AI VIET NAM
@aivietnam.edu.vn

AI VIET NAM – AI COURSE 2025

Step-by-Step: Decision Tree Classification

Vũ Yến Linh Dương Đình Thắng Đinh Quang Vinh

I. Dẫn nhập

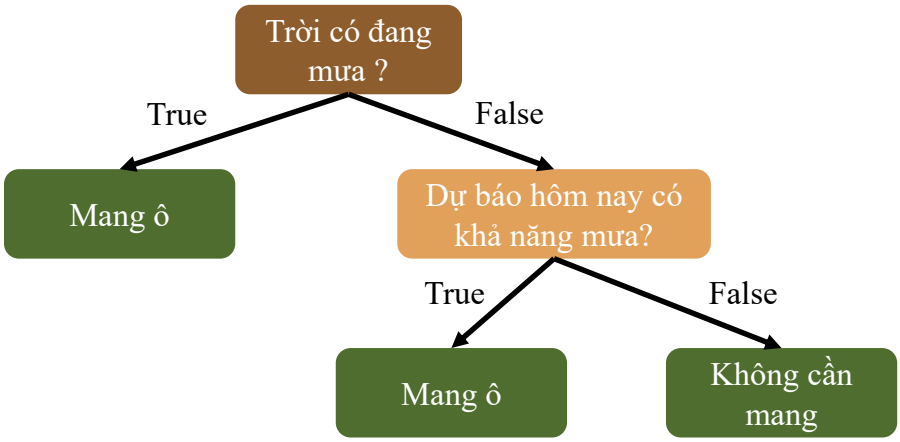
Có bao giờ chúng ta tự hỏi, máy tính đưa ra quyết định như thế nào khi phải phân loại email, dự đoán khách hàng tiềm năng hay xác định xem một bệnh nhân có triệu chứng nào đó hay không? Một trong những công cụ trực quan và dễ hiểu nhất để giúp máy thực hiện những nhiệm vụ này chính là **Decision Tree** (Tạm dịch: Cây quyết định).

Hình 1: Decision Tree là một dạng mô hình giúp đưa ra quyết định một cách logic, dễ hiểu và ứng dụng trong nhiều lĩnh vực thực tiễn.

Mục lục

I.1. Decision Tree là gì?

Hãy tưởng tượng ta đang vẽ một sơ đồ: khởi đầu bằng một câu hỏi chung nhất, ví dụ mỗi khi ra khỏi nhà ta phân vân có nên mang theo ô hay không, lúc này ta sẽ xem xét tình hình thời tiết với câu hỏi "Trời có đang mưa?". Mỗi khi trả lời "đúng" hoặc "sai", ta rẽ nhánh tiếp theo với câu hỏi mới, và cứ tiếp tục như vậy cho đến khi ta chốt được quyết định cuối cùng. Decision Tree chính là mô hình máy học mô phỏng quy trình đó. Khi cây dùng để phân loại dữ liệu thành các nhãn rời rạc, ta gọi nó là **Classification Tree** (Cây phân loại). Nếu dùng để dự đoán giá trị số liên tục, ta gọi là **Regression Tree** (Cây hồi quy). Trong bài này, chúng ta sẽ tập trung vào tìm hiểu **Classification Tree**.



Hình 2: Minh họa Classification Tree đơn giản để xác định có nên mang ô, xét lần lượt điều kiện mưa hiện tại và dự báo.

Decision Tree có thể xử lý đồng thời ba dạng dữ liệu:

- **Dữ liệu rời rạc** (Categorical data): với các biến mang giá trị hữu hạn như "Đúng/Sai", cây sẽ tạo nhánh tương ứng cho từng giá trị.
- **Dữ liệu liên tục** (Continuous/Numeric data): với các biến số thực như "Tuổi" hay "Điểm". Thuật toán sẽ tìm ngưỡng (threshold) tối ưu.
- **Dữ liệu hỗn hợp** (Mixed data): khi tập dữ liệu chứa cả biến rời rạc và biến liên tục, Decision Tree vẫn có thể kết hợp cả hai.

Thời tiết	Tốc độ gió	Mang ô
Mưa	Nhẹ	Có
Nhiều mây	Mạnh	Không
Nắng	Nhẹ	Không

Dữ liệu rời rạc
(Categorical data)

Nhiệt độ	Độ ẩm	Mang ô
22	95	Có
25	60	Không
30	50	Không

Dữ liệu liên tục
(Continuous/Numeric data)

Thời tiết	Độ ẩm	Mang ô
Mưa	95	Có
Nhiều mây	60	Không
Nắng	50	Không

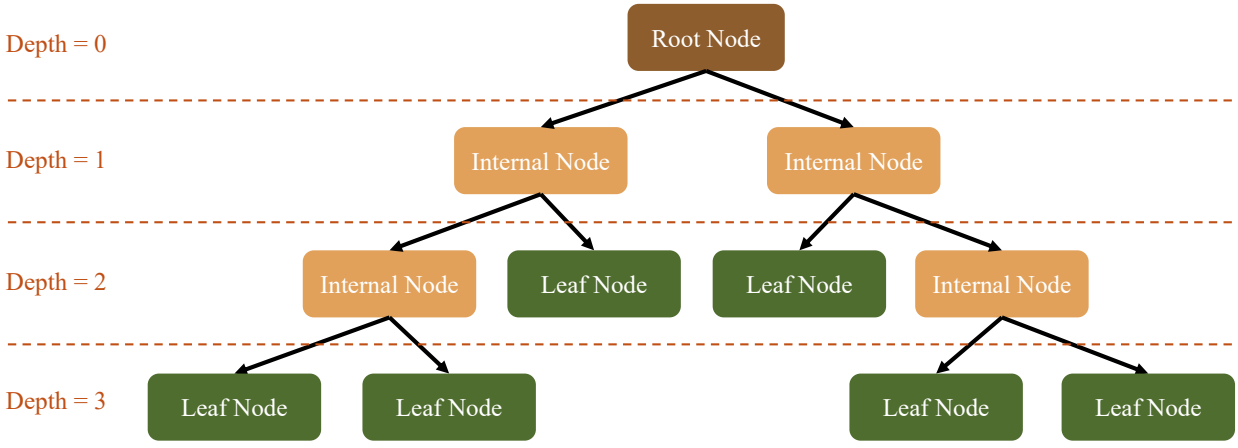
Dữ liệu hỗn hợp
(Mixed data)

Hình 3: Minh họa các dạng dữ liệu của Decision Tree.

I.2. Một số thuật ngữ chính

Bảng thuật ngữ chính

Thuật ngữ	Định nghĩa
Root node (nốt gốc)	Nốt đầu tiên, là vị trí khởi đầu, nằm ở đỉnh cao nhất của cây.
Internal node (nốt phân nhánh)	Mỗi nốt thực hiện một điều kiện kiểm tra, có các nhánh con.
Split (phân tách)	Thao tác chia tập dữ liệu tại một Internal node dựa trên giá trị hoặc ngưỡng của thuộc tính, tạo ra các nhánh con.
Branch (nhánh)	Đường kết nối giữa các nút, thể hiện kết quả True/False hoặc giá trị của biến rời rạc.
Leaf node (nốt lá)	Nốt cuối cùng, không còn nhánh con, đưa ra nhãn phân loại.
Depth (độ sâu)	Độ sâu của cây (hay chiều cao cây), là số nhánh dài nhất tính từ root đến một leaf.
Pure (thuần/ tinh khiết)	Là node mà tất cả các mẫu trong nó đều thuộc cùng một nhãn.
Impure (không thuần/không tinh khiết)	Là node có hai hoặc nhiều hơn các nhãn khác nhau xuất hiện trong tập mẫu.



Hình 4: Minh họa các thuật ngữ được sử dụng trong một Decision Tree.

II. Thuật toán phân tách trong Decision Tree: Entropy & Gini Impurity

Để thuận tiện theo dõi và đảm bảo rõ ràng, bảng dưới đây tập hợp các ký hiệu toán học quan trọng sẽ được sử dụng xuyên suốt tài liệu. Mỗi ký hiệu đi kèm với lời giải thích ngắn gọn giúp nắm bắt ý nghĩa và công thức liên quan một cách nhanh chóng.

Bảng ký hiệu toán học

Ký hiệu	Ý nghĩa
X, x_i	Biến ngẫu nhiên X và giá trị thứ i trong tập giá trị $\{x_i\}$.
$p_i = P(X = x_i)$	Xác suất để X nhận giá trị x_i.
$\mu = \mathbb{E}[X]$	Kỳ vọng (mean) của X, $\sum_i x_i\, p_i$.
$\sigma^2 = \mathrm{Var}(X)$	Phương sai của X, $\sum_i (x_i - \mu)^2\, p_i$.
$\sigma = \sqrt{\sigma^2}$	Độ lệch chuẩn của X.
S	Tập dữ liệu ban đầu.
$\mathcal{C}$	Tập các nhãn (classes) trong S.
p_c	Xác suất có nhãn c trong S.
$I(E)$	Lượng thông tin (information content) của biến cố E (đơn vị: bit).
$H(S)$	Entropy của S, đo độ hỗn độn của một tập dữ liệu.
$\mathrm{Vals}(A)$	Tập các giá trị (hoặc ngưỡng) của thuộc tính A.
S_v	Tập con của S gồm các mẫu có $A = v$.
$\mathrm{IG}(S, A)$	Information Gain: độ giảm trung bình của độ hỗn độn (entropy) khi phân tách tập S theo thuộc tính A.
$G(S)$	Gini Impurity: mức "không tinh khiết"/"không thuần" (impurity) của tập S theo thước đo Gini, phản ánh độ không thuần của nhãn trong S.
$\mathrm{GG}(S, A)$	Gini Gain: độ giảm impurity theo Gini khi phân tách tập S theo thuộc tính A, tương tự IG nhưng dùng thước đo Gini.
$\lvert S\rvert$, $\lvert S_v\rvert$	Kích thước tập S và tập con S_v.
`True`/`False`	Nhãn rẽ nhánh cho kết quả kiểm tra điều kiện tại mỗi node.

II.1. Nhắc lại: Mean, Variance và Standard Deviation

Cho biến ngẫu nhiên rời rạc X với các giá trị $\{x_i\}$ và xác suất $p_i = P(X = x_i)$.

- **Mean (Kỳ vọng)**
$$\mu = \mathbb{E}[X] = \sum_i x_i \, p_i.$$

- **Variance (Phương sai)**
$$\sigma^2 = \mathrm{Var}(X) = \mathbb{E}\Big[(X - \mu)^2\Big] = \sum_i (x_i - \mu)^2 \, p_i.$$

- **Standard Deviation (Độ lệch chuẩn)**
$$\sigma = \sqrt{\sigma^2} = \sqrt{\sum_i (x_i - \mu)^2 \, p_i}.$$

II.2. Entropy & Information Gain

Trong một sự kiện E: Một chiếc túi trong đó có 9 viên bi đỏ và 1 viên bi xanh, xác suất lấy ra:

$$P(\text{đỏ}) = \frac{9}{10} = 0.9, \qquad P(\text{xanh}) = \frac{1}{10} = 0.1.$$

Vì xác suất nhỏ hơn, ta thấy ngạc nhiên hơn khi rút được viên bi xanh so với viên bi đỏ. Vậy làm thế nào để đo lường được "độ ngạc nhiên" này?

II.2.1. Đề xuất ban đầu

$$\mathrm{Surprise}(E) = \frac{1}{P(E)}.$$

Hàm này thỏa điều kiện: sự kiện càng hiếm ($P(E) \downarrow$) thì "độ ngạc nhiên" càng cao $\mathrm{Surprise}(E) \uparrow$. Tuy nhiên, tồn tại một số vấn đề

1. Đơn vị mơ hồ: $P(E) = 0.1 \Rightarrow \mathrm{Surprise} = 10$. vậy "10" này là đơn vị gì? Ta không thể nói là 10 ngạc nhiên.

2. Không cộng dồn được: với hai biến cố độc lập E_1, E_2, ta mong
$$\mathrm{Surprise}(E_1 \cap E_2) = \mathrm{Surprise}(E_1) + \mathrm{Surprise}(E_2),$$
nhưng thực tế
$$\mathrm{Surprise}(E_1 \cap E_2) = \frac{1}{P(E_1 \cap E_2)} = \frac{1}{P(E_1)P(E_2)} = \mathrm{Surprise}(E_1) \times \mathrm{Surprise}(E_2).$$

Từ các tiên đề lý thuyết thông tin (Information Theory) của Shannon (1948) [1], [2], ông đã chỉ ra rằng một hàm $I(p)$ đo thông tin của xác suất p phải thỏa:

1. Tính chất liên tục theo p.
2. I(p) giảm khi p tăng.
3. *Tính cộng dồn* cho biến cố độc lập: $I(p_1 p_2) = I(p_1) + I(p_2)$.

Thì duy nhất (ngoại trừ hệ số tỉ lệ) có

$$\boxed{I(E) \;=\; -\log_2 P(E)} \quad \text{(đơn vị: bit)}.$$

Với E_1, E_2 độc lập, thỏa phép cộng dồn:

$$I(E_1 \cap E_2) = -\log_2\Big(P(E_1)P(E_2)\Big) = I(E_1) + I(E_2).$$

Như vậy Shannon đã chứng minh chỉ có hàm $(-log_p)$ mới thỏa mãn được các điều kiện trên. Về mặt toán học, ta có thể dùng bất kì cơ số nào để định nghĩa hàm thông tin (base e, base 2,...), ở đây ta sẽ dùng cơ số 2 tức log_2 để đúng theo chuẩn mực của lý thuyết thông tin.

II.2.2. Entropy & Information Gain

Từ hàm $I(E)$, **Entropy** H của một tập dữ liệu S với phân phối xác suất $\{p_c\}_{c\in C}$ được định nghĩa:

$$\boxed{H(S) = -\sum_{c\in\mathcal{C}} p_c \log_2 p_c} \quad \text{(đơn vị: bit)},$$

trong đó

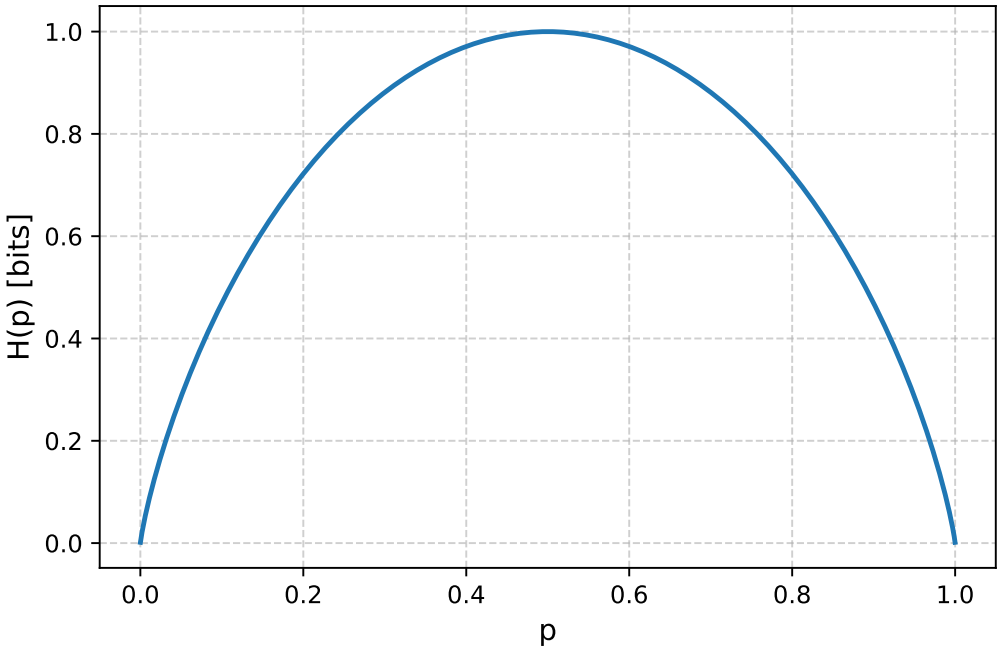
- S là tập dữ liệu ban đầu.
- $\mathcal{C}$ là tập các nhãn (classes).
- p_c là xác suất có nhãn c trong S.

Entropy càng lớn thì tập dữ liệu càng "không thuần", càng khó phân tách.

Lưu ý:

- $H(S) = 0$ khi S chỉ có 1 lớp, được gọi là thuần (Pure).
- $H(S) = 1$ đạt cực đại khi S có 2 nhãn phân phối đều (phân loại nhị phân).
- Với $\mathcal{C} > 2$, tức nhiều hơn 2 nhãn mà dữ liệu phân phối đều thì $H(S) = log_2|\mathcal{C}|$

Để hiểu tại sao lại như vậy, ta nhìn vào hình minh họa bên dưới



Hình 5: Đồ thị của hàm entropy $H(p) = -p\log_2 p - (1-p)\log_2(1-p)$ với $p \in (0, 1)$.

Từ đồ thị của hàm Entropy, ta có thể thấy:

- Khi p tiến gần đến 0 hoặc 1($p \to 0$ hoặc $p \to 1$), thì $H(p)$ tiến gần đến 0 ($H(p) \to 0$). Điều này phản ánh rằng nếu một nhãn gần như chắc chắn xảy ra (xác suất gần 0 hoặc gần 1), thì không có gì để "ngạc nhiên".
- Hàm đạt cực đại $H(p) = 1$ tại $p = 0.5$. Chú ý rằng giá trị 1 (bit) chỉ đúng trong trường hợp tập dữ liệu phân phối đều khi chỉ có 2 nhãn/2 khả năng xảy ra (phân loại nhị phân) ví dụ như "Đúng/Sai", "Có/Không", "Dương tính/Âm tính". Với nhiều lớp ($|\mathcal{C}| > 2$), ví dụ như "Tích cực/Trung Tính/Tiêu cực" thì giá trị cực đại của Entropy sẽ lớn hơn 1 (cụ thể là $\log_2|\mathcal{C}|$) và xảy ra khi các nhãn phân phối đều.
- Một cách trực quan, chiều cao của đường cong tại mỗi p cho biết mức độ "hỗn độn trung bình" của nhãn: càng gần hai đầu (nhãn thuần), Entropy càng thấp; càng gần trung tâm (phân phối đều), Entropy càng cao.

Khi phân tách tập mẫu S theo thuộc tính A thành các nhóm con $\{S_v\}$ với $v \in \text{Vals}(A)$, **Information Gain** được định nghĩa

$$\boxed{\text{IG}(S, A) = H(S) - \sum_{v \in \text{Vals}(A)} \frac{|S_v|}{|S|} H(S_v)},$$

trong đó

- A là thuộc tính ta định tách.
- S_v là tập con của S gồm các mẫu có $A = v$.

- $\frac{|S_v|}{|S|}$ là trọng số theo kích thước tập con.

Information Gain chính là độ giảm trung bình của entropy sau khi tách theo thuộc tính A. Thuộc tính có giá trị IG càng lớn thì khả năng phân tách càng tốt.

II.3. Gini Impurity & Gini Gain

Theo Breiman et al. (1984), **Gini Impurity** của tập mẫu S được định nghĩa bởi

$$\boxed{G(S) \;=\; \sum_{c\in\mathcal{C}} p_c(1-p_c) \;=\; 1-\sum_{c\in\mathcal{C}} p_c^2},$$

trong đó

- $\mathcal{C}$ là tập các nhãn (classes).
- p_c là xác suất có nhãn c trong S.
- $G(S)$ đo mức "không tinh khiết" hoặc "không thuần" (impurity) trong phân phối nhãn.

Khi tách S theo thuộc tính A thành các tập con $\{S_v\}$ với $v \in \text{Vals}(A)$, **Gini Gain** có dạng:

$$\boxed{\text{GG}(S,A) = G(S) \;-\; \sum_{v\in\text{Vals}(A)} \frac{|S_v|}{|S|}\, G(S_v)},$$

trong đó

- S_v là tập con của S gồm các mẫu có $A = v$.
- $\frac{|S_v|}{|S|}$ là trọng số theo kích thước tập con.
- $\text{GG}(S,A)$ đo mức giảm impurity thu được nhờ phân tách theo A.

Lưu ý:

- $G(S) = 0$ khi S đạt độ thuần (Pure).
- $G(S) = 0.5$ khi S có 2 nhãn phân phối đều (phân loại nhị phân).
- Với $\mathcal{C} > 2$, tức nhiều hơn 2 nhãn mà dữ liệu phân phối đều thì $G(S) = 1 - \frac{1}{|\mathcal{C}|}$.

Sau khi tìm hiểu qua về 2 thuật toán phân tách, ta có thể tổng hợp được cơ bản

Bảng so sánh Entropy/Information Gain và Gini/Gini Gain

Giống nhau	Khác nhau
• Đều đánh giá chất lượng phân tách dựa trên mức độ "thuần" (Pure) của nhãn. • Đều sử dụng trọng số $\frac{\|S_v\|}{\|S\|}$ khi kết hợp giá trị trên các nhánh con.	• Entropy/Information Gain sử dụng hàm log để đo "độ ngạc nhiên" trung bình; Gini/Gini Gain không dùng log, độ phức tạp tính toán thấp hơn. • Entropy nhạy hơn với các nhãn có tần suất nhỏ (hiếm); Gini Impurity ưu tiên tạo các node thuần và ít dao động theo tần suất.

II.4. Xây dựng một Classification Tree (Step by Step)

- **Bước 1:** Tính thống kê gốc
  - Tính Root Entropy $H(S)$ (Entropy của cột nhãn) cho toàn bộ mẫu.
  - Tương tự tính Root Gini $G(S)$ cho toàn bộ mẫu.
- **Bước 2:** Tính *Gain* cho từng thuộc tính
  - Với Entropy:
$$\mathrm{IG}(S, A) = H(S) - \sum_{v \in \mathrm{Vals}(A)} \frac{|S_v|}{|S|} H(S_v).$$
  - Với Gini:
$$\mathrm{GG}(S, A) = G(S) - \sum_{v \in \mathrm{Vals}(A)} \frac{|S_v|}{|S|} G(S_v).$$
- **Bước 3:** Chọn thuộc tính phân nhánh
  - Chọn thuộc tính có IG hoặc GG cao nhất để phân tách.
  - Phân tách dữ liệu, lặp lại Bước 1–3 cho mỗi node con đến khi các node thuần/tinh khiết (Pure) hoặc hết thuộc tính.
- **Bước 4:** Ứng dụng cây để dự đoán
  - Với dữ liệu mới, từ root node đi theo nhánh phù hợp đến leaf node, lấy nhãn dự đoán.

III. Thực hành

III.1. Mô tả bài toán

Cho bộ dữ liệu gồm 8 hồ sơ thí sinh sau, mỗi dòng là một thí sinh với ba thông tin đầu vào và kết quả trúng tuyển đại học.

STT	Điểm Tốt Nghiệp	Chứng chỉ IELTS	Cộng Điểm Dân Tộc	Đậu Đại Học
1	12.0	Không	Không	Không
2	14.5	Không	Có	Không
3	16.0	Không	Không	Không
4	18.0	Không	Có	Không
5	20.0	Có	Không	Có
6	22.0	Không	Không	Không
7	24.0	Có	Có	Có
8	26.0	Không	Không	Có

Bảng 1: Bộ dữ liệu điểm tốt nghiệp và kết quả trúng tuyển đại học.

Phân tích sơ bộ dữ liệu

- **Điểm Tốt Nghiệp** (numeric): Thuộc tính này cho biết khả năng học vấn của thí sinh, và vì là biến liên tục nên khi tách nhánh ta sẽ phải tìm ngưỡng (threshold) phù hợp.
- **Chứng chỉ IELTS** (categorical): Nhóm thí sinh theo việc có hay không chứng chỉ IELTS. Đây là biến rời rạc với hai giá trị "Có" hoặc "Không".
- **Cộng Điểm Dân Tộc** (categorical): Cho biết thí sinh có được cộng điểm ưu tiên theo chính sách dân tộc hay không. Giá trị cũng là "Có" hoặc "Không".
- **Đậu Đại Học** (label): Nhãn mục tiêu mà chúng ta muốn dự đoán. "Có" nếu thí sinh trúng tuyển, ngược lại là "Không".

Yêu cầu: xây dựng Classification Tree để dự đoán một sinh viên có trúng tuyển đại học không từ ba thuộc tính trên, theo hai cách:

1. Entropy & Information Gain
2. Gini Impurity & Gini Gain

Sau khi hoàn thiện, hãy thử dự đoán kết quả cho dữ liệu của một học sinh mới:

$$(\text{Điểm Tốt Nghiệp} = 21.0,\ \text{Chứng chỉ IELTS} = \text{Không},\ \text{Cộng Điểm Dân Tộc} = \text{Có}).$$

III.2. Lời giải (Solution)

Bước 1. Tính thống kê gốc

Tập dữ liệu ban đầu S_0 đã cho gồm 8 học sinh, với nhãn "Có" = 3 và "Không" = 5, ta có thể viết $S_0 = \{3\,\text{Có}, 5\,\text{Không}\}$; như vậy ta tính được:

$$p_{\text{Có}} = \frac{3}{8}, \quad p_{\text{Không}} = \frac{5}{8}.$$

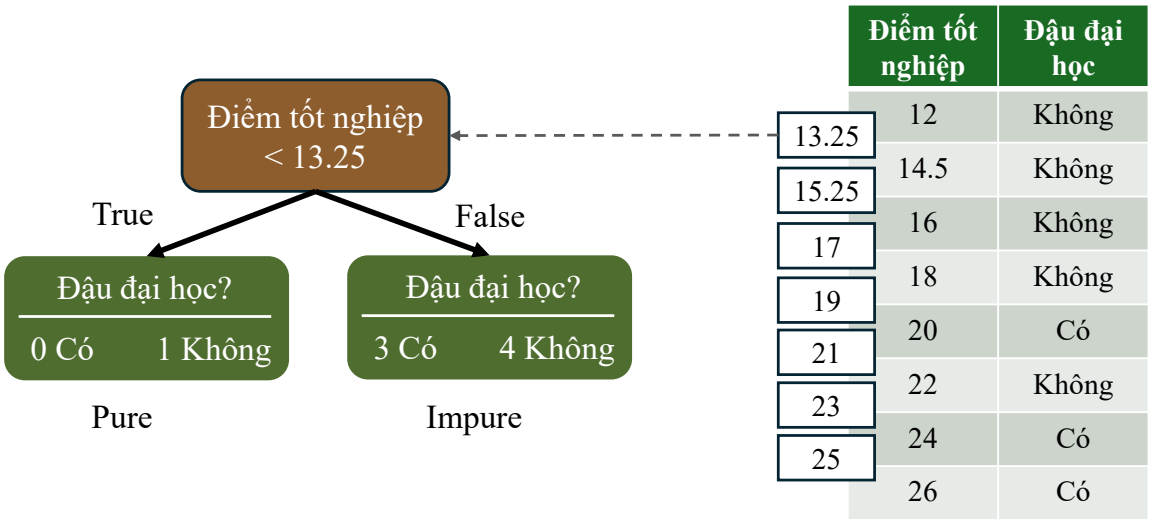
$$H(S_0) = -\sum_{c\in\mathcal{C}} p_c \log_2 p_c = -\frac{3}{8}\log_2\frac{3}{8} - \frac{5}{8}\log_2\frac{5}{8} = 0.9544.$$

$$G(S_0) = 1 - \sum_{c\in\mathcal{C}} p_c^2 = 1 - \left(\left(\frac{3}{8}\right)^2 + \left(\frac{5}{8}\right)^2\right) = 0.4688.$$

Bước 2. Tính Gain cho từng thuộc tính

1) Điểm Tốt Nghiệp (numeric):

1. Với dữ liệu dạng số hay liên tục, ta cần sắp xếp theo thứ tự từ giá trị nhỏ đến lớn. Vì dữ liệu cho sẵn đã theo thứ tự đó, ta không cần thay đổi gì thêm.
2. Sau đó lấy giá trị trung bình của từng cặp liền kề, ta được 7 ngưỡng mới.
3. Cuối cùng tính IG/GG cho từng ngưỡng đó:



Hình 6: Minh họa các bước xử lý dữ liệu dạng numeric.

Với ngưỡng 13.25, xét "Điểm tốt nghiệp < 13.25". Đi theo nhánh True, ta thấy chỉ có 1 người có điểm tốt nghiệp bằng 12 là thỏa mãn nhỏ hơn 13.25. Và ứng với điểm đó là nhãn "Không" đậu đại học. Do node này chỉ chứa 1 nhãn duy nhất là "Không", ta gọi đây là node thuần hoặc node tinh khiết (Pure), và đối với node thuần thì Entropy của nó bằng 0 như đã trình bày trong phần Lưu ý mục II.2.2.

$$H(\text{True}) = 0$$

Với nhánh False, tức "Điểm tốt nghiệp ≥ 13.25" ta thấy có tới 7 người có điểm số đáp ứng điều kiện đó, trong 7 người đó lại có 3 người đậu đại học và 4 người không, ứng với 3 nhãn "Có" và 4 nhãn "Không". Do node này chứa cả hai nhóm học sinh "Có" và "Không" đậu đại học, ta gọi node này là không thuần hoặc không tinh khiết (Impure).

$$H(\text{False}) = -\frac{3}{7}\log_2\frac{3}{7} - \frac{4}{7}\log_2\frac{4}{7} = 0.9852$$

Ta tính được Information Gain:

$$\begin{aligned} IG\Big(S_0,\ \text{Điểm tốt nghiệp} < 13.25\Big) &= H(S_0) - \left(\frac{S_{\text{True}}}{S}H(\text{True}) + \frac{S_{\text{False}}}{S}H(\text{False})\right) \\ &= 0.9544 - \left(\frac{1}{8}\times 0 + \frac{7}{8}\times 0.9852\right) \\ &= 0.9544 - 0.8620 \\ &= 0.0924 \end{aligned}$$

Với Gini Impurity, khi gặp node thuần thì Gini của nó cũng bằng 0 (xem lại phần Lưu ý mục II.3.):

$$G(\text{True}) = 0$$

$$G(\text{False}) = 1 - \left(\left(\frac{3}{7}\right)^2 + \left(\frac{4}{7}\right)^2\right) = 0.4898$$

$$\begin{aligned} GG\Big(S_0,\ \text{Điểm tốt nghiệp} < 13.25\Big) &= G(S_0) - \left(\frac{S_{\text{True}}}{S}G(\text{True}) + \frac{S_{\text{False}}}{S}G(\text{False})\right) \\ &= 0.4688 - \left(\frac{1}{8}\times 0 + \frac{7}{8}\times 0.4898\right) \\ &= 0.4688 - 0.4286 \\ &= 0.0402 \end{aligned}$$

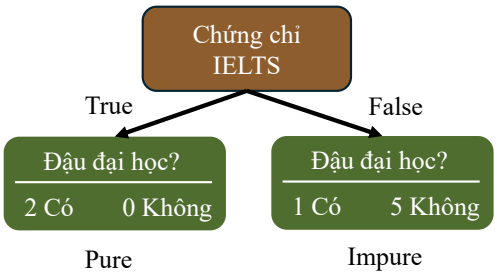
Tương tự, ta tính lần lượt cho từng ngưỡng và thu bảng kết quả IG và GG

Ngưỡng Điểm	**Information Gain**	**Gini Gain**
13.25	0.0924	0.0402
15.25	0.2044	0.0938
17.00	0.3476	0.1688
19.00	**0.5488**	**0.2812**
21.00	0.1589	0.1021
23.00	0.4669	0.2605
25.00	0.1992	0.1116

Bảng 2: So sánh Information Gain và Gini Gain cho các ngưỡng điểm trên tập S_0.

4. Chọn ngưỡng có giá trị IG/GG cao nhất để làm đại diện cho thuộc tính điểm tốt nghiệp: 19.00 (IG=0.5488, GG=0.2812).

2) Chứng chỉ IELTS:



Điểm tốt nghiệp	Chứng chỉ IELTS	Cộng điểm dân tộc	Đậu đại học
12	Không	Không	Không
14.5	Không	Có	Không
16	Không	Không	Không
18	Không	Có	Không
20	Có	Không	Có
22	Không	Không	Không
24	Có	Có	Có
26	Không	Không	Có

Hình 7: Minh hoạ quá trình tách nhánh dữ liệu với thuộc tính "Chứng chỉ IELTS".

Với nhánh True, node này chỉ chứa 1 nhãn duy nhất là "Có", là node thuần, Entropy có dạng

$$H(\text{True}) = 0.$$

Với nhánh False,

$$H(\text{False}) = -\frac{1}{6}\log_2\frac{1}{6} - \frac{5}{6}\log_2\frac{5}{6} = 0.650.$$

Ta tính được Information Gain:

$$\begin{aligned} IG\big(S_0, \text{Chứng chỉ IELTS}\big) &= H(S_0) - \left(\frac{2}{8}\times 0 + \frac{6}{8}\times 0.650\right) \\ &= 0.9544 - 0 - 0.4875 \\ &= 0.4669 \end{aligned}$$

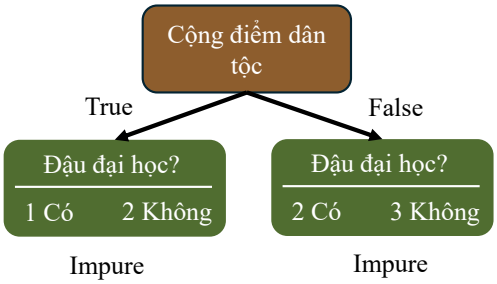
Với Gini Impurity:

$$G(\text{True}) = 0.$$

$$G(\text{False}) = 1 - \left(\left(\frac{1}{6}\right)^2 + \left(\frac{5}{6}\right)^2\right) = 0.2778.$$

$$\begin{aligned} GG\big(S_0, \text{Chứng chỉ IELTS}\big) &= G(S_0) - \left(\frac{2}{8}\times 0 + \frac{6}{8}\times 0.2778\right) \\ &= 0.4688 - 0 - 0.20835 \\ &= 0.2604 \end{aligned}$$

3) Cộng Điểm Dân Tộc:



Điểm tốt nghiệp	Chứng chỉ IELTS	Cộng điểm dân tộc	Đậu đại học
12	Không	Không	Không
14.5	Không	Có	Không
16	Không	Không	Không
18	Không	Có	Không
20	Có	Không	Có
22	Không	Không	Không
24	Có	Có	Có
26	Không	Không	Có

Hình 8: Minh hoạ quá trình tách nhánh dữ liệu với thuộc tính "Cộng điểm dân tộc".

Thuộc tính này sau khi phân tách thì không có node thuần nào, ta áp dụng công thức tính toán như bình thường.

Với Entropy:

$$H(\text{True}) = -\frac{1}{3}\log_2\frac{1}{3} - \frac{2}{3}\log_2\frac{2}{3} = 0.9183.$$

$$H(\text{False}) = -\frac{2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2\frac{3}{5} = 0.9710.$$

$$\begin{aligned} IG\Big(S_0, \text{Cộng điểm dân tộc}\Big) &= 0.9544 - \left(\frac{3}{8}\times 0.9183 + \frac{5}{8}\times 0.9710\right) \\ &= 0.0032 \end{aligned}$$

Với Gini Impurity:

$$G(\text{True}) = 1 - \left(\left(\frac{1}{3}\right)^2 + \left(\frac{2}{3}\right)^2\right) = 0.4444.$$

$$G(\text{False}) = 1 - \left(\left(\frac{2}{5}\right)^2 + \left(\frac{3}{5}\right)^2\right) = 0.4800.$$

$$\begin{aligned} GG\Big(S_0, \text{Cộng điểm dân tộc}\Big) &= 0.4688 - \left(\frac{3}{8}\times 0.4444 + \frac{5}{8}\times 0.4800\right) \\ &= 0.0021 \end{aligned}$$

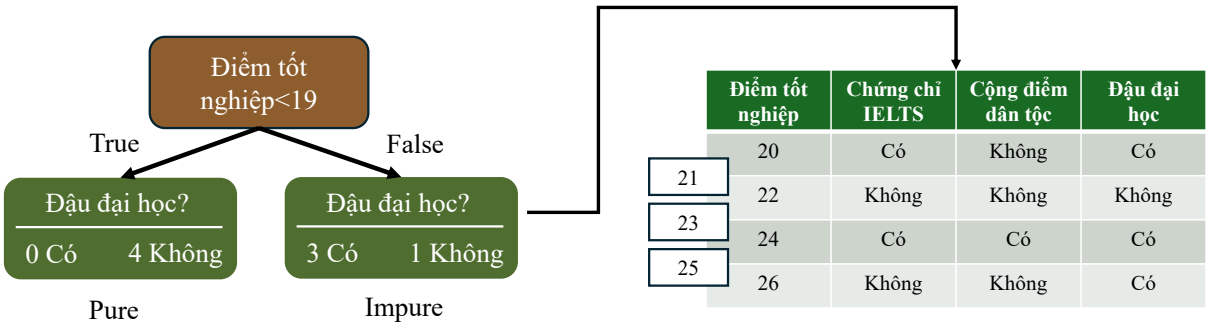
Bước 3. Chọn thuộc tính phân nhánh:

Thuộc tính	Information Gain	Gini Gain
Điểm Tốt Nghiệp < 19	**0.5488**	**0.2812**
Chứng chỉ IELTS	0.4669	0.2604
Cộng Điểm Dân Tộc	0.0032	0.0021

Bảng 3: So sánh Information Gain và Gini Gain cho ba thuộc tính trên tập S_0.

Dựa theo kết quả từ bước 2, ta chọn được thuộc tính có IG/GG cao nhất "Điểm Tốt Nghiệp < 19" để làm Root node của Decision Tree và thực hiện phân nhánh lần thứ nhất

Decision Tree cấp 1 (depth=1):



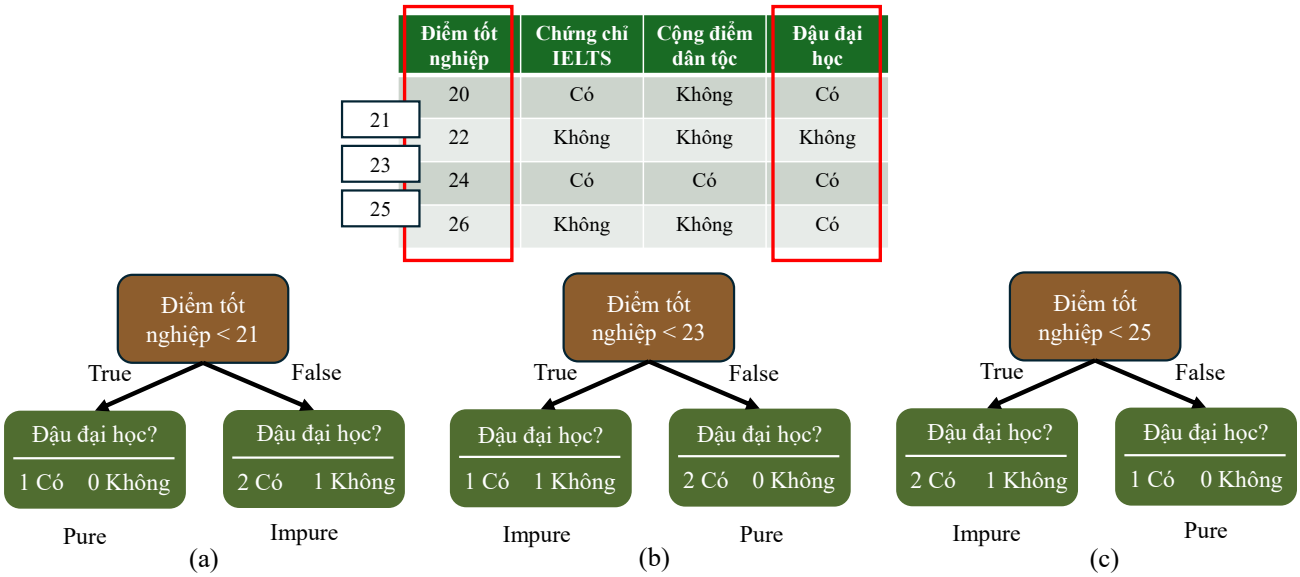
Hình 9: Minh hoạ quá trình tách nhánh dữ liệu của Decision Tree với thuộc tính "Điểm tốt nghiệp<19".

Lá bên nhánh "True" với $S_{\text{True}} = \{0\,\text{Có}, 4\,\text{Không}\}$ đã đạt được độ thuần và không cần phải phân tách tiếp nữa. Tuy nhiên lá bên nhánh "False" tức "Điểm $\geq$ 19" vẫn còn chưa tinh khiết $S_{\text{False}} = \{3\,\text{Có}, 1\,\text{Không}\}$, ta tiếp tục dựa trên dữ liệu của những thí sinh nằm trong nhóm đó để tìm thuộc tính phân tách. Tức lặp lại các bước 1-3.

Gọi dữ liệu của nhóm "Điểm $\geq$ 19" là S_1, nhìn vào bảng dữ liệu của nhóm này trên Hình 9 ta thấy có 3 nhãn "Có", và 1 nhãn "Không", tức $S_1 = \{3\,\text{Có}, 1\,\text{Không}\}$, ta có :

$$H(S_1) = -\frac{3}{4}\log_2\frac{3}{4} - \frac{1}{4}\log_2\frac{1}{4} = 0.8113, \quad G(S_1) = 1 - \Big((\frac{3}{4})^2 + (\frac{1}{4})^2\Big) = 0.3750.$$

1) Điểm Tốt Nghiệp (numeric):



Hình 10: Minh hoạ quá trình tách nhánh dữ liệu với các ngưỡng thuộc tính "Điểm tốt nghiệp" của tập dữ liệu S_1.

Nhìn vào Hình 10, với mỗi ngưỡng đều có một phân nhánh đạt độ thuần, Entropy và Gini của node đó đều bằng 0, lúc này việc tính toán dễ dàng hơn rất nhiều. Ta có thể tính trực tiếp (a):

$$\begin{aligned} IG\big(S_1, \text{Điểm tốt nghiệp} < 21\big) &= H(S_1) - \left(\frac{1}{4}\cdot 0 + \frac{3}{4}\left(-\frac{2}{3}\log_2\frac{2}{3} - \frac{1}{3}\log_2\frac{1}{3}\right)\right) \\ &= 0.8113 - 0 - \frac{3}{4}\times 0.9182 \\ &= 0.1226 \end{aligned}$$

$$\begin{aligned} GG\big(S_1, \text{Điểm tốt nghiệp} < 21\big) &= G(S_1) - \left(\frac{1}{4}\cdot 0 + \frac{3}{4}\left(1 - \left(\frac{2}{3}\right)^2 - \left(\frac{1}{3}\right)^2\right)\right) \\ &= 0.3750 - 0 - \frac{3}{4}\times 0.4444 \\ &= 0.0417 \end{aligned}$$

Với hình 10 (b), nhánh True với số lượng nhãn "Có" và "Không" bằng nhau, dữ liệu trong node này rơi vào tình trạng 2 nhãn phân phối đều như ta đã tìm hiểu (xem lại phần Lưu ý mục II.2.2. và II.3.), khi đó Entropy sẽ bằng 1, và Gini sẽ bằng 0.5, áp dụng công thức cũng sẽ ra kết quả như vậy:

$$H(\text{True}) = -\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2} = 1, \quad G(\text{True}) = 1 - \left(\left(\frac{1}{2}\right)^2 + \left(\frac{1}{2}\right)^2\right) = 0.5.$$

$$IG\big(S_1, \text{Điểm tốt nghiệp} < 23\big) = H(S_1) - \left(\frac{2}{4}\cdot 1 + \frac{2}{4}\cdot 0\right) = 0.8113 - \frac{2}{4} = 0.3113$$

$$GG\big(S_1, \text{Điểm tốt nghiệp} < 23\big) = G(S_1) - \left(\frac{2}{4}\cdot 0.5 + \frac{2}{4}\cdot 0\right) = 0.3750 - 0.25 = 0.1250$$

Với hình 10 (c):

$$IG\Big(S_1, \text{Điểm tốt nghiệp} < 25\Big) = H(S_1) - \Big(\frac{3}{4}\Big(-\frac{2}{3}\log_2\frac{2}{3} - \frac{1}{3}\log_2\frac{1}{3}\Big) + \frac{1}{4}\cdot 0\Big)$$
$$= 0.8113 - \frac{3}{4}\cdot 0.9182$$
$$= 0.1226.$$

$$GG\Big(S_1, \text{Điểm tốt nghiệp} < 25\Big) = G(S_1) - \Big(\frac{3}{4}\Big(1 - (\frac{2}{3})^2 - (\frac{1}{3})^2\Big) + \frac{1}{4}\cdot 0\Big)$$
$$= 0.3750 - \frac{3}{4}\cdot 0.4444$$
$$= 0.0417.$$

So sánh Gain (IG&GG):

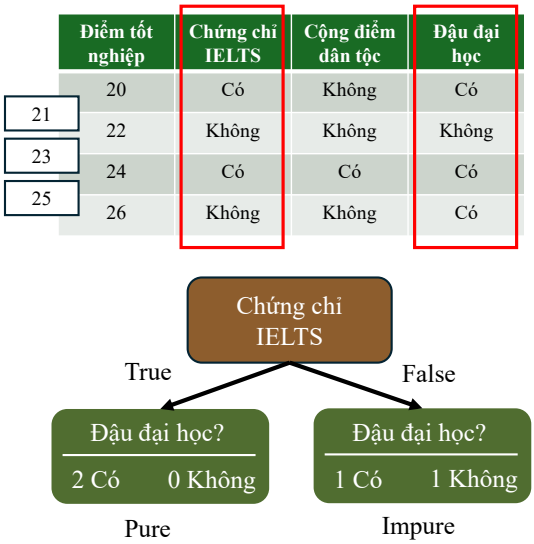
Ngưỡng Điểm	Information Gain	Gini Gain
Điểm < 21.0	0.1226	0.0417
Điểm < 23.0	**0.3113**	**0.1250**
Điểm < 25.0	0.1226	0.0417

IG và GG cao nhất đạt được với ngưỡng "Điểm tốt nghiệp < 23.0".

2) Chứng chỉ IELTS:
Tương tự, ta xét thuộc tính "Chứng chỉ IELTS" trong tập S_1:

$$IG\Big(S_1, \text{IELTS}\Big) = H(S_1) - \Big(\frac{2}{4}\cdot 0 + \frac{2}{4}\cdot 1\Big)$$
$$= 0.8113 - 0.5$$
$$= 0.3113.$$

$$GG\Big(S_1, \text{IELTS}\Big) = G(S_1) - \Big(\frac{2}{4}\cdot 0 + \frac{2}{4}\cdot 0.5\Big)$$
$$= 0.3750 - 0.25$$
$$= 0.1250.$$



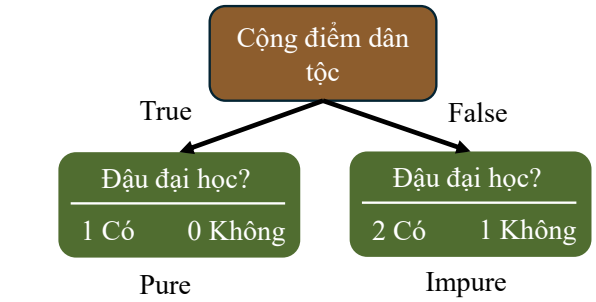
Hình 11: Minh hoạ quá trình tách nhánh dữ liệu với thuộc tính "Chứng chỉ IELTS" của tập S_1.

3) Cộng điểm dân tộc:
Tương tự, ta xét thuộc tính "Cộng điểm dân tộc" trong tập S_1:

$$IG\big(S_1, \text{Cộng điểm dân tộc}\big) = H\big(S_1\big) - \Big(\frac{1}{4}\cdot 0 + \frac{3}{4}\Big(-\frac{2}{3}\log_2\frac{2}{3} - \frac{1}{3}\log_2\frac{1}{3}\Big)\Big) = 0.8113 - 0 - \frac{3}{4}\times 0.9182 = 0.1226.$$

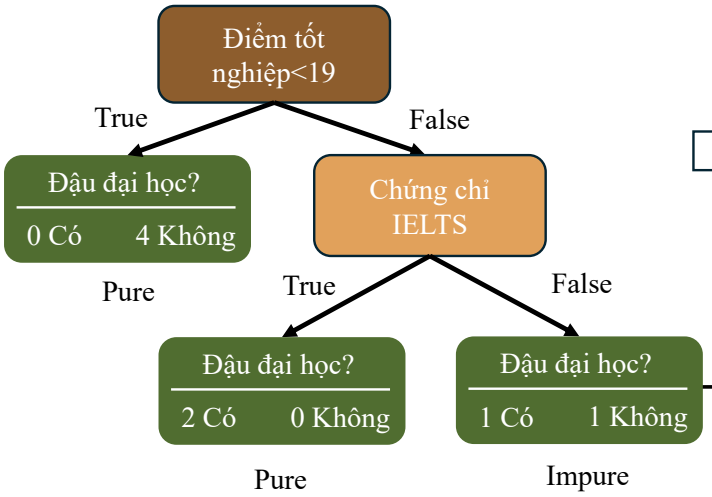
$$GG\big(S_1, \text{Cộng điểm dân tộc}\big) = G\big(S_1\big) - \Big(\frac{1}{4}\cdot 0 + \frac{3}{4}\Big(1 - \Big(\frac{2}{3}\Big)^2 - \Big(\frac{1}{3}\Big)^2\Big)\Big) = 0.3750 - 0 - \frac{3}{4}\times 0.4444 = 0.0417.$$

	Điểm tốt nghiệp	Chứng chỉ IELTS	Cộng điểm dân tộc	Đậu đại học
21	20	Có	Không	Có
23	22	Không	Không	Không
25	24	Có	Có	Có
	26	Không	Không	Có



Hình 12: Minh hoạ quá trình tách nhánh dữ liệu với thuộc tính "Cộng điểm dân tộc" của tập S_1.

Decision Tree cấp 2 (depth=2)
So sánh IG và GG của các thuộc tính tập S_1, ta thấy 2 thuộc tính "Điểm Tốt Nghiệp < 23.0" và "Chứng chỉ IELTS" có giá trị IG và GG lớn nhất và bằng nhau. Tình huống này gọi là tie-break, ta sẽ ưu tiên chọn thuộc tính có dạng dữ liệu categorical "Chứng chỉ IELTS" để đơn giản quá trình.

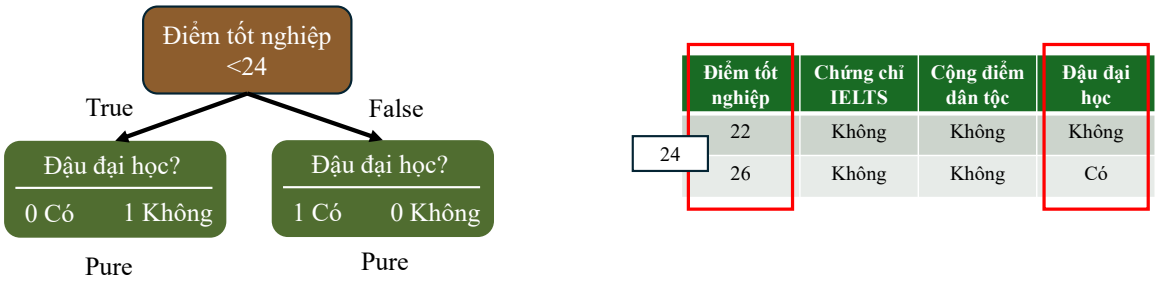


	Điểm tốt nghiệp	Chứng chỉ IELTS	Cộng điểm dân tộc	Đậu đại học
24	22	Không	Không	Không
	26	Không	Không	Có

Hình 13: Minh hoạ quá trình tách nhánh dữ liệu với thuộc tính "Chứng chỉ IELTS" của Decision Tree cấp 2.

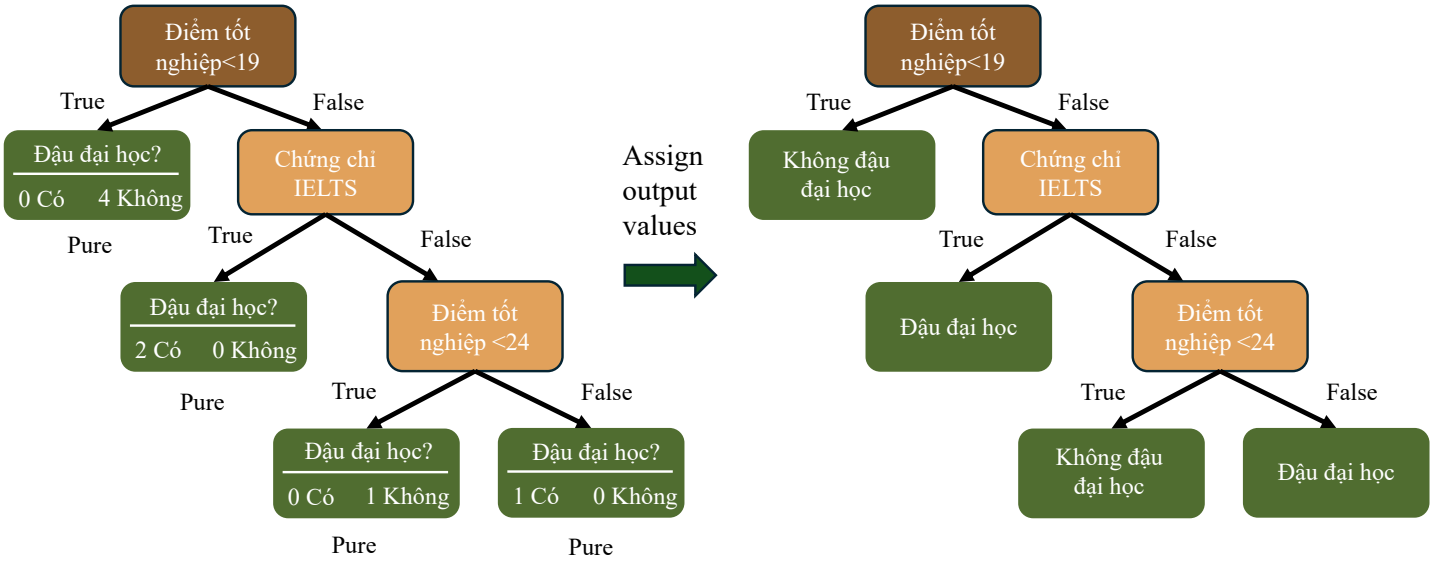
Như vậy ta chỉ còn duy nhất một nhóm cuối bên nhánh False chưa đạt độ thuần, do 2 thí sinh này đều không có chứng chỉ IELTS và không có điểm cộng, ta chỉ xét thuộc tính còn lại là "Điểm

Tốt Nghiệp".



Hình 14: Minh hoạ quá trình tách nhánh dữ liệu với thuộc tính "Điểm tốt nghiệp < 24".

Ta đạt được các node tinh khiết, lúc này không còn gì để phân tách nữa, có thể vẽ được Decision Tree hoàn chỉnh với depth = 3 như sau



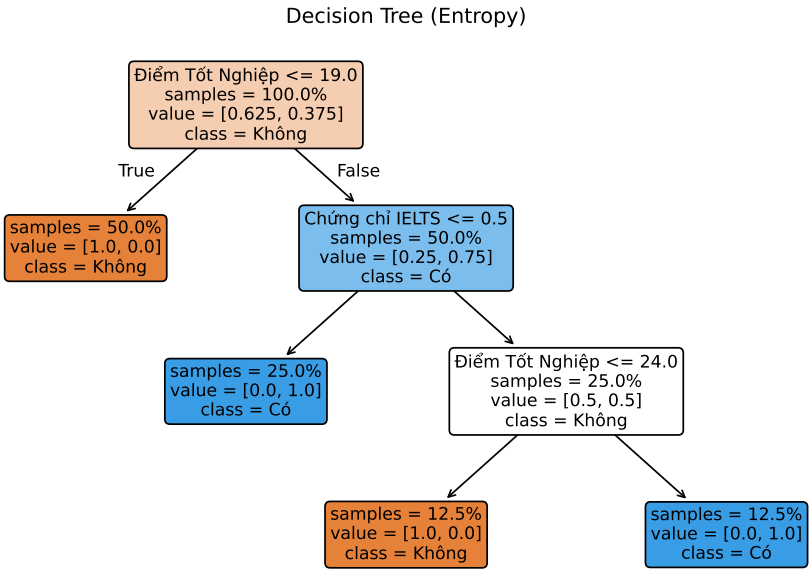
Hình 15: Minh hoạ Decision Tree hoàn chỉnh và gắn giá trị đầu ra cho từng lá.

Trước khi sử dụng Decision Tree để dự đoán cho một mẫu mới, chúng ta phải gán giá trị đầu ra (label) cho mỗi leaf node (nốt lá). Nói một cách đơn giản, ta chỉ cần nhìn vào tập con dữ liệu huấn luyện của node đó, nhãn nào có số lượng nhiều hơn thì gán nhãn đó cho node. Như minh hoạ trong Hình 15, vì tất cả mẫu trong leaf node đều thuần về cùng một nhãn (Pure), ta chỉ việc chọn nhãn đó làm kết quả dự đoán.

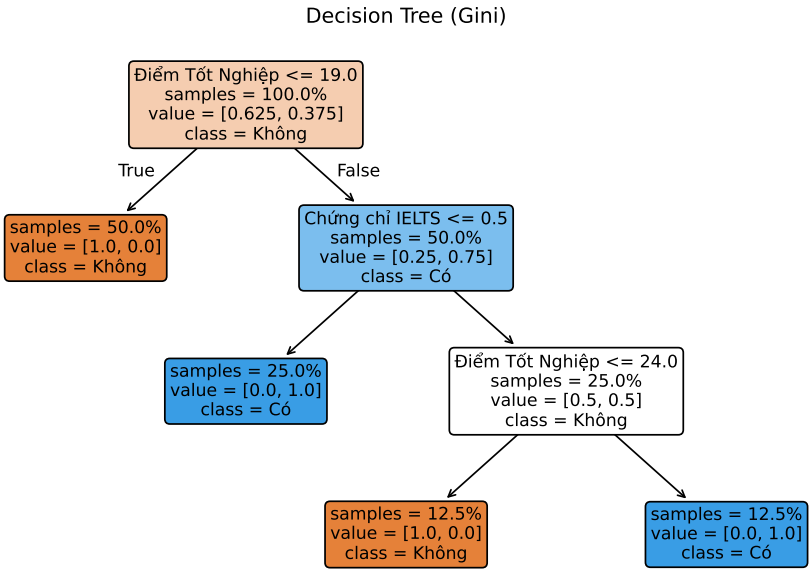
Bước 4. Ứng dụng cây để dự đoán:
Xét thí sinh mới:

$$(\text{Điểm Tốt Nghiệp} = 21.0,\ \text{Chứng chỉ IELTS} = \text{Không},\ \text{Cộng Điểm Dân Tộc} = \text{Có}).$$

$21.0 < 19.0 \rightarrow \text{False}$, IELTS=Không $\rightarrow$ False, $21.0 < 24.0 \rightarrow \text{True} \Rightarrow$ **Không đậu đại học**



Hình 16: Decision Tree xây dựng với Entropy bằng thư viện scikit-learn (biến nhị phân đã được mã hoá 0/1).



Hình 17: Decision Tree xây dựng với Gini Impurity bằng thư viện scikit-learn (biến nhị phân đã được mã hoá 0/1).

IV. Câu hỏi trắc nghiệm

1. Trong cây quyết định, node không còn nhánh con và đưa ra nhãn cuối cùng được gọi là gì?

   (a) Root node.

   (b) Internal node.

   (c) Branch.

   (d) Leaf node.

2. Tiêu chí nào sử dụng hàm $\log_2$ để đo mức "độ ngạc nhiên" trung bình?

   (a) Gini Impurity.

   (b) Variance.

   (c) Entropy.

   (d) Standard Deviation.

3. Chỉ số nào được định nghĩa $G(S) = 1 - \sum_c p_c^2$?

   (a) Information Gain.

   (b) Entropy.

   (c) Mean Squared Error.

   (d) Gini Impurity.

4. Information Gain đo đại lượng nào sau đây?

   (a) Tổng số mẫu ở mỗi node.

   (b) Mức giảm Entropy sau khi tách.

   (c) Khoảng cách Euclid giữa hai điểm.

   (d) Tốc độ hội tụ của thuật toán.

5. Khi xử lý dữ liệu liên tục (Continuous/Numeric data), cây quyết định thường áp dụng thao tác nào?

   (a) Bỏ qua thuộc tính.

   (b) Thêm một nhánh cho mỗi giá trị rời rạc.

   (c) Chọn ngưỡng (threshold) tối ưu.

   (d) Mã hoá one-hot rồi tách nhánh.

6. Thuật ngữ nào chỉ độ sâu tối đa (chiều cao lớn nhất) mà cây quyết định được phép đạt tới?

   (a) minimum leaf size.

   (b) maximum depth.

(c) splitting criterion.

(d) branch factor.

7. Theo các tiên đề của Shannon, hàm thông tin $I(p)$ phải thỏa tính chất nào?

   (a) Tỉ lệ thuận với p.

   (b) Phụ thuộc hệ đơn vị đo.

   (c) Liên tục và có tính cộng dồn.

   (d) Luôn bằng $1/p$.

8. Cho tập dữ liệu:

ID	Tiếp xúc F0	Kết quả xét nghiệm
1	Có	Dương tính
2	Không	Dương tính
3	Không	Âm tính

   Hãy tính Root Entropy $H(S)$ của cột nhãn (đơn vị bit).

   (a) 0.6667.

   (b) 0.8113.

   (c) 0.9183.

   (d) 1.0000.

9. Vẫn dựa trên tập dữ liệu ở câu 8, tính Information Gain khi phân tách theo thuộc tính Tiếp xúc F0.

   (a) 0.0000.

   (b) 0.2516.

   (c) 0.3333.

   (d) 0.9183.

10. Với cùng tập dữ liệu ở Câu 8, hãy tính Gini Impurity $G(S)$ của cột nhãn tại nút gốc.

    (a) 0.2222.

    (b) 0.3333.

    (c) 0.4444.

    (d) 0.6667.

V. Tài liệu tham khảo

[1] C. E. Shannon, "A mathematical theory of communication", *Bell System Technical Journal*, vol. 27, no. 3–4, pp. 379–423, 623–656, 1948. DOI: 10.1002/j.1538-7305.1948.tb01338.x.

[2] MIT OpenCourseWare, *15.097 Prediction: Machine Learning and Statistics (Spring 2012) Lecture 08: Decision Trees*, Accessed: 2025-08-06, 2012.

[3] L. Breiman et al., *Classification and Regression Trees*. Monterey, CA: Wadsworth & Brooks/Cole, 1984.

[4] J. R. Quinlan, "Induction of decision trees", *Machine Learning*, vol. 1, no. 1, pp. 81–106, 1986. DOI: 10.1023/A:1022643204877.

[5] J. R. Quinlan, *C4.5: Programs for Machine Learning*. San Mateo, CA: Morgan Kaufmann, 1993.

Phụ lục

1. **Datasets:** Các file dataset được đề cập trong bài có thể được tải tại đây.
2. **Hint:** Các file code gợi ý có thể được tải tại đây.
3. **Solution:** Các file code cài đặt hoàn chỉnh và phần trả lời nội dung trắc nghiệm có thể được tải tại đây.
4. **Rubric:**

Mục	Kiến Thức	Đánh Giá
I.	- Khái niệm Decision Tree. - Các thuật ngữ cơ bản.	- Giải thích chính xác khái niệm Decision Tree. - Liệt kê và định nghĩa đúng các thuật ngữ cơ bản (root, internal node, leaf, branch).
II.	- Thuật toán xây dựng Classification Tree. - Công thức Entropy và Gini Impurity. - Định nghĩa Information Gain và Gini Gain. - Các bước giải một bài toán Classification với Decision Tree.	- Mô tả đầy đủ các bước xây dựng tree. - Hiểu các thành phần công thức tính Entropy và Gini Impurity. - Hiểu sự khác biệt giữa Entropy và Gini Impurity.
III.	- Thực hành bài toán.	- Triển khai pipeline step by step. - Ứng dụng Decision Tree.