AI VIET NAM – COURSE 2024

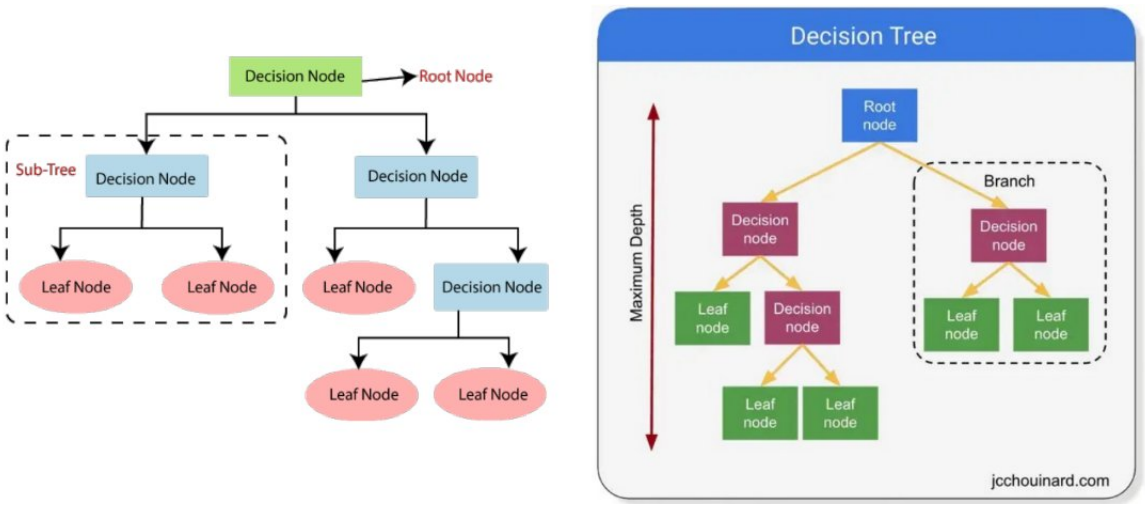
Decision Tree for Classification

Ngày 17 tháng 2 năm 2024

Nguồn dữ liệu:	Decision Tree for Classification
Từ khóa:	Decision Tree, Gini Impurity, Entropy
Người tóm tắt:	Nguyễn Tấn Phát

1 Decision Tree.

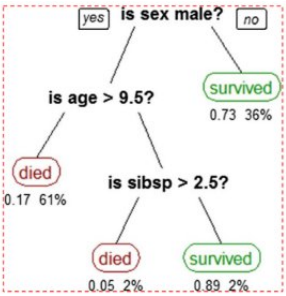
- *Cây quyết định* (Decision Tree): là một cây phân cấp có cấu trúc được dùng để phân lớp các đối tượng dựa vào dãy các luật. Các thuộc tính (features) của đối tượng có thể thuộc các kiểu dữ liệu khác nhau như Nhị phân (Binary), Định danh (Nominal), Thứ tự (Ordinal), Số lượng (Quantitative), trong đó thuộc tính phân lớp phải có kiểu dữ liệu là Binary hoặc Ordinal.

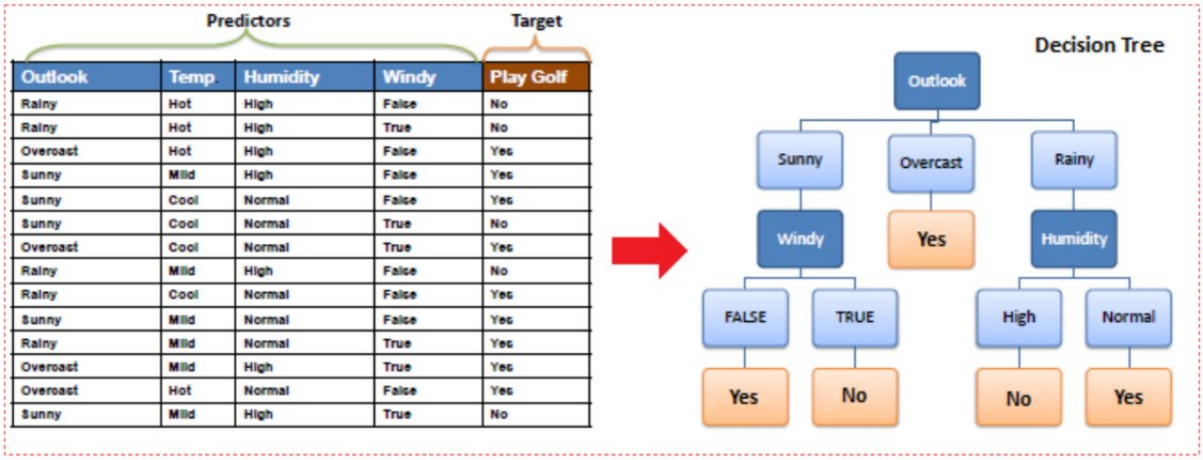


Minh họa cấu trúc của một cây quyết định cơ bản.

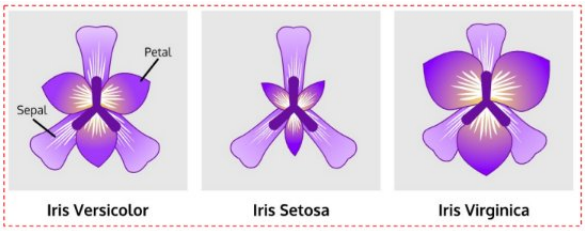
- Ví dụ về cây quyết định trong bài toán phân loại:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S

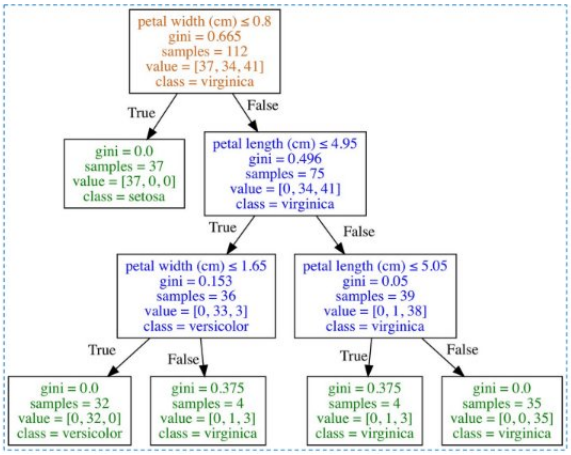




Outlook	Temp	Humidity	Windy	Play Golf
Rainy	Hot	High	False	No
Rainy	Hot	High	True	No
Overcast	Hot	High	False	Yes
Sunny	Mild	High	False	Yes
Sunny	Cool	Normal	False	Yes
Sunny	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Rainy	Mild	High	False	No
Rainy	Cool	Normal	False	Yes
Sunny	Mild	Normal	False	Yes
Rainy	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Sunny	Mild	High	True	No

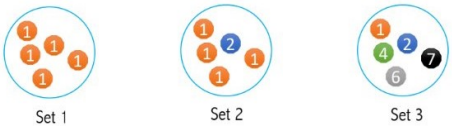


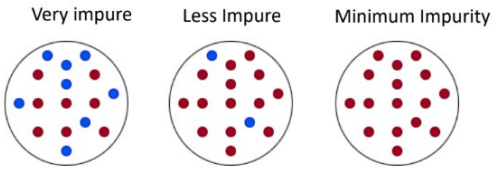
	sepal_length	sepal_width	petal_length	petal_width	species
0	5.1	3.5	1.4	0.2	Iris-setosa
1	4.9	3.0	1.4	0.2	Iris-setosa
2	4.7	3.2	1.3	0.2	Iris-setosa
3	4.6	3.1	1.5	0.2	Iris-setosa
4	5.0	3.6	1.4	0.2	Iris-setosa



- *Độ không thuần khiết* (Impurity) của một tập dữ liệu được minh họa như sau:

	Impurity Score
Set 1	Thấp
Set 2	Trung Bình
Set 3	Cao





- Để xây dựng một cây quyết định, ta tiến hành chia nhỏ dần dataset ban đầu theo hướng tạo ra những dataset nhỏ hơn có Impurity giảm nhiều nhất. Để có tiêu chí chọn ra nút gốc (root node) ở mỗi bước, ta có thể sử dụng giá trị *Gini Impurity* hoặc *Entropy - Information Gain*.

2 Xây dựng Decision Tree với giá trị Gini Impurity.

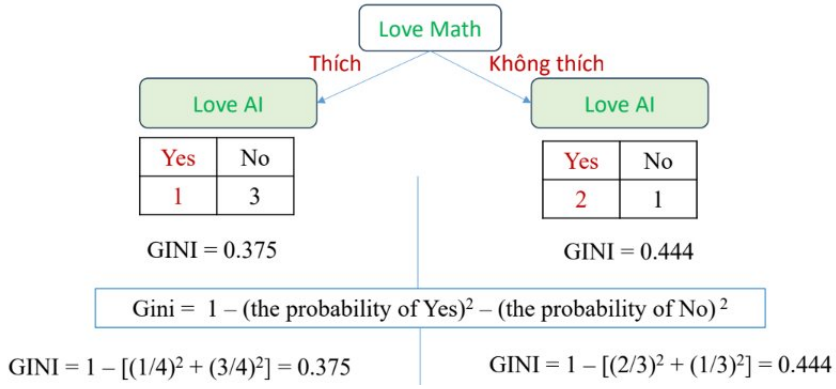
- Cho dataset D chứa các mẫu lấy từ k lớp. Ta có công thức Gini Impurity:

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

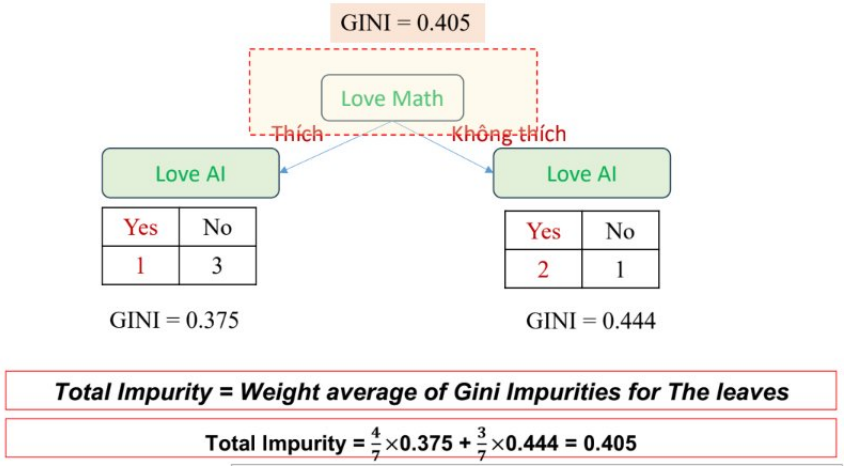
với p_i là xác suất một mẫu thuộc vào lớp thứ i.

	Count		Probability		Gini Impurity
	n_1	n_2	p_1	p_2	$1 - p_1^2 - p_2^2$
Node A	0	10	0	1	$1 - 0^2 - 1^2 = 0$
Node B	3	7	0.3	0.7	$1 - 0.3^2 - 0.7^2 = 0.42$
Node C	5	5	0.5	0.5	$1 - 0.5^2 - 0.5^2 = 0.5$

- Đối với các nút lá, ta tính giá trị Gini dựa theo công thức ở trên:

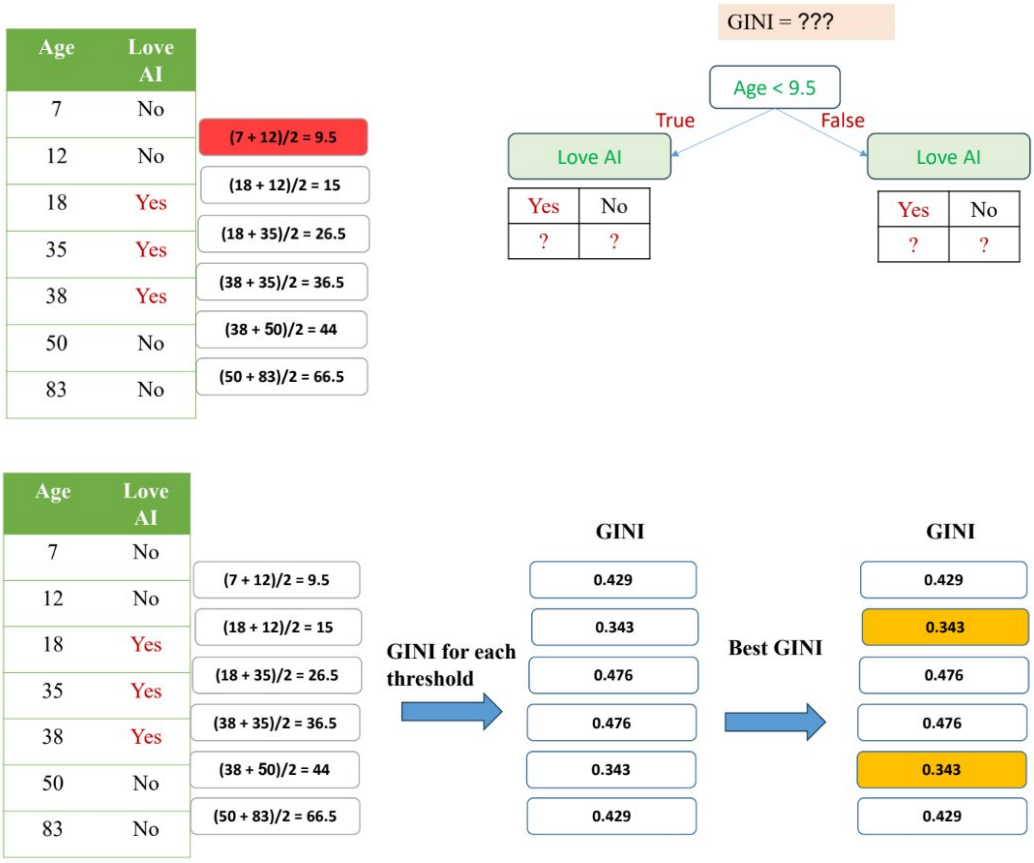


- Đối với các nút gốc, giá trị Gini là trung bình có trọng số các giá trị Gini tại các nút lá:



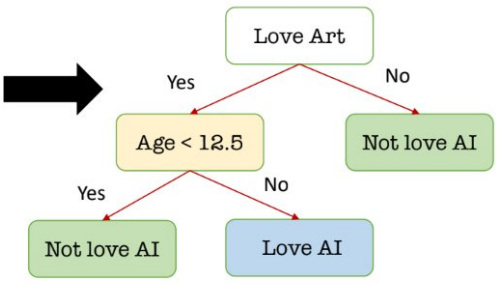
- Lưu ý: Trường hợp dataset có thuộc tính thuộc kiểu dữ liệu Ordinal (chẳng hạn như "Tuổi"), ta sắp xếp lại dữ liệu và tính giá trị Gini dựa theo việc chia các *ngưỡng* (threshold). Cách thực hiện

như bên dưới:



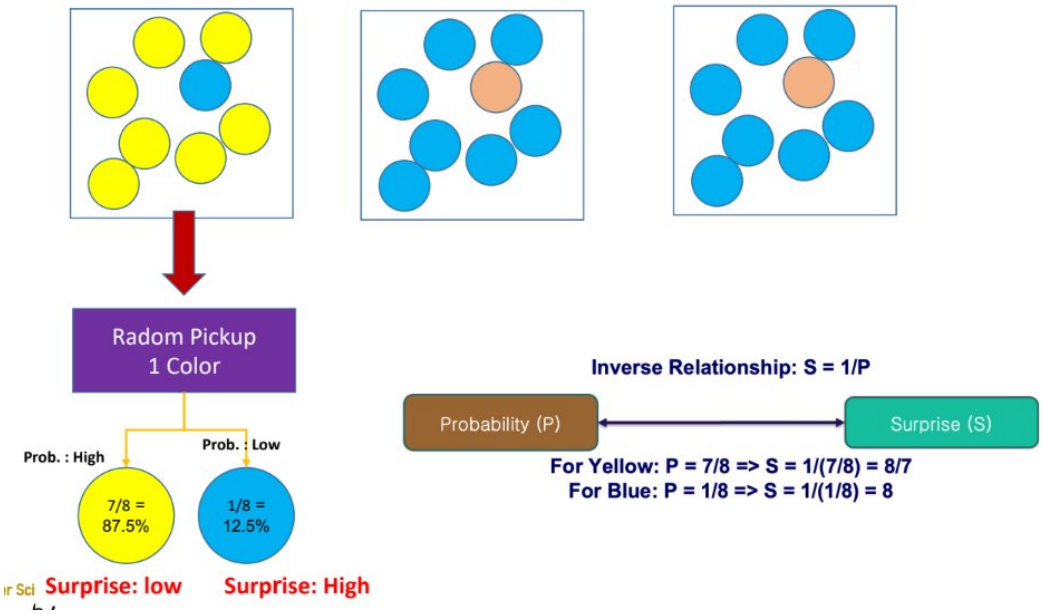
- Tại mỗi thời điểm chọn nút làm gốc, ta thử lần lượt từng thuộc tính, tính giá trị Gini thu được và so sánh, thuộc tính nào có giá trị Gini nhỏ nhất thì ta chọn nó làm nút gốc. Kết quả minh họa:

Love Math	Love Art	Age	Love AI
Yes	Yes	7	No
Yes	No	12	No
No	Yes	18	Yes
No	Yes	35	Yes
Yes	Yes	38	Yes
Yes	No	50	No
No	No	83	No



3 Xây dựng Decision Tree với giá trị Entropy - Information Gain.

- *Độ ngạc nhiên* (Suprise): $S = \dfrac{1}{P}$ tỉ lệ nghịch với xác suất P.



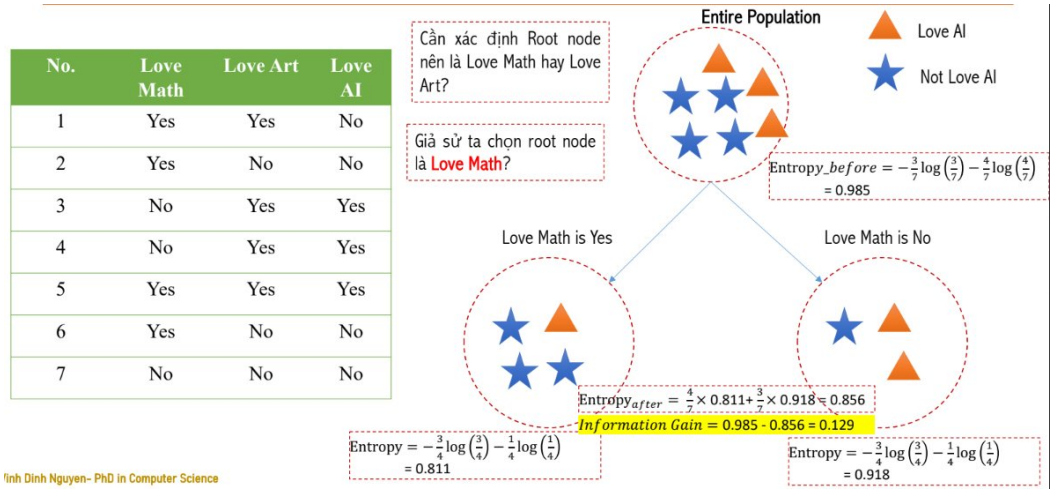
Vấn đề xảy ra khi dataset gồm các mẫu thuộc cùng một lớp ($P = 1$).
Khi đó $S = 1/1 = 1$ (không phù hợp vì độ ngạc nhiên phải nên bằng 0). vì thế ta sử dụng hàm logarit để tính toán về sau.

- Cho dataset D chứa các mẫu lấy từ k lớp. Ta có công thức Entropy:

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

với p_i là xác suất một mẫu thuộc vào lớp thứ i.

- Ta thực hiện xây dựng cây quyết định với ý tưởng giống với phần Gini Impurity, chỉ khác ở việc so sánh để chọn nút gốc. Nút gốc sẽ là thuộc tính có giá trị *Information Gain* $= E_{\text{parent}} - E_{\text{children}}$ lớn nhất.



4 Đánh giá về Decision Tree.

- Cây quyết định là một thuật toán đơn giản và phổ biến. Thuật toán này được sử dụng rộng rãi bởi những **ưu điểm** của nó:
  - 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.
  - Dữ liệu đầu vào có thể là dữ liệu missing, không cần chuẩn hóa hoặc tạo biến giả.
  - Có thể làm việc với cả dữ liệu số và dữ liệu phân loại.
  - Có thể xác thực mô hình bằng cách sử dụng các kiểm tra thống kê.
  - Có khả năng làm việc với dữ liệu lớn.
- Kèm với đó, cây quyết định cũng có những **nhược điểm** cụ thể:
  - Mô hình cây quyết định phụ thuộc rất lớn vào dữ liệu. 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.
  - Cây quyết định hay gặp vấn đề *overfitting*.