AI VIET NAM – Course 2024

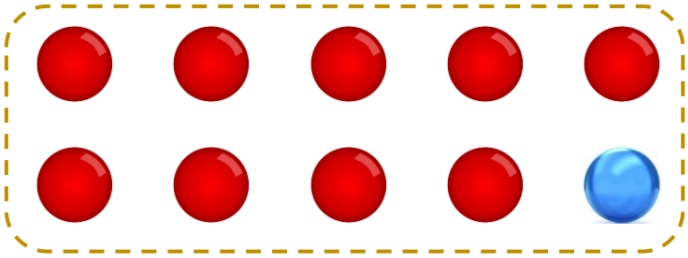
# Decision Tree

Xuân Hiệp

Ngày 15 tháng 2 năm 2024

## 1. Cây quyết định cho bài toán Phân Loại:

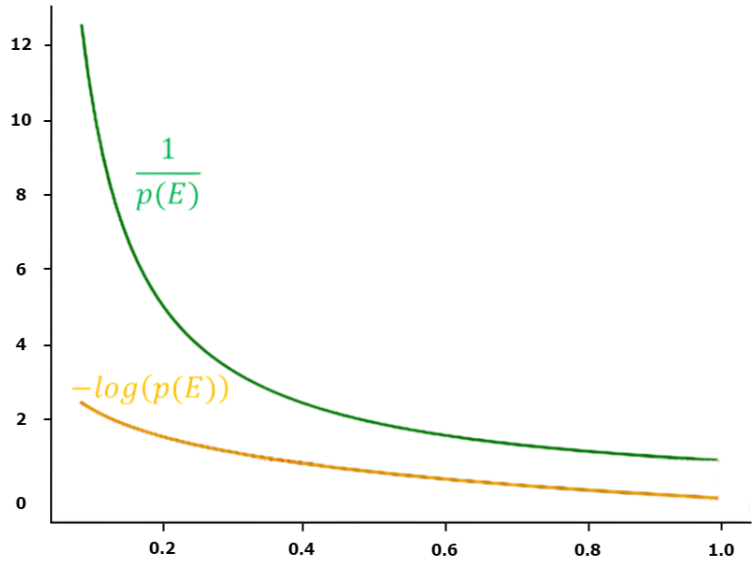
### 1.1 Entropy:



Hình 1: Ví dụ lấy một quả bóng bất kỳ từ một tập hợp

- Để nắm bắt được cách vận dụng **Cây quyết định** cho bài toán **phân loại** ta cần tìm hiểu về Enrtropy, một khái niệm mới thường được nhắc đến mỗi khi nhắc đến **Cây quyết định**.
- Đầu tiên, ta cùng xem qua 1 ví dụ đơn giản:
  - Theo hình 1, ta có một tập hợp gồm 10 quả bóng, trong đó gồm 9 quả đỏ và 1 quả xanh. Ta gọi E là phép kiểm thử lấy 1 quả bóng bất kỳ từ tập hợp. Ta đưa ra 2 phép kiểm thử:
  - A: Lấy 1 quả bóng màu đỏ
  - B: Lấy 1 quả bóng màu xanh
  - Từ đó, xác suất của A và B tương ứng là P(A)=0.9, P(B)=0.1
  - Theo kết quả trên, khi ta lấy 1 quả bóng từ tập hợp, khả năng lấy được quả bóng đỏ sẽ rất cao và lấy bóng xanh sẽ rất thấp. Tuy nhiên, dù rất nhỏ, vẫn sẽ có trường hợp ta lấy được bóng xanh. Từ đó, ta có khái niệm Surprise, cho ta thấy mức độ ngạc nhiên với 1 phép kiểm thử và có giá trị tỷ lệ nghịch với xác xuất. Công thức tính được như sau:

$$S(E) = \frac{1}{P(E)}$$




Hình 2: Khoảng giá trị của Surpise

– Từ hình 2, ta thấy được giá trị S $\in$ [1, $\infty$], từ đó xác xuất càng lớn thì S càng nhỏ và khi P(E)=1 thì S(E)=1. Ta thấy trường hợp này rõ ràng rất vô lý với khái niệm Surprise, dó đó, ta cần chuẩn hóa lại với công thức sau (S $\in$ [0, $\infty$]):

$$S(E) = log(\frac{1}{P(E)}) = -log(P(E))$$


Hình 3: Ví dụ tung 1 đồng xu

– Tiếp theo, ta cùng xem qua 1 ví dụ khác:
– Theo hình 3, ta tung 1 đồng xu và quy đổi sự kiện tung đồng xu về dạng số (Head - 1 và Tail - 0).
– Ta có: P(H)=P(T)=0.5. Vậy sau n lần tung, tổng giá trị xác suất là:

$$P(H) * n * 1 + P(T) * n * 1$$

– Tổng quát hơn, đối với các thực nghiệm thực tế, ta gọi $X_i \in \{0,1\}$ ứng với các trường

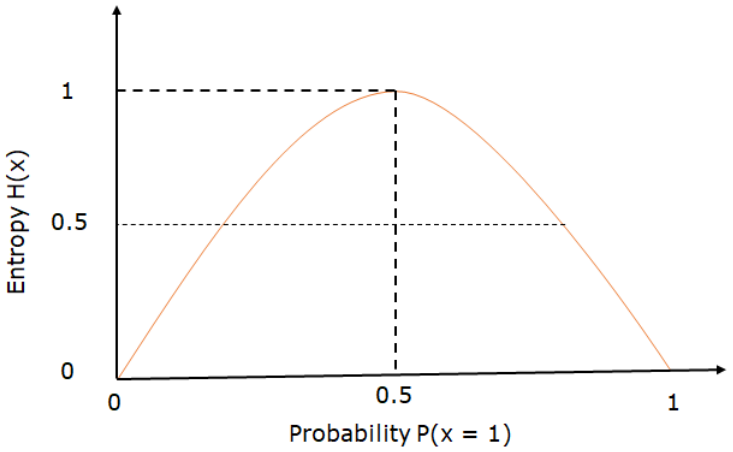
hợp thực nghiệm. Khi đó tổng giá trị xác suất:

$$\sum_i X_i P(X_i) = X_1 * P(X_1) + X_2 * P(X_2)$$

- Ta có S(E)= -log(P(E)). Do đó, ta giá trị trung bình Surprise là:

$$\overline{E}(S) = -\sum P(X) * log(P(X))$$

- Từ đó ta gọi giá trị trung bình Surprise, (hay còn gọi là Entropy) là công thức đo mức độ không tinh khiết (Impurity) của dữ liệu.




Hình 4: Khoảng giá trị của Entropy

- Theo hình 4, ta thấy với 1 tập hợp gồm 2 loại (ví dụ bóng đỏ, bóng xanh) thì giá trị Entropy cao nhất khi P(đỏ)=P(xanh).

### 1.2 Information Gain:

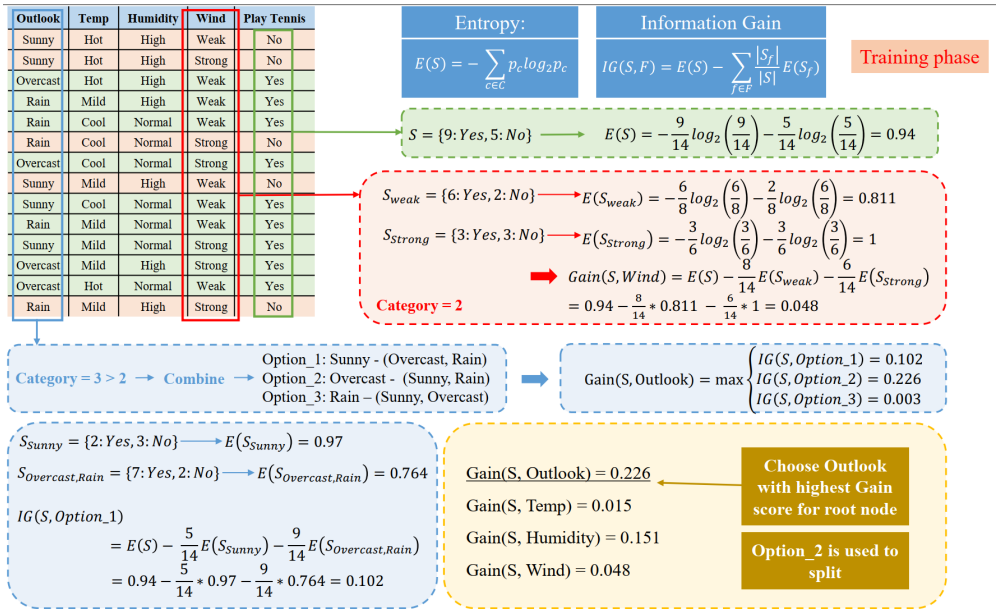
- Dùng đo lường việc giảm độ không tinh khiết và cũng là yếu tố quyết định thuộc tính nào nên được chọn làm **nút gốc** cho **Cây quyết định**. Ta có công thức sau:

$$IG(S, F) = E(S) - \sum_{f \in F} \frac{|S_f|}{|S|} * E(S_f)$$

Trong đó:

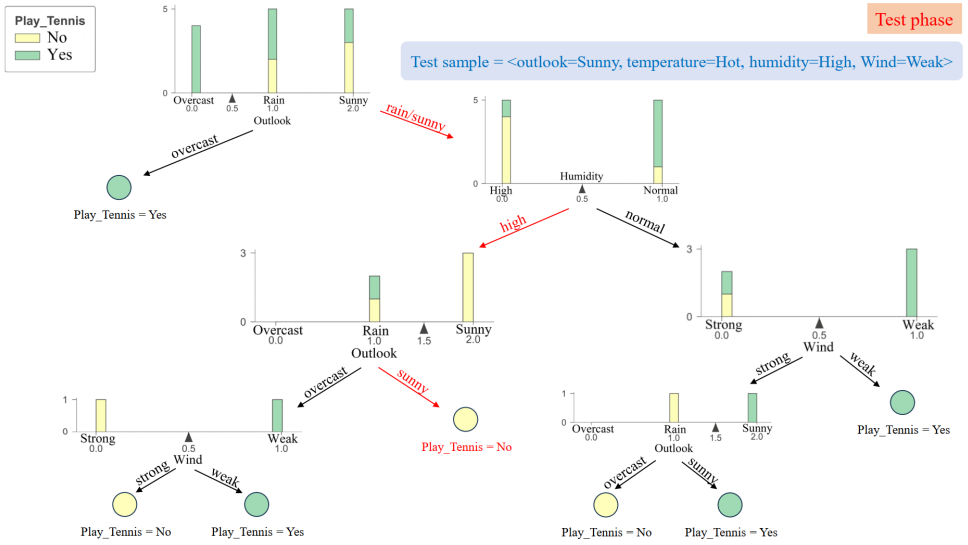
- E(S) là entropy tổng với S là tập hợp gốc ban đầu
- $S_f$ là các tập hợp con tách từ S với số lượng dữ liệu khác nhau. Từ đó, ta tính entropy cho các tập con $E(S_f)$ với các trọng số $\frac{|S_f|}{|S|}$ tương ứng
- Khi **Infomation Gain** tăng thì Entropy sẽ giảm. Đây là tiêu chí để quyết định việc chọn đặc trưng nào để phân nhánh. Ta cần tính **Infomation Gain** của tất cả các nhánh có thể xảy ra và chọn nhánh có **Infomation Gain** lớn nhất. Quá trình này được lặp lại nhiều lần cho đến khi các nhánh chỉ chứa 1 loại dữ liệu duy nhất.

### 1.3 Cây quyết định với dữ liệu rời rạc:




Hình 5: Tính Information Gain với dữ liệu rời rạc

- Đầu tiên, ta tính entropy của tập hợp gốc (cột Play Tennis). Trong bảng dữ liệu, ngoài cột Play Tennis, ta có 4 cột, mỗi cột lại bao gồm các giá trị thông tin khác nhau. Để xây dựng **Cây quyết định**, ta phải tính **Information Gain** cho tất cả các trường hợp có thể có từ giá trị thông tin của các cột.
- Trong ví dụ trên, ta chọn cột Wind (Weak, Strong), cột Outlook (Sunny, Overcast, Rain). Để đơn giản bài toán, ta tách nhánh theo dạng cây nhị phân với từng giá trị của mỗi cột. Ta tính tương tự với các cột khác, từ đó, ta có **Information Gain**=0.226 lớn nhất (Ứng với của cột Outlook và Option 2)




Hình 6: Cây quyết định với dữ liệu rời rạc

- Sau khi xây dựng xong cây quyết định, ta có thể thử với 1 sample đơn giản:

$$< outlook = Sunny, temperature = Hot, humidity = High, Wind = Weak >$$

- Sau khi kiểm tra, kết quả trả về là Play Tennis=No (Không chơi Tennis)

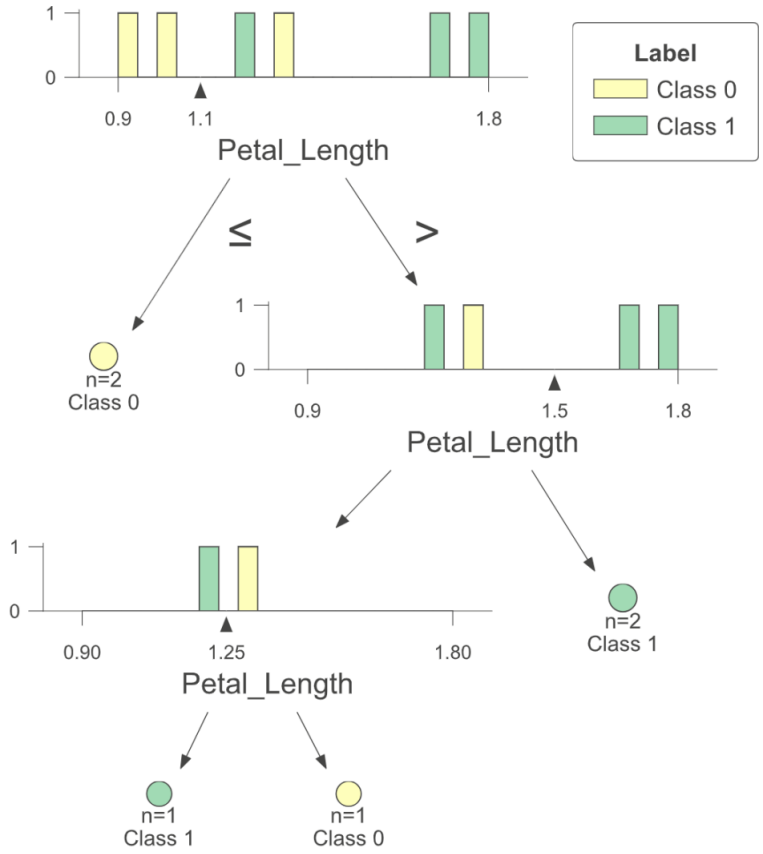
### 1.4 Cây quyết định với dữ liệu liên tục:

| Petal_Length | Label |
|---|---|
| 1 | 0 |
| 1.3 | 0 |
| 0.9 | 0 |
| 1.7 | 1 |
| 1.8 | 1 |
| 1.2 | 1 |

| Petal_Length | Label | Mean | Entropy | Information Gain |
|---|---|---|---|---|
| 0.9 | 0 | 0.95 | 0.8091 | 0.1909 |
| 1.0 | 0 | 1.1 | 0.5409 | 0.4591 |
| 1.2 | 1 | 1.25 | 0.9183 | 0.0817 |
| 1.3 | 0 | 1.5 | 0.5409 | 0.4591 |
| 1.7 | 1 | 1.75 | 0.8091 | 0.1909 |
| 1.8 | 1 | Total Entropy = 1 | | |

Hình 7: Dữ liệu liên tục

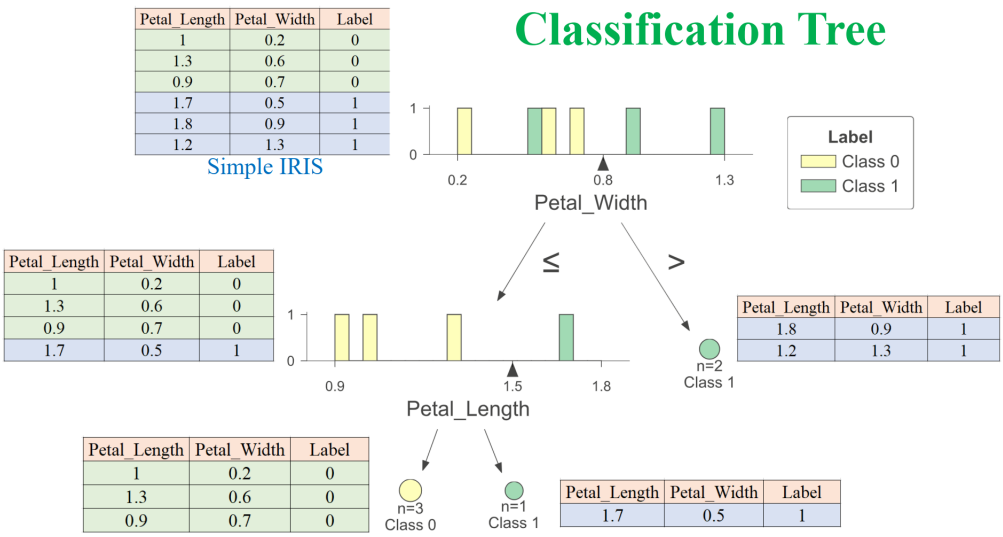
- Để xây dựng **Cây quyết định**, ta tạo điều kiện phân nhánh, ta lấy giá trị trung bình của 2 giá trị Petal Length kế tiếp nhau trong bảng. Sau đó, ta tính **Information Gain** theo từng điều kiện phân nhánh đã tạo.




Hình 8: Cây quyết định với dữ liệu liên tục

- Theo hình 7, với Mean $\in \{1.1, 1.5\}$ ta được Gain cao nhất. Do đó, ta chọn Mean=1.1 làm điều kiện phân nhánh. Làm tương tự với các bước tiếp theo, ta được cây quyết định như hình 8. Ta thử ví dụ với Petal Length=1.5, theo **Cây quyết định** sẽ cho ra Class 0

### 1.5 Cây quyết định với bảng dữ liệu có nhiều cột

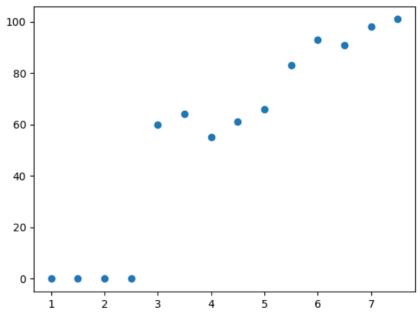



Hình 9: Cây quyết định với dữ liệu nhiều cột

- Tương tự với dữ liệu 1 cột, ta làm tuần tự các bước, lấy Mean cho tất cả các cột, tính **Information Gain** theo tất cả các giá trị Mean, lấy Gain lớn nhất và xây dựng Cây quyết định như hình 9

## 2. Decision Tree cho bài toán Hồi Quy:

| Experience | Salary |
|---|---|
| 1 | 0 |
| 1.5 | 0 |
| 2 | 0 |
| 2.5 | 0 |
| 3 | 60 |
| 3.5 | 64 |
| 4 | 55 |
| 4.5 | 61 |
| 5 | 66 |
| 5.5 | 83 |
| 6 | 93 |
| 6.5 | 91 |
| 7 | 98 |
| 7.5 | 101 |



**When Experience = 5.3,**

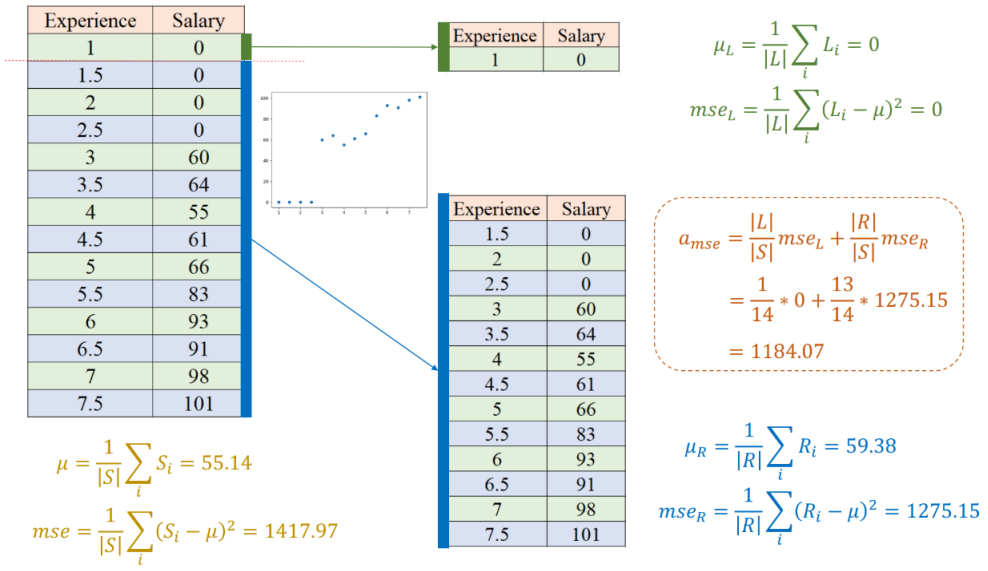
**Salary = ?**

Hình 10: Bảng dữ liệu cho bài toán Hồi quy

- Với bài toán này, ta mong muốn nhóm các Experience gần nhau lại, để tìm giá trị Mean tốt nhất mô tả được các giá trị Experience. Đầu tiên, ta tính Mean của các giá trị Experience. Sau đó, ta áp dụng công thức Mean Squared Error để thấy được sự chênh lệch dữ liệu với giá trị Mean. Độ chênh lệch càng thấp thì các giá trị Experience càng gần Mean, khi đó, giá trị Salary dự đoán được sẽ càng chính xác.

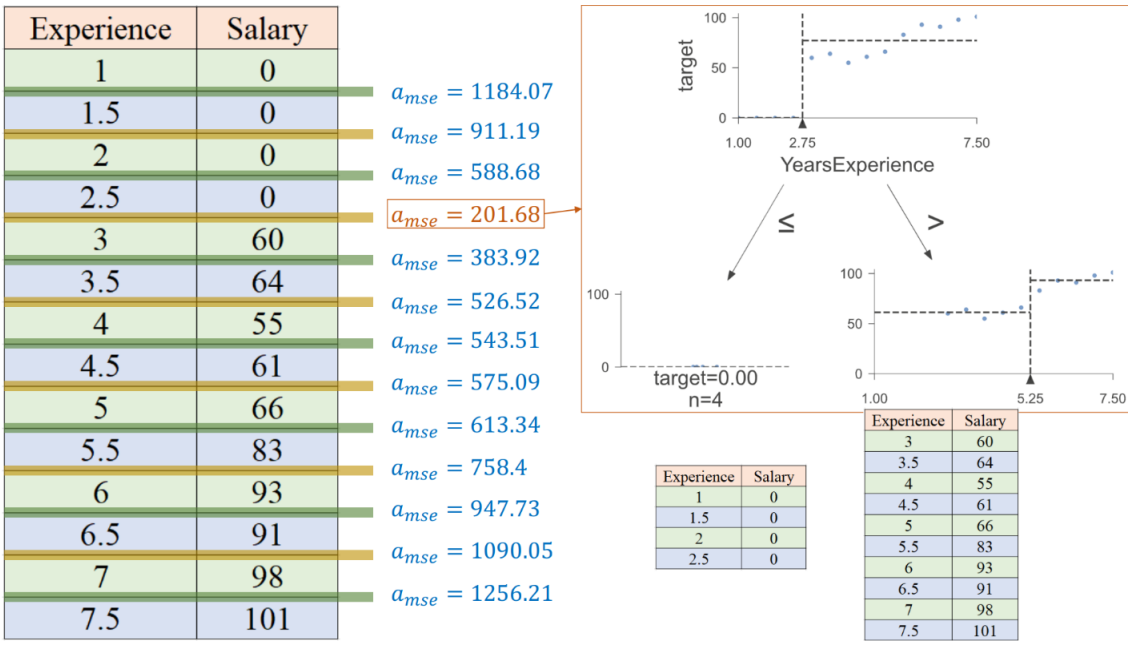
- Công thức Mean Squared Error (L là tổng số giá trị và $L_i$ là giá trị thứ i):

$$mse_L = \frac{1}{|L|} * \sum_i (L_i - Mean)^2$$




Hình 11: Xây dựng cây quyết định cho bài toán Hồi quy

- Để chọn điều kiện phân nhánh tốt nhất, tách bảng dữ liệu theo các giá trị lẻ, sau đó tính giá trị Mean tổng theo dữ liệu.




Hình 12: Cây quyết định cho bài toán Hồi quy

- Sau khi tìm được điều kiện với Mean tốt nhất, ta xây dựng cây quyết định theo điều kiện phân nhánh hình 11

**- Hết -**