**Lý thuyết về CART và ID3**

1. **CART (Classification and Regression Trees)**
   * **Đặc điểm:** CART là thuật toán cây quyết định sử dụng chỉ số Gini để đo độ thuần nhất của các nút.
   * **Gini Index:** Gini Index là một thước đo độ bất bình đẳng hoặc độ "tạp" của dữ liệu ở mỗi nút của cây. Gini Index cho biết xác suất mà một điểm dữ liệu được chọn ngẫu nhiên sẽ bị phân loại sai nếu nó được gán nhãn theo phân phối hiện tại. Giá trị Gini càng thấp thì dữ liệu càng thuần nhất.
   * **Ứng dụng:** CART có thể áp dụng cho cả bài toán phân loại (Classification Trees) và hồi quy (Regression Trees).
   * **Quy trình:** CART sẽ chọn đặc trưng phân chia dữ liệu bằng cách tối thiểu hóa chỉ số Gini tại mỗi nút.
2. **ID3 (Iterative Dichotomiser 3)**
   * **Đặc điểm:** ID3 là thuật toán cây quyết định sử dụng Information Gain (Lợi ích Thông tin) để chọn thuộc tính chia dữ liệu.
   * **Information Gain:** Lợi ích Thông tin là sự giảm bớt Entropy, đo lường mức độ hỗn loạn của dữ liệu, khi tách tập dữ liệu theo một thuộc tính. Entropy càng cao thì dữ liệu càng hỗn loạn, và quá trình chọn thuộc tính dựa trên việc giảm Entropy tối đa.
   * **Ứng dụng:** ID3 chủ yếu được sử dụng cho các bài toán phân loại.
   * **Quy trình:** ID3 chọn thuộc tính có Lợi ích Thông tin cao nhất để phân chia dữ liệu tại mỗi nút.

So sánh giữa 2 phương pháp :   


**Lựa chọn và ứng dụng**

* CART phù hợp cho các bài toán có dữ liệu lớn, có thể chứa cả biến định lượng và định tính, và cho phép linh hoạt chọn nhị phân.
* ID3 thích hợp cho dữ liệu phân loại và đơn giản hơn, tuy nhiên có thể kém hiệu quả với các thuộc tính có nhiều giá trị.

Thực hành :

https://github.com/HaVanQuy-depzai/baitapnop/commit/e24e9c5eac440ca26257256ac369fc836a7d0067