**1. Kiến trúc của cây quyết định**

Cây quyết định bao gồm các thành phần chính:

* **Nút gốc (Root Node)**: Là nút đầu tiên của cây, đại diện cho toàn bộ dữ liệu và thường chứa toàn bộ dữ liệu huấn luyện.
* **Nút bên trong (Internal Nodes)**: Các nút này đại diện cho một câu hỏi hay điều kiện (condition) dựa trên giá trị của một đặc trưng nào đó.
* **Cành (Branch)**: Là đường nối giữa các nút, đại diện cho kết quả của điều kiện phân chia.
* **Lá (Leaf Nodes)**: Đây là các nút cuối cùng của cây, đưa ra quyết định phân loại hoặc giá trị dự đoán trong bài toán hồi quy.

**2. Quá trình xây dựng cây**

Cây quyết định xây dựng dựa trên việc liên tục chia nhỏ dữ liệu, sao cho các nhóm con (sau mỗi lần chia) có độ thuần nhất cao hơn (ít hỗn hợp giữa các lớp khác nhau). Mục tiêu của cây là tìm ra cách phân chia tối ưu giúp tăng độ chính xác của mô hình.

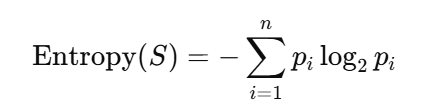
**Các bước xây dựng cây:**

1. **Chọn đặc trưng để phân chia**: Tại mỗi nút, cây quyết định tìm ra đặc trưng nào có độ thuần nhất cao nhất khi sử dụng để phân chia.
2. **Tính toán độ đo để chọn đặc trưng phân chia**:
   * Với **CART (Classification and Regression Trees)**: sử dụng **Gini Index**.
   * Với **ID3 (Iterative Dichotomiser 3)**: sử dụng **Information Gain** dựa trên Entropy.
3. **Phân chia dữ liệu**: Chia dữ liệu thành hai nhóm (hoặc nhiều nhóm) dựa trên điều kiện phân chia.
4. **Đệ quy xây dựng các nhánh con**: Lặp lại quá trình với các nhánh mới sinh ra cho đến khi đạt được điều kiện dừng.

**3. Các tiêu chí lựa chọn đặc trưng phân chia**

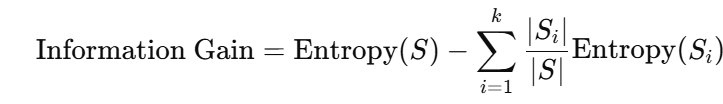
**3.1 Entropy và Information Gain (ID3)**

* **Entropy**: Đo độ hỗn loạn hoặc không chắc chắn của dữ liệu. Độ entropy cao khi dữ liệu gồm nhiều lớp hỗn hợp, và thấp khi dữ liệu có độ thuần nhất cao.



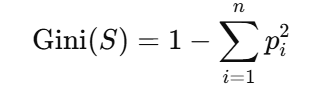
Trong đó pi là tỷ lệ của lớp i trong tập dữ liệu S.

* **Information Gain**: Đo lường mức độ giảm entropy sau khi chia dữ liệu theo một đặc trưng.



Trong đó Si là nhóm con sau khi chia dữ liệu dựa trên đặc trưng đó.

**3.2 Gini Index (CART)**

* **Gini Index**: Đo độ thuần khiết của một tập dữ liệu. Giá trị của Gini sẽ thấp hơn khi các mẫu trong tập dữ liệu thuộc về một lớp duy nhất và cao hơn khi các lớp khác nhau trộn lẫn với nhau. CART cố gắng giảm thiểu chỉ số Gini để tối ưu phân chia dữ liệu.

**4. Điều kiện dừng**

Cây quyết định sẽ dừng phân chia ở một nhánh khi đạt một trong các điều kiện sau:

* **Độ sâu tối đa**: Độ sâu của cây đạt ngưỡng cho trước.
* **Kích thước tối thiểu của nhóm con**: Số lượng mẫu trong nhóm con đạt ngưỡng cho trước.
* **Không có cải thiện trong phép đo chất lượng**: Khi không còn cải thiện về độ thuần nhất.

**5. Ưu và nhược điểm của cây quyết định**

**Ưu điểm:**

* **Dễ hiểu và trực quan**: Cây quyết định có thể trực quan hóa và giải thích một cách dễ dàng.
* **Ít yêu cầu tiền xử lý**: Không yêu cầu chuẩn hóa hoặc chuẩn hóa dữ liệu.
* **Có thể xử lý dữ liệu phân loại và dữ liệu liên tục**: Cây quyết định có thể chia dữ liệu liên tục dựa trên ngưỡng và dữ liệu phân loại dựa trên từng lớp cụ thể.

**Nhược điểm:**

* **Dễ bị overfitting**: Cây quyết định có xu hướng phức tạp khi dữ liệu huấn luyện nhiều nhiễu, dẫn đến quá khớp.
* **Nhạy cảm với dữ liệu huấn luyện**: Các thay đổi nhỏ trong dữ liệu huấn luyện có thể dẫn đến thay đổi lớn trong cấu trúc cây.
* **Khả năng tổng quát kém**: Với dữ liệu phức tạp, cây quyết định đơn lẻ có thể không hiệu quả bằng các mô hình khác (thường dùng cây ngẫu nhiên hoặc rừng ngẫu nhiên).