ĐẠI NAM
UNIVERSITY

Bài 5

THUẬT TOÁN CÂY QUYẾT ĐỊNH

Giảng viên: TS. Nguyễn Ngọc Giang

SĐT: 0862411011

Email: giangnn.cntt@dainam.edu.vn

OUTLINE

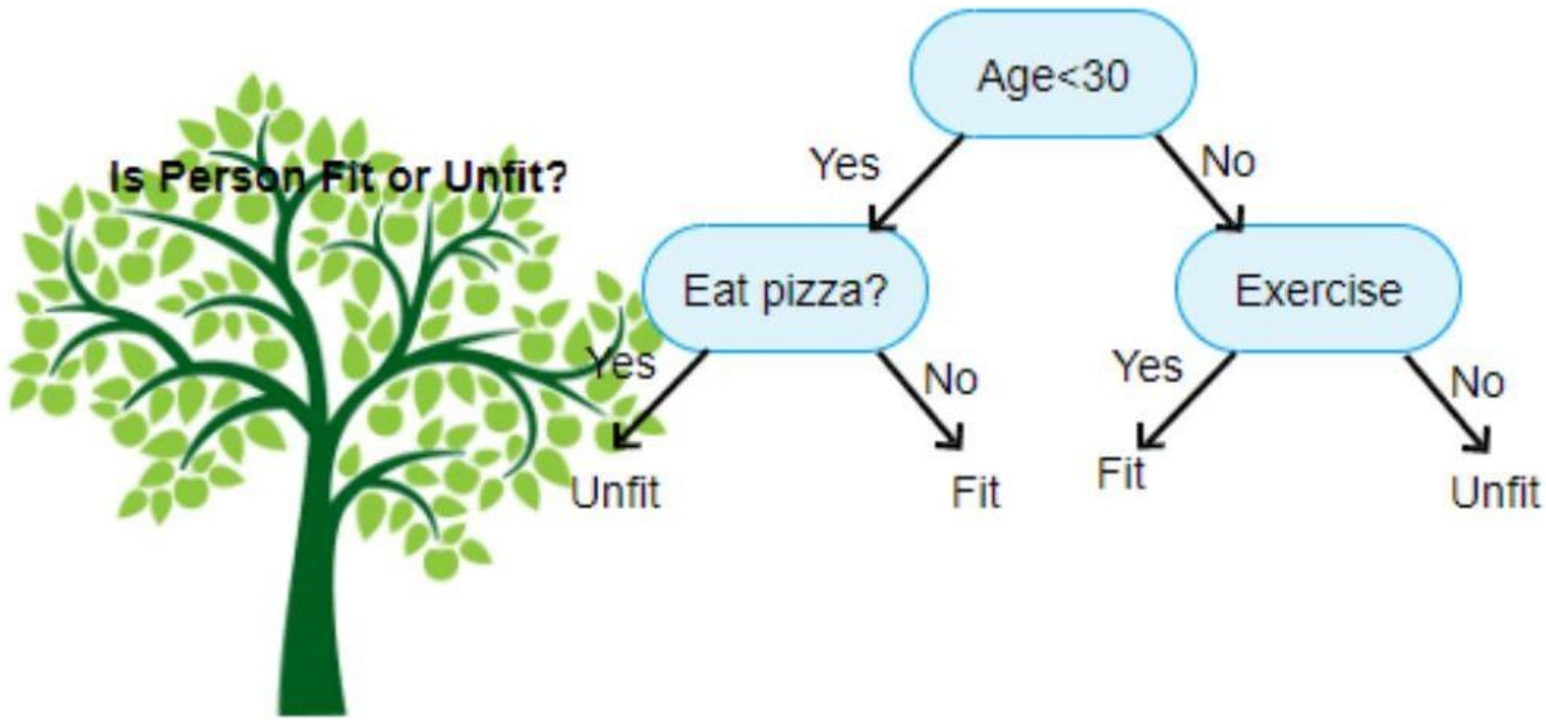
- Giới thiệu
- Ứng dụng của cây quyết định
- Cấu trúc cây quyết định
- Tạo cây quyết định
- Cách tính ID3
- Cách tính Gini Index
- Luyện tập

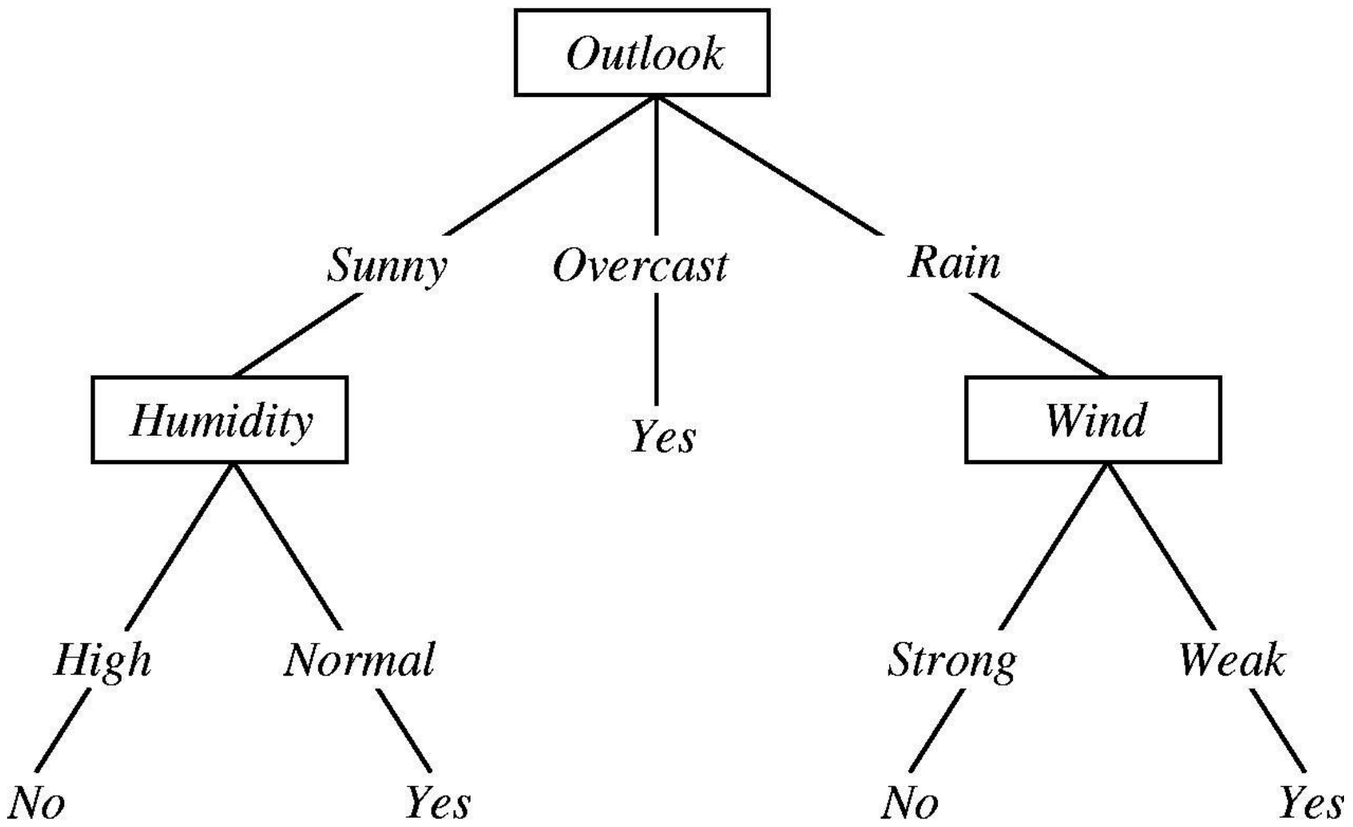
- **Bài toán phân loại (Classification Problem):** Là một loại bài toán trong học máy, thực hiện phân loại các đối tượng vào các nhóm hoặc lớp khác nhau dựa trên các đặc trưng hoặc thuộc tính của chúng.

- **Bài toán hồi quy (Regression Problem):** Là một loại bài toán học máy, thực hiện dự đoán giá trị của một biến liên tục dựa trên một hoặc nhiều biến độc lập (hoặc đặc trưng). Gồm 2 loại biến: Độc lập và Phụ thuộc.

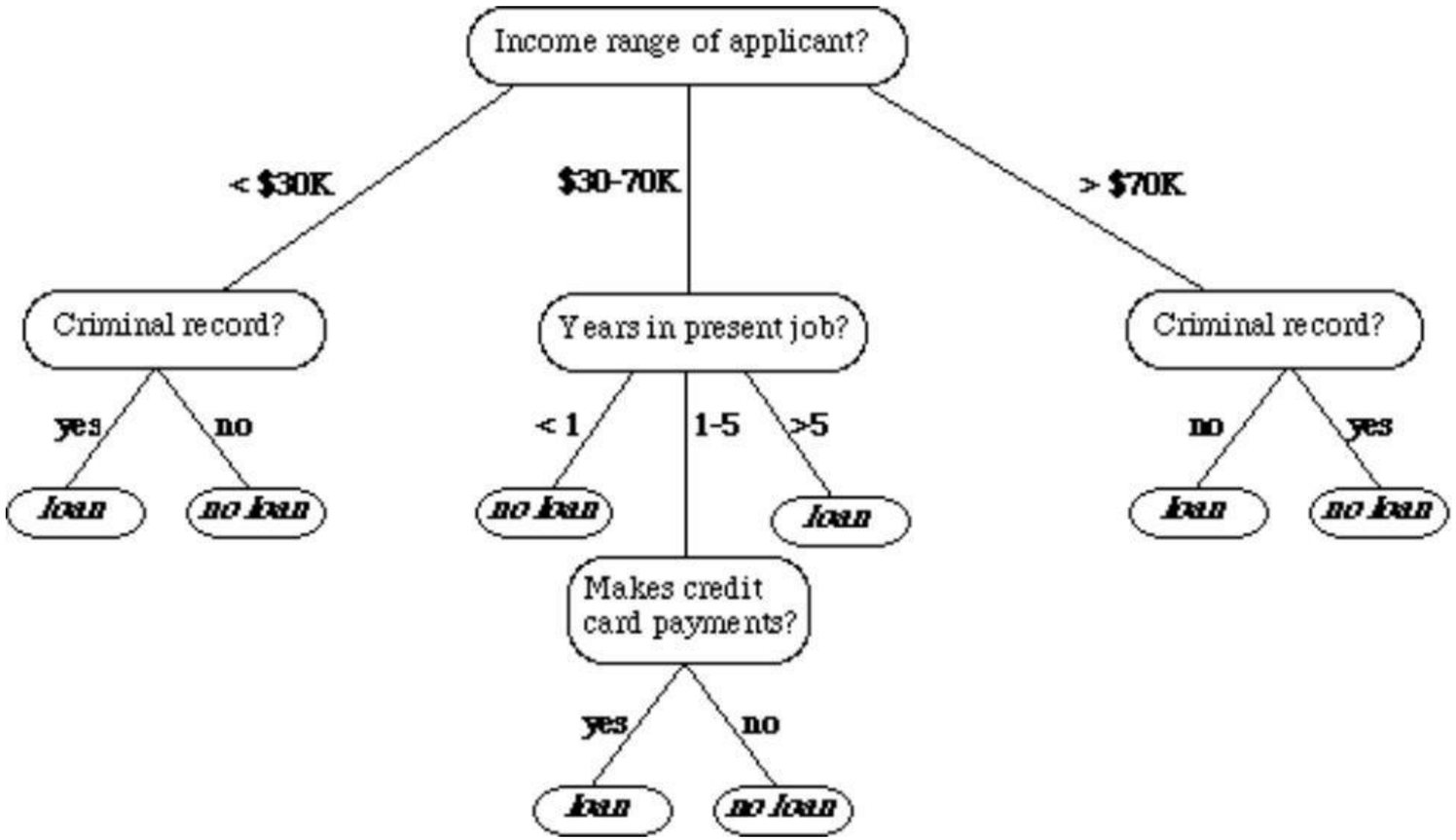
  - Hồi quy tuyến tính (Linear Regression)
  - Hồi quy logistic (Logistic Regression)
  - Hồi quy bội (Multiple Regression)
  - Hồi quy không tuyến tính (Non-linear Regression)

- **Thuật toán cây quyết định (Decision Tree Algorithm):** **Là một phương pháp phổ biến trong học máy dùng để phân loại và hồi quy.**





outlook	temp.	humidity	windy	play
sunny	hot	high	false	no
sunny	hot	high	true	no
overcast	hot	high	false	yes
rainy	mild	high	false	yes
rainy	cool	normal	false	yes
rainy	cool	normal	true	no
overcast	cool	normal	true	yes
sunny	mild	high	false	no
sunny	cool	normal	false	yes
rainy	mild	normal	false	yes
sunny	mild	normal	true	yes
overcast	mild	high	true	yes
overcast	hot	normal	false	yes
rainy	mild	high	true	no



Income range of applicant	Criminal record	Years in present job	Makes credit card payments	Class	Prediction
$25K	no	6	no	loan	
$40K	yes	2	yes	loan	
$80K	no	7	yes	No loan	
$55K	no	8	no	No loan	

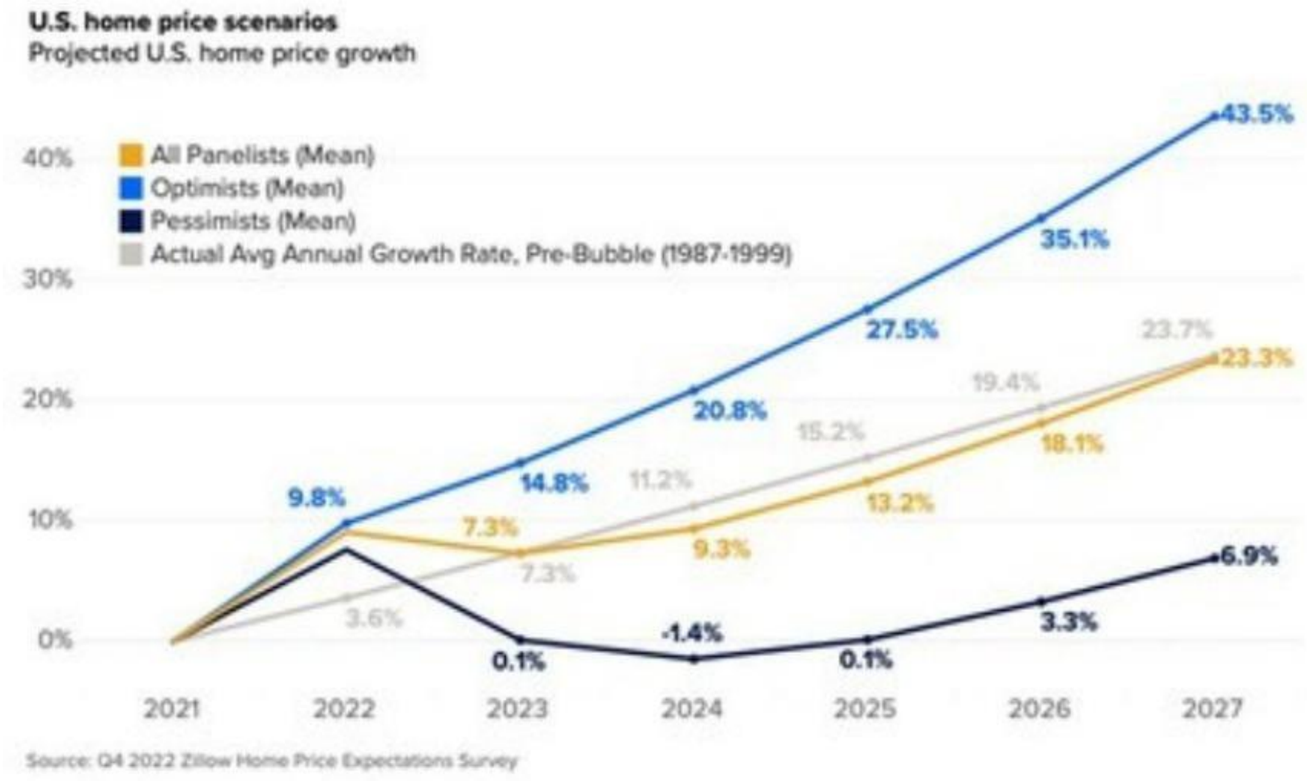
ỨNG DỤNG CỦA CÂY TRUYẾT ĐỊNH

- **Dự báo: Thị trường, nhu cầu, sự bất thường**
- **Chẩn đoán, hiệu suất, sự phù hợp...**

2024 Home Price Forecasts

2024 Forecasts from 11/2023 vs. Current Forecasts

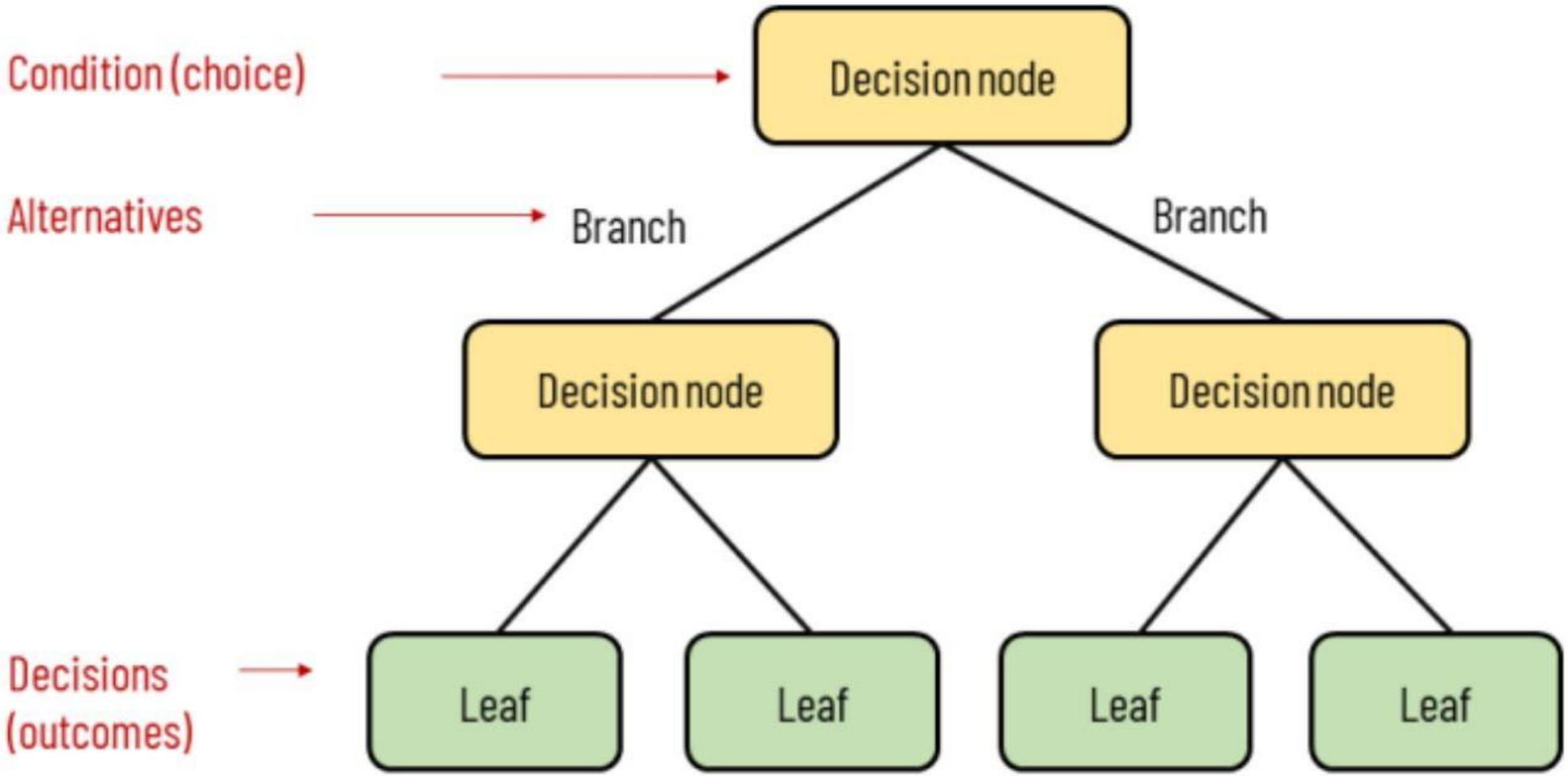
Entity	Original Forecast	Current Forecast
Goldman Sachs	1.9%	**5.0%**
Mortgage Bankers Association	1.1%	**4.1%**
Zillow	0.2%	**3.5%**
Fannie Mae	2.8%	**3.2%**
Freddie Mac	2.6%	**2.8%**
Home Price Expectations Survey	2.2%	**2.4%**
National Association of Realtors	0.7%	**1.9%**



CẤU TRÚC CÂY QUYẾT ĐỊNH

- **Nút gốc (Root Node):** Nút đầu tiên của cây, chứa toàn bộ dữ liệu và đưa ra câu hỏi đầu tiên để phân chia dfi liệu.
- **Nút quyết định (Decision Node):** Các nút bên trong cây nơi dữ liệu được phân chia dựa trên giá trị của một đặc trưng. Mỗi nút quyết định đại diện cho một đặc trưng của dfi liệu.
- **Nút lá (Leaf Node):** Các nút ở cuối cây, đại diện cho các lớp hoặc giá trị dự đoán (cho bài toán phân loại hoặc hồi quy).
- **Nhánh (Branch):** Kết nối gifia các nút, đại diện cho kết quả của câu hỏi hoặc quyết định ở nút quyết định.

CẤU TRÚC CÂY QUYẾT ĐỊNH



- **Chọn đặc trưng phân chia:** Chọn đặc trưng nào để phân chia dữ liệu tại mỗi nút. Quyết định này thường dựa trên các tiêu chí như Entropy, Gini Index, hoặc Gain Ratio.
- **Chia dữ liệu:** Dựa trên đặc trưng đã chọn, chia dữ liệu thành các nhóm con tương ứng với các giá trị của đặc trưng đó.
- **Lặp lại:** Lặp lại quá trình cho môi nhóm con, tiếp tục chia dữ liệu cho đến khi tất cả dữ liệu được phân loại vào các nút lá hoặc đạt đến điều kiện dừng.
- **Tạo cây:** Kết hợp tất cả các quyết định và phân chia để tạo thành cây quyết định.

Các tham số để lựa chọn đặc trưng phân tách dữ liệu:

- **ID3 (Iterative Dichotomiser 3): Thường dùng cho bài toán phân lớp nhị phân**

- **Entropy**: Đo lường độ hỗn loạn hoặc không đồng nhất trong dữ liệu

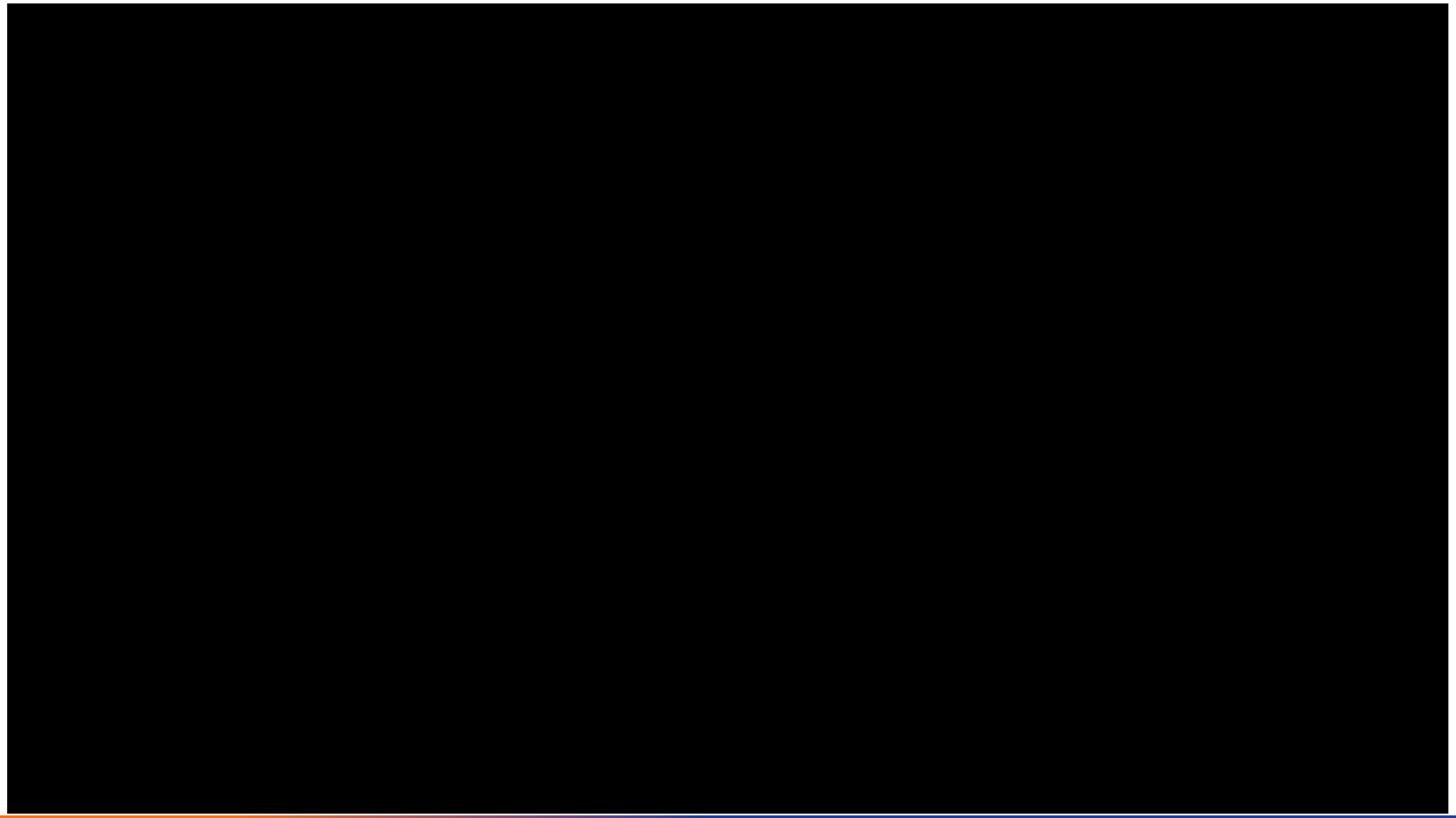
- **Information Gain**: Đo sự giảm của Entropy khi phân tác dữ liệu dựa trên một đặc trưng. Đặc trưng với Information Gain cao nhất thường được chọn để phân chia dữ liệu tại mỗi nút của cây

- **Gini Index: Thường dùng cho bài toán hồi quy**

- Là chỉ số đo lường mức độ hỗn loạn hoặc độ không đồng nhất của các lớp trong dữ liệu

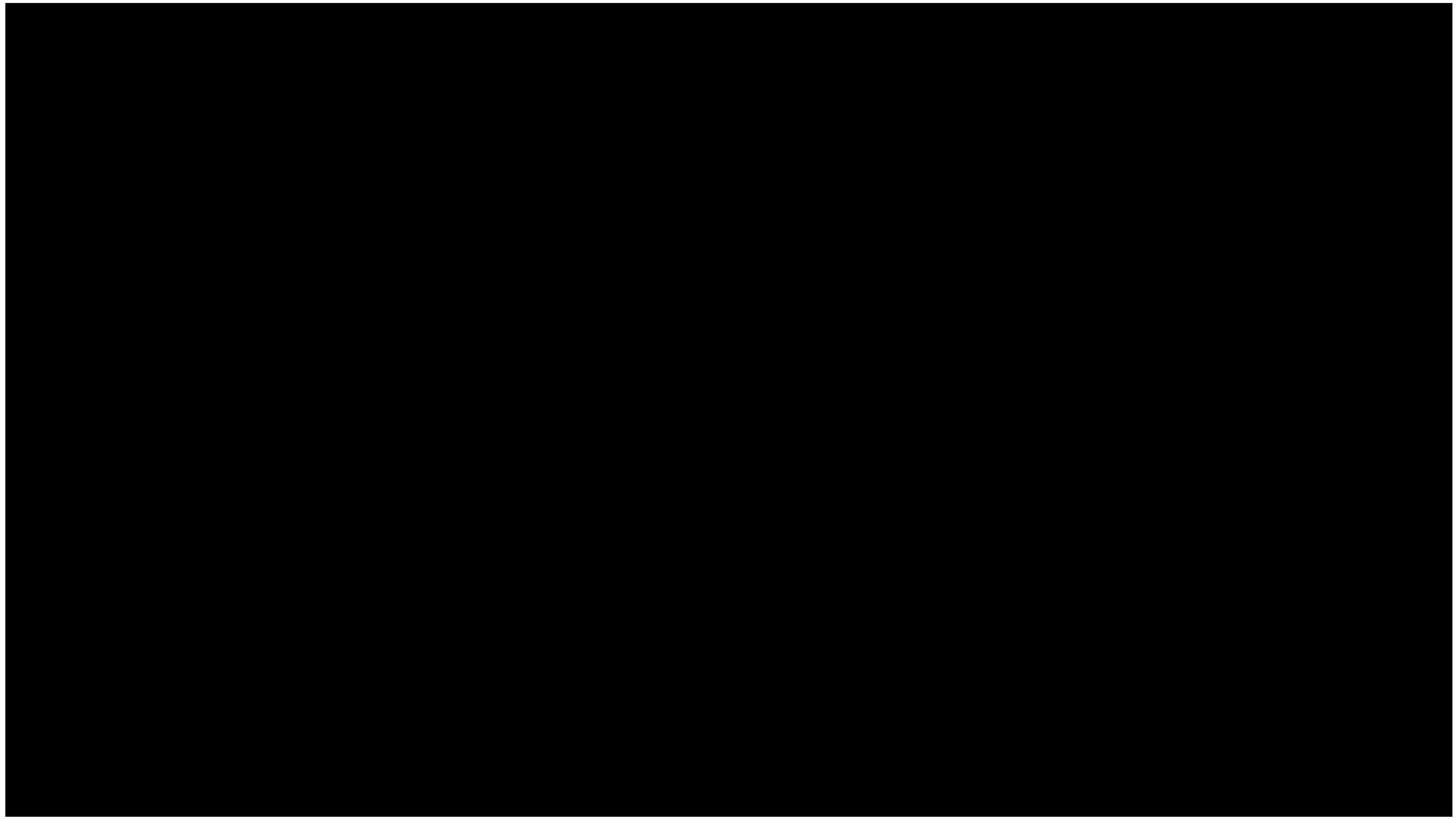
StatQuest: Decission Tree

https://www.youtube.com/watch?v=_L39rN6gz7Y



StatQuest: Entropy

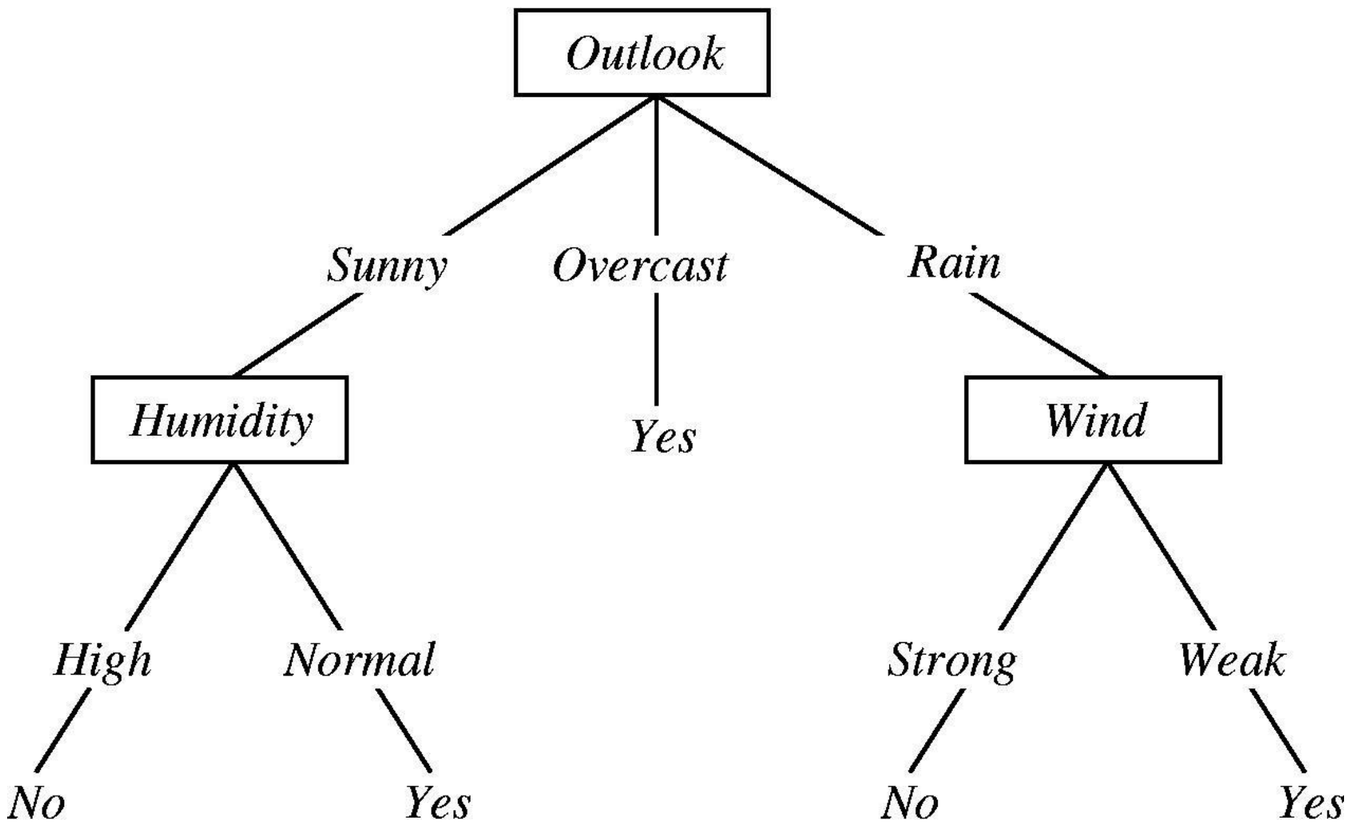
https://www.youtube.com/watch?v=YtebGVx-Fxw



StatQuest: Regression Tree

https://www.youtube.com/watch?v=g9c66TUylZ4

TẠO CÂY QUYẾT ĐỊNH



outlook	temp.	humidity	windy	play
sunny	hot	high	false	no
sunny	hot	high	true	no
overcast	hot	high	false	yes
rainy	mild	high	false	yes
rainy	cool	normal	false	yes
rainy	cool	normal	true	no
overcast	cool	normal	true	yes
sunny	mild	high	false	no
sunny	cool	normal	false	yes
rainy	mild	normal	false	yes
sunny	mild	normal	true	yes
overcast	mild	high	true	yes
overcast	hot	normal	false	yes
rainy	mild	high	true	no

- **Tính Entropy và Informantion Gain:**

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

$$I(x, \mathcal{S}) = \sum_{k=1}^{K} \frac{m_k}{N} E(\mathcal{S}_k)$$

$$G(x, \mathcal{S}) = E(\mathcal{S}) - I(x, \mathcal{S})$$

outlook	temp.	humidity	windy	play
sunny	hot	high	false	no
sunny	hot	high	true	no
overcast	hot	high	false	yes
rainy	mild	high	false	yes
rainy	cool	normal	false	yes
rainy	cool	normal	true	no
overcast	cool	normal	true	yes
sunny	mild	high	false	no
sunny	cool	normal	false	yes
rainy	mild	normal	false	yes
sunny	mild	normal	true	yes
overcast	mild	high	true	yes
overcast	hot	normal	false	yes
rainy	mild	high	true	no

https://machinelearningcoban.com/2018/01/14/id3/

- **Tính Gini Index:**

$$Gini(D) = 1 - \sum_{i=1}^{k} p_i^2$$

$$\text{Gini}_{\text{after}}(D, \text{feature}) = \sum_{j=1}^{m} \frac{|D_j|}{|D|} \text{Gini}(D_j)$$

$$\text{Gini Gain} = \text{Gini}(D) - \text{Gini}_{\text{after}}(D, \text{feature})$$

outlook	temp.	humidity	windy	play
sunny	hot	high	false	no
sunny	hot	high	true	no
overcast	hot	high	false	yes
rainy	mild	high	false	yes
rainy	cool	normal	false	yes
rainy	cool	normal	true	no
overcast	cool	normal	true	yes
sunny	mild	high	false	no
sunny	cool	normal	false	yes
rainy	mild	normal	false	yes
sunny	mild	normal	true	yes
overcast	mild	high	true	yes
overcast	hot	normal	false	yes
rainy	mild	high	true	no

https://www.learndatasci.com/glossary/gini-impurity/

Tóm tắt


PRACTICE
ENGLISH
EVERYDAY
30 Minutes

BÀI TẬP VỀ NHÀ

- **Tìm dữ liệu và lập trình dự báo giá nhà với thuật toán Decision Tree.**


Thank
you!