Đinh Huy Tú – 20211722  
  
Thuật toán **Cây Quyết Định (Decision Tree)** là một phương pháp học máy phân loại và hồi quy dựa trên việc chia dữ liệu thành các nút (node) và nhánh (branch) theo các thuộc tính đầu vào. Các thuật toán cây quyết định chính bao gồm:  
 **Thuật toán Cây Quyết Định (Decision Tree)**:

* **CART (Classification and Regression Trees)**: Thuật toán cây quyết định này dùng tiêu chí **Gini** để phân chia dữ liệu tại các nút. Mục tiêu là tạo ra các nút phân chia có độ tinh khiết cao nhất (tức là các nhóm dữ liệu thuộc cùng một lớp càng nhiều càng tốt).
* **ID3 (Iterative Dichotomiser 3)**: Đây cũng là một thuật toán cây quyết định, nhưng nó sử dụng **Entropy và Information Gain** để xác định các điểm phân chia dữ liệu. Entropy đo lường mức độ hỗn loạn, và Information Gain chọn thuộc tính có khả năng phân tách dữ liệu tốt nhất.

 **Đặc trưng HOG (Histogram of Oriented Gradients)**:

* Trước khi áp dụng cây quyết định cho phân lớp ảnh, đoạn mã sử dụng **HOG (Histogram of Oriented Gradients)** để trích xuất đặc trưng từ ảnh. HOG là một phương pháp phổ biến để trích xuất đặc trưng của hình ảnh, giúp mô tả các cạnh và hình dạng trong ảnh. Điều này biến dữ liệu ảnh thành các vector đặc trưng số, giúp cây quyết định xử lý dễ dàng hơn.

 **Phân lớp trên dữ liệu IRIS và ảnh nha khoa**:

* Đầu tiên, thuật toán cây quyết định được áp dụng trên dữ liệu **IRIS** – một bộ dữ liệu phổ biến trong học máy chứa thông tin về ba loài hoa iris khác nhau.
* Sau đó, thuật toán này được áp dụng cho **ảnh nha khoa**, sử dụng các đặc trưng HOG của ảnh.

### 2. Cách Hoạt Động

#### Bước 1: Tạo Cây Quyết Định

Thuật toán cây quyết định được xây dựng thông qua các bước sau:

1. **Chọn Đặc Trưng Tốt Nhất để Chia**:
   * Để tạo ra các nút trong cây, thuật toán sẽ chọn đặc trưng mà sẽ phân chia dữ liệu tốt nhất.
   * **Tiêu chí chọn lựa**:
     + **Chỉ số Gini (Gini Index)**: Được tính toán dựa trên xác suất của một nhãn được chọn. Chỉ số Gini sẽ có giá trị từ 0 (tốt nhất) đến 1 (xấu nhất).
     + **Độ đo Thông tin (Entropy)**: Được tính toán để xác định mức độ không chắc chắn của một hệ thống. Giá trị của Entropy nằm trong khoảng từ 0 đến 1, với 0 đại diện cho sự thuần khiết hoàn toàn và 1 cho sự hỗn loạn hoàn toàn.
     + **Thông tin tăng (Information Gain)**: Được tính toán bằng cách lấy độ đo Entropy của toàn bộ dữ liệu trừ đi Entropy của dữ liệu sau khi chia theo đặc trưng.
2. **Chia Dữ Liệu**:
   * Sau khi chọn đặc trưng tốt nhất, dữ liệu sẽ được chia thành các nhánh dựa trên giá trị của đặc trưng đó. Quá trình này sẽ tiếp tục cho đến khi một trong các điều kiện dừng được thỏa mãn (ví dụ: khi một nút chứa chỉ một lớp hoặc khi chiều sâu của cây đạt đến một giới hạn).

#### Bước 2: Dự Đoán

Khi cây quyết định đã được xây dựng, việc dự đoán một nhãn cho một mẫu mới sẽ được thực hiện bằng cách đi từ nút gốc xuống nút lá theo các quy tắc đã được xác định trong cây.

### 3. Đặc Điểm Nổi Bật

* **Trực quan hóa**: Cây quyết định có thể được trực quan hóa dễ dàng, giúp cho việc phân tích và hiểu rõ hơn về quá trình ra quyết định.
* **Khả năng xử lý dữ liệu phi tuyến tính**: Cây quyết định có thể xử lý dữ liệu phi tuyến tính mà không cần phải chuyển đổi thành dạng tuyến tính như một số mô hình khác.
* **Chạy nhanh**: Thời gian huấn luyện và dự đoán thường nhanh, đặc biệt với các tập dữ liệu lớn.

### 4. Ưu và Nhược Điểm

#### Ưu điểm:

* Dễ hiểu và dễ giải thích.
* Không cần phải chuẩn hóa dữ liệu (như chuẩn hóa hay chuẩn hóa Z).
* Có thể xử lý cả dữ liệu số và dữ liệu phân loại.
* Có khả năng quản lý dữ liệu bị thiếu.

#### Nhược điểm:

* Cây quyết định có thể dễ dàng bị overfitting (quá khớp) với dữ liệu huấn luyện nếu không được điều chỉnh đúng cách (chẳng hạn như cắt tỉa cây).
* Nhạy cảm với thay đổi nhỏ trong dữ liệu, có thể tạo ra một cây hoàn toàn khác nếu dữ liệu bị thay đổi một chút.
* Có thể không hoạt động tốt trong các trường hợp dữ liệu phân loại có mối quan hệ phức tạp.

### 5. Kết luận

Cây quyết định là một trong những thuật toán phổ biến trong học máy do tính trực quan và dễ áp dụng. Chúng có thể được sử dụng cho nhiều bài toán khác nhau, từ phân loại đến hồi quy, và có thể hoạt động tốt trong nhiều tình huống, cây quyết định được sử dụng để phân loại dữ liệu hoa Iris và ảnh nha khoa, cho thấy tính linh hoạt và khả năng áp dụng của thuật toán này.