đó chưa thực sự tốt.
- Từ nhận xét này, ta cần có một hàm số đo *độ tinh khiết* (*purity*), hoặc *độ vẩn đục* (*impurity*) của một phép phân chia.
- Hàm số này sẽ cho giá trị thấp nhất nếu dữ liệu trong mỗi *child node* nằm trong cùng một class (tinh khiết nhất), và cho giá trị cao nếu mỗi *child node* có chứa dữ liệu thuộc nhiều class khác nhau.

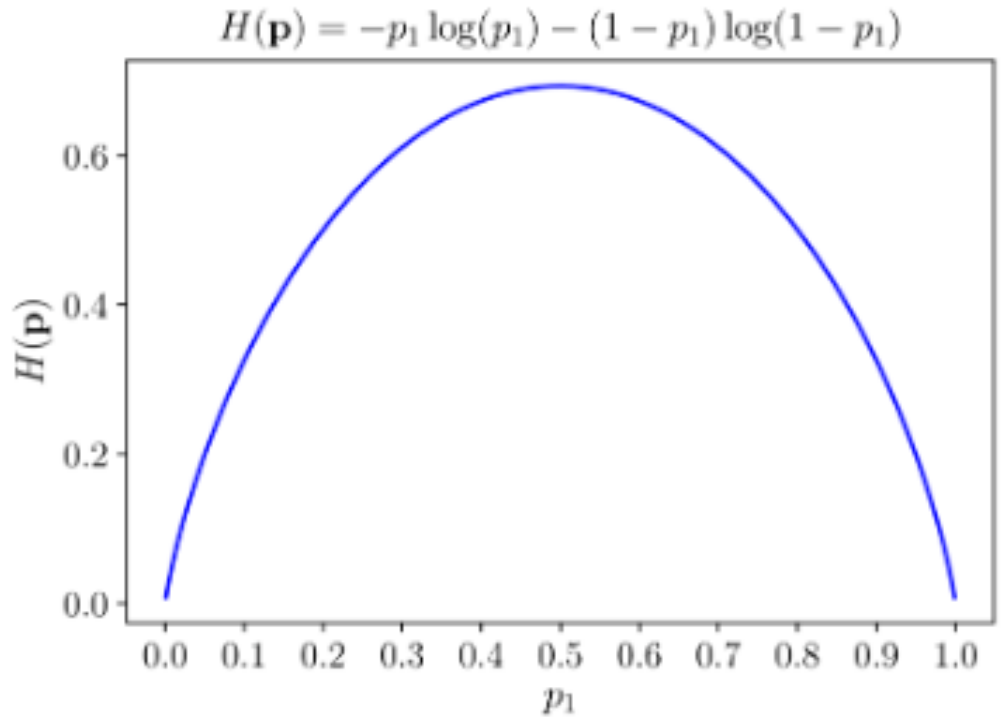
Hàm số entropy

Cho một phân phối xác suất của một biến rời rạc x có thể nhận n giá trị khác nhau $x_1, x_2, \ldots, x_n$. Giả sử rằng xác suất để x nhận các giá trị này là $p_i = p(x = x_i)$ với $0 \leq p_i \leq 1, \sum_{i=1}^{n} p_i = 1$. Ký hiệu phân phối này là $\mathbf{p} = (p_1, p_2, \ldots, p_n)$. Entropy của phân phối này được định nghĩa là

$$H(\mathbf{p}) = -\sum_{i=1}^{n} p_i \log(p_i) \qquad (1)$$

trong đó $\log$ là logarit tự nhiên (*Một số tài liệu dùng logarit cơ số 2, nhưng giá trị của $H(\mathbf{p})$ chỉ khác đi bằng cách nhân với một hằng số.*) và quy ước $0 \log(0) = 0$.

Xét một ví dụ với $n = 2$ được cho trên Hình 3. Trong trường hợp $\mathbf{p}$ là *tinh khiết* nhất, tức một trong hai giá trị p_i bằng 1, giá trị kia bằng 0, entropy của phân phối này là $H(\mathbf{p}) = 0$. Khi $\mathbf{p}$ là *vẩn đục* nhất, tức cả hai giá trị $p_i = 0.5$, hàm entropy đạt giá trị cao nhất.



Hình 3: Đồ thị của hàm entropy với $n = 2$.

- Trong ID3, *tổng có trọng số của entropy tại các leaf-node* sau khi xây dựng decision tree được coi là hàm mất mát của decision tree đó.
- Các trọng số ở đây tỉ lệ với số điểm dữ liệu được phân vào mỗi node

- . Công việc của ID3 là tìm các cách phân chia hợp lý (thứ tự chọn thuộc tính hợp lý) sao cho hàm mất mát cuối cùng đạt giá trị càng nhỏ càng tốt.
- Bài toán xây dựng một decision tree bằng ID3 có thể chia thành các bài toán nhỏ, trong mỗi bài toán, ta chỉ cần chọn ra thuộc tính giúp cho việc phân chia đạt kết quả tốt nhất.
- Mỗi bài toán nhỏ này tương ứng với việc phân chia dữ liệu trong một *non-leaf node*. Chúng ta sẽ xây dựng phương pháp tính toán dựa trên mỗi node này.

Xét một bài toán với C class khác nhau. Giả sử ta đang làm việc với một *non-leaf node* với các điểm dữ liệu tạo thành một tập $\mathcal{S}$ với số phần tử là $|\mathcal{S}| = N$. Giả sử thêm rằng trong số N điểm dữ liệu này, $N_c, c = 1, 2, \ldots, C$ điểm thuộc vào class c. Xác suất để mỗi điểm dữ liệu rơi vào một class c được xấp xỉ bằng $\frac{N_c}{N}$ (maximum likelihood estimation). Như vậy, entropy tại node này được tính bởi:

$$H(\mathcal{S}) = -\sum_{c=1}^{C} \frac{N_c}{N} \log\left(\frac{N_c}{N}\right) \qquad (2)$$

Tiếp theo, giả sử thuộc tính được chọn là x. Dựa trên x, các điểm dữ liệu trong S được phân ra thành K child node $S_1, S_2, \ldots, S_K$ với số điểm trong mỗi child node lần lượt là $m_1, m_2, \ldots, m_K$. Ta định nghĩa

$$H(x, S) = \sum_{k=1}^{K} \frac{m_k}{N} H(S_k) \qquad (3)$$

là tổng có trọng số entroy của mỗi child node–được tính tương tự như (2). Việc lấy trọng số này là quan trọng vì các node thường có số lượng điểm khác nhau.

Tiếp theo, ta định nghĩa *information gain* dựa trên thuộc tính $\boldsymbol{x}$:

$$G(\boldsymbol{x}, \mathcal{S}) = H(\mathcal{S}) - H(\boldsymbol{x}, \mathcal{S})$$

Trong ID3, tại mỗi node, thuộc tính được chọn được xác định dựa trên:

$$\boldsymbol{x}^* = \arg\max_{\boldsymbol{x}} G(\boldsymbol{x}, \mathcal{S}) = \arg\min_{\boldsymbol{x}} H(\boldsymbol{x}, \mathcal{S})$$

tức thuộc tính khiến cho *information gain* đạt giá trị lớn nhất.

Ví dụ

- Bảng dữ liệu này mô tả mối quan hệ giữa thời tiết trong 14 ngày (bốn cột đầu, không tính cột id) và việc một đội bóng có chơi bóng hay không (cột cuối cùng). Nói cách khác, ta phải dự đoán giá trị ở cột cuối cùng nếu biết giá trị của bốn cột còn lại.

id	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

Có bốn thuộc tính thời tiết:

1. *Outlook* nhận một trong ba giá trị: sunny, overcast, rainy.
2. *Temperature* nhận một trong ba giá trị: hot, cool, mild.
3. *Humidity* nhận một trong hai giá trị: high, normal.
4. *Wind* nhận một trong hai giá trị: weak, strong.

(Tổng cộng có $\mathbf{3 \times 3 \times 2 \times 2 = 36}$ loại thời tiết khác nhau, trong đó 14 loại được thể hiện trong bảng.)

- Đây có thể được coi là một bài toán dự đoán liệu đội bóng có chơi bóng không dựa trên các quan sát thời tiết. Ở đây, các quan sát đều ở dạng categorical. Cách dự đoán dưới đây tương đối đơn giản và khá chính xác, có thể không phải là cách ra quyết định tốt nhất:
- Nếu *outlook* = *sunny* và *humidity* = *high* thì *play* = *no*.
- Nếu *outlook* = *rainy* và *windy* = *true* thì *play* = *no*.
- Nếu *outlook* = *overcast* thì *play* = *yes*.
- Ngoài ra, nếu *humidity* = *normal* thì *play* = *yes*.
- Ngoài ra, *play* = *yes*.
- Chúng ta sẽ cùng tìm thứ tự các thuộc tính bằng thuật toán ID3.

Trong 14 giá trị đầu ra ở Bảng trên, có năm giá trị bằng *no* và chín giá trị bằng *yes*. Entroy tại *root node* của bài toán là:

$$H(S) = -\frac{5}{14}\log\left(\frac{5}{14}\right) - \frac{9}{14}\log\left(\frac{9}{14}\right) \approx 0.65$$

Xét thuộc tính *outlook*. Thuộc tính này có thể nhận một trong ba giá trị *sunny, overcast, rainy*. Mỗi một giá trị sẽ tương ứng với một *child node*. Gọi tập hợp các điểm trong mỗi child node này lần lượt là $\mathcal{S}_s, \mathcal{S}_o, \mathcal{S}_r$ với tương ứng m_s, m_o, m_r phần tử. Sắp xếp lại Bảng ban đầu theo thuộc tính outlook ta đạt được ba Bảng nhỏ sau đây.

id	outlook	temperature	humidity	wind	play
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

id	outlook	temperature	humidity	wind	play
3	overcast	hot	high	weak	yes
7	overcast	cool	normal	strong	yes
12	overcast	mild	high	strong	yes
13	overcast	hot	normal	weak	yes

id	outlook	temperature	humidity	wind	play
4	rainy	mild	high	weak	yes
5	rainy	cool	normal	weak	yes
6	rainy	cool	normal	strong	no
10	rainy	mild	normal	weak	yes
14	rainy	mild	high	strong	no

$$H(\mathcal{S}_s) = -\frac{2}{5}\log\left(\frac{2}{5}\right) - \frac{3}{5}\log\left(\frac{3}{5}\right) \approx 0.673$$

$$H(\mathcal{S}_o) = 0$$

$$H(\mathcal{S}_r) = -\frac{3}{5}\log\left(\frac{2}{5}\right) - \frac{3}{5}\log\left(\frac{3}{5}\right) \approx 0.673$$

$$H(outlook, \mathcal{S}) = \frac{5}{14}H(\mathcal{S}_s) + \frac{4}{14}H(\mathcal{S}_o) + \frac{5}{14}H(\mathcal{S}_r) \approx 0.48$$

id	outlook	temperature	humidity	wind	play
1	sunny	hot	high	weak	no
2	sunny	hot	high	strong	no
3	overcast	hot	high	weak	yes
13	overcast	hot	normal	weak	yes

id	outlook	temperature	humidity	wind	play
4	rainy	mild	high	weak	yes
8	sunny	mild	high	weak	no
10	rainy	mild	normal	weak	yes
11	sunny	mild	normal	strong	yes
12	overcast	mild	high	strong	yes
14	rainy	mild	high	strong	no

id	outlook	temperature	humidity	wind	play
5	rainy	cool	normal	weak	yes
6	rainy	cool	normal	strong	no
7	overcast	cool	normal	strong	yes
9	sunny	cool	normal	weak	yes

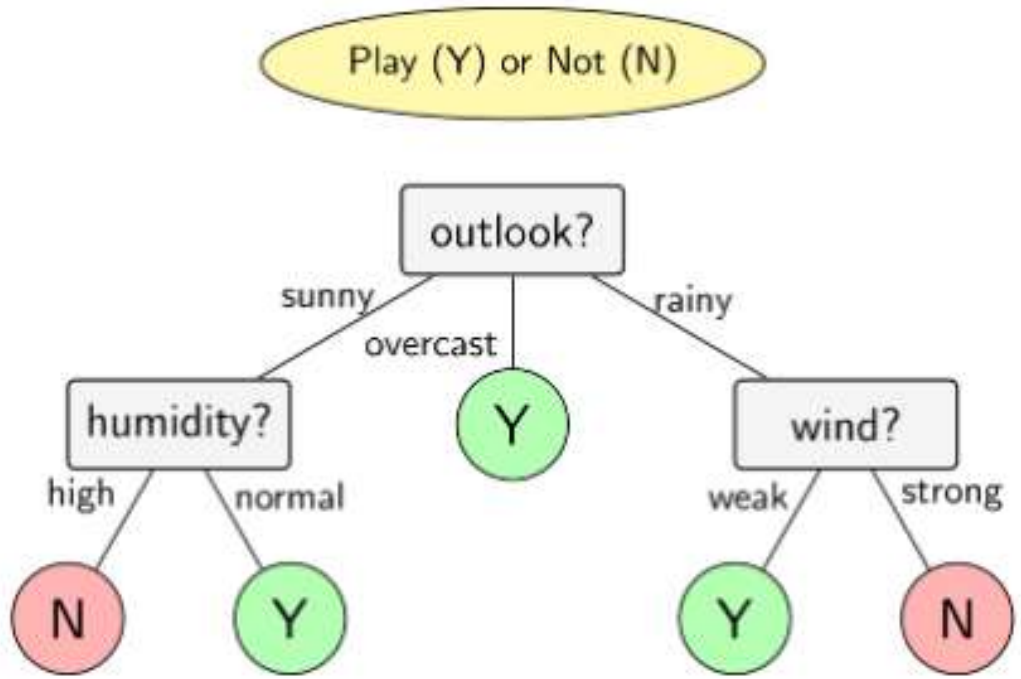
$$H(\mathcal{S}_h) = -\frac{2}{4}\log\left(\frac{2}{4}\right) - \frac{2}{4}\log\left(\frac{2}{4}\right) \approx 0.693$$

$$H(\mathcal{S}_m) = -\frac{4}{6}\log\left(\frac{4}{6}\right) - \frac{2}{6}\log\left(\frac{2}{6}\right) \approx 0.637$$

$$H(\mathcal{S}_c) = -\frac{3}{4}\log\left(\frac{3}{4}\right) - \frac{1}{4}\log\left(\frac{1}{4}\right) \approx 0.562$$

$$H(temperature, \mathcal{S}) = \frac{4}{14}H(\mathcal{S}_h) + \frac{6}{14}H(\mathcal{S}_m) + \frac{4}{14}H(\mathcal{S}_c) \approx 0.631$$

$$H(humidity, \mathcal{S}) \approx 0.547, \quad H(wind, \mathcal{S}) \approx 0.618$$


Play (Y) or Not (N)
outlook?
sunny
overcast
rainy
humidity?
Y
wind?
high
normal
weak
strong
N
Y
Y
N

- https://colab.research.google.com/drive/1RnMrHx1UeElBLVWdjoq-O5QpRJnypTOy?usp=sharing

Bộ dữ liệu **Cleveland Heart Disease**

1. age

- **Mô tả**: Tuổi bệnh nhân (tính theo năm).
- **Ý nghĩa**: Tuổi càng cao, nguy cơ mắc bệnh tim mạch càng lớn do xơ vữa động mạch và suy giảm chức năng tim theo tuổi.

2. sex

- **Mô tả**: Giới tính bệnh nhân.
  - `1 = Nam`
  - `0 = Nữ`
- **Ý nghĩa**: Nam giới có nguy cơ mắc bệnh tim mạch cao hơn ở độ tuổi trung niên, nhưng phụ nữ sau mãn kinh thì nguy cơ cũng tăng đáng kể.

3. cp (Chest pain type – Loại đau ngực)

- **Mã hóa**:
  - `1 = Đau thắt ngực điển hình` (typical angina)
  - `2 = Đau thắt ngực không điển hình` (atypical angina)
  - `3 = Đau không liên quan tim` (non-anginal pain)
  - `4 = Không đau ngực` (asymptomatic)
- **Ý nghĩa**: Một trong những chỉ báo lâm sàng mạnh nhất. Bệnh nhân có **typical angina** thường có nguy cơ cao hơn.

4. trestbps (Resting blood pressure – Huyết áp nghỉ)

- **Đơn vị**: mm Hg.
- **Ý nghĩa**: Huyết áp cao kéo dài là yếu tố nguy cơ chính gây bệnh tim mạch, dẫn đến suy tim và bệnh mạch vành.

5. chol (Serum cholesterol – Nồng độ cholesterol trong máu)

- **Đơn vị**: mg/dl.
- **Ý nghĩa**: Cholesterol cao gây xơ vữa động mạch, tăng nguy cơ tắc nghẽn mạch vành.

6. fbs (Fasting blood sugar – Đường huyết khi đói)

- **Mã hóa**:
  - `1 = fbs > 120 mg/dl`
  - `0 = fbs ≤ 120 mg/dl`
- **Ý nghĩa**: Đái tháo đường (đường huyết cao) là yếu tố nguy cơ hàng đầu của bệnh tim mạch.

7. restecg (Resting electrocardiographic results – Điện tâm đồ khi nghỉ)

- **Mã hóa**:
  - `0 = Bình thường`
  - `1 = Có sóng ST-T bất thường` (ví dụ: đảo ngược sóng T, ST chênh)
  - `2 = Phì đại thất trái theo tiêu chuẩn Estes`
- **Ý nghĩa**: Điện tâm đồ bất thường cho thấy dấu hiệu suy tim, nhồi máu cũ hoặc phì đại tim.

8. thalach (Maximum heart rate achieved – Nhịp tim tối đa đạt được)

- **Đơn vị**: beats per minute (bpm).
- **Ý nghĩa**: Nhịp tim tối đa thấp khi gắng sức thường liên quan đến suy giảm chức năng tim.

9. exang (Exercise induced angina – Đau thắt ngực khi gắng sức)

- **Mã hóa**:
  - `1 = Có`
  - `0 = Không`
- **Ý nghĩa**: Xuất hiện đau ngực khi gắng sức thường là dấu hiệu bệnh mạch vành.

10. oldpeak (ST depression induced by exercise relative to rest – Độ chênh ST so với lúc nghỉ)

- **Đơn vị**: mm.
- **Ý nghĩa**: Giá trị cao cho thấy giảm tưới máu cơ tim khi gắng sức → dấu hiệu thiếu máu cơ tim.

11. slope (Slope of the peak exercise ST segment – Độ dốc ST khi gắng sức)

- **Mã hóa**:
  - `1 = Dốc lên` (upsloping)
  - `2 = Bằng phẳng` (flat)
  - `3 = Dốc xuống` (downsloping)
- **Ý nghĩa**: "Flat" hoặc "downsloping" thường liên quan mạnh đến bệnh tim.

12. ca (Number of major vessels – Số động mạch lớn bị hẹp)

- **Giá trị**: từ 0 đến 3 (đếm được qua chụp X-quang/fluoroscopy).
- **Ý nghĩa**: Số mạch vành hẹp nhiều → nguy cơ bệnh tim càng cao.

13. thal (Thalassemia test result – Kiểm tra thallium stress test)

- **Mã hóa**:
  - `3 = Bình thường`
  - `6 = Lỗi cố định (fixed defect)`
  - `7 = Lỗi có thể hồi phục (reversible defect)`
- **Ý nghĩa**: "Reversible defect" cho thấy thiếu máu cơ tim khi gắng sức → chỉ báo nguy cơ cao.

14. target

- **Mục tiêu dự đoán**:
  - `0 = Không mắc bệnh tim`
  - `1–4 = Mức độ bệnh tim` (1 = nhẹ, 4 = nặng)
- Thường được nhị phân hóa: `0 = Không bệnh`, `1 = Có bệnh`.

- Nhóm **lâm sàng**: tuổi, giới, đau ngực, huyết áp, cholesterol, đường huyết.
- Nhóm **chẩn đoán y khoa**: điện tâm đồ, nhịp tim, ST depression, slope, thallium test.
- Nhóm **chỉ báo hình ảnh**: số mạch vành bị hẹp (ca).