**I. Decision Tree**

**1. Khái niệm**

- Cây quyết định (Decision Tree) là một cây phân cấp có cấu trúc được dùng để phân lớp các đối tượng dựa vào dãy các luật (series of rules). Khi cho dữ liệu về các đối tượng gồm các thuộc tính cùng với lớp (classes) của nó, cây quyết định sẽ sinh ra các luật để dự đoán lớp của các đối tượng chưa biết (unseen data).

- Thuật toán cây quyết định thuộc họ thuật toán học có giám sát. Không giống như các thuật toán học có giám sát khác, thuật toán cây quyết định cũng có thể được sử dụng để giải các bài toán hồi quy và phân loại

- Các loại cây quyết định dựa trên mục tiêu mà chúng ta có, nó có thể có 2 loại:

+ Cây quyết định biến phân loại: Cây quyết định có biến mục tiêu phân loại nên được gọi là cây quyết định biến phân loại

+ Cây quyết định biến liên tục: Cây quyết định có biến mục tiêu liên tục nên được gọi là cây quyết định biến liên tục

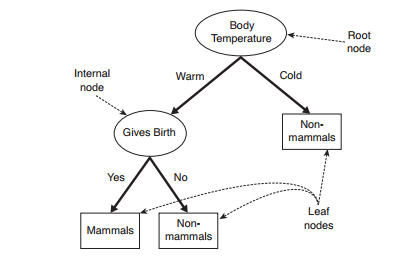
**2. Cấu trúc Decision Tree**

- Decision Tree gồm 3 phần chính: 1 node gốc (root node), những node lá (leaf nodes) và các nhánh của nó (branches):

+ Node gốc là điểm bắt đầu của cây quyết định, nó đại diện cho toàn bộ quần thể hoặc mẫu và phần này được chia thành 2 hay nhiều tập đông nhất

+ Node lá là các node không phân chia, và được gọi là node lá hoặc node đấu cuối

+ Nhánh: một phần con của toàn bộ cây được gọi là nhánh hoặc cây con



- Một số tiêu chí có thể được lựa chọn trong Decision Tree:

+ Entropy  
+ Information gain  
+ Gini index  
+ Gain Ratio  
+ Reduction in Variance  
+ Chi-Square

**3. Các thuật toán liên quan tới Decition Tree**

- ID3 (Iterative Dichotomiser 3): Xây dụng cây quyết định bằng cách sử dụng phương pháp tìm kiếm tham lam từ trên xuống thông qua không gian của các nhánh có thể mà không cần quay lại. Có thể tóm tắt như sau:

+ Lấy tất cả các thuộc tính chưa được sử dụng và đếm entropy liên quan đến mẫu thử của các thuộc tính đó

+ Chọn thuộc tính có entropy lớn nhất

+ Nối node với thuộc tính đó

- C4.5: là thuật toán phân lớp dữ liệu dựa trên cây quyết định hiệu quả và phổ biến trong những ứng dụng khai phá cơ sở dữ liệu có kích thước nhỏ. C4.5 xây dựng cây quyết định từ tập training data tương tự như ID3:

+ Tập training data S gồm các mẫu đã được phân  loại sẵn s1, s2,… Mỗi mẫu si = x1,x2,… với x1, x2 là 1 vector biểu diễn cho thuộc tính hoặc đặc điểm của mẫu. Một tập vector C = c1,c2,… với c1, c2 biểu diễn cho mỗi lớp mà mỗi mẫu thuộc về.

+ Với mỗi lớp của cây, C4.5 chọn một thuộc tính của dữ liệu mà phân chia tập các mẫu thành nhiều tập con, được nâng cao chất lượng một cách hiệu quả nhất. Tiêu chuẩn của nó là thu được **information gain** (sự khác biệt về entropy) – kết quả từ việc chọn một thuộc tính cho việc chia tách dữ liệu. Quyết định đưa ra dựa trên tập thuộc tính với information gain được chuẩn hóa cao nhất. Thuật toán C4.5 sau đó sẽ lặp lại với các danh sách con nhỏ hơn.

- Ngoài ID3, C4.5 thì còn một số thuật toán khác như:

+ Thuật toán CHAID: tạo cây quyết định bằng cách sử dụng thống kê chi-square để xác định các phân tách tối ưu. Các biến mục tiêu đầu vào có thể là số (liên tục) hoặc phân loại.

+ Thuật toán C&R: sử dụng phân vùng đệ quy để chia cây. Tham biến mục tiêu có thể dạng số hoặc phân loại.

+ MARS

+ Conditional Inference Trees

**4. Ưu, nhược điểm của Decision Tree**

**4.1. Ưu điểm**

- Cây quyết định là một thuật toán đơn giản và phổ biến. Thuật toán này được sử dụng rộng rãi bới những lợi ích của nó:

+ Mô hình sinh ra các quy tắc dễ hiểu cho người đọc, tạo ra bộ luật với mỗi nhánh lá là một luật của cây.

+ Dữ liệu đầu vào có thể là là dữ liệu missing, không cần chuẩn hóa hoặc tạo biến giả

+ Có thể làm việc với cả dữ liệu số và dữ liệu phân loại

+ Có thể xác thực mô hình bằng cách sử dụng các kiểm tra thống kê

+ Có khả năng là việc với dữ liệu lớn

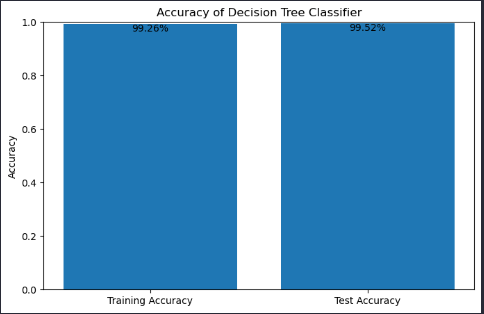
**4.2. Nhược điểm**

- Kèm với đó, cây quyết định cũng có những nhược điểm cụ thể:

+ Mô hình cây quyết định phụ thuộc rất lớn vào dữ liệu của bạn. Thạm chí, với một sự thay đổi nhỏ trong bộ dữ liệu, cấu trúc mô hình cây quyết định có thể thay đổi hoàn toàn.

+ Cây quyết định hay gặp vấn đề [overfitting](https://trituenhantao.io/tu-dien-thuat-ngu/overfitting/)

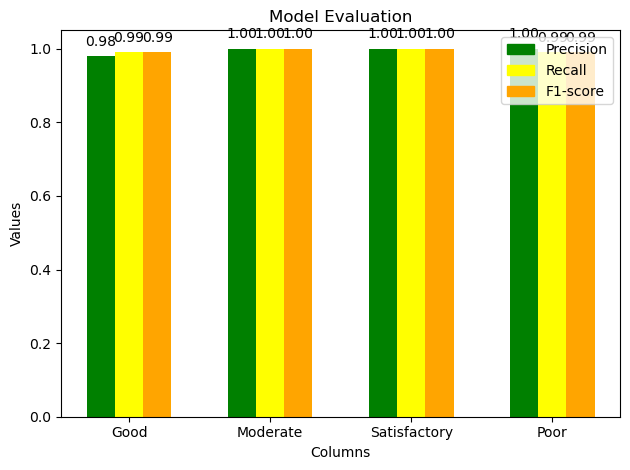
**II. Đánh giá mô hình**



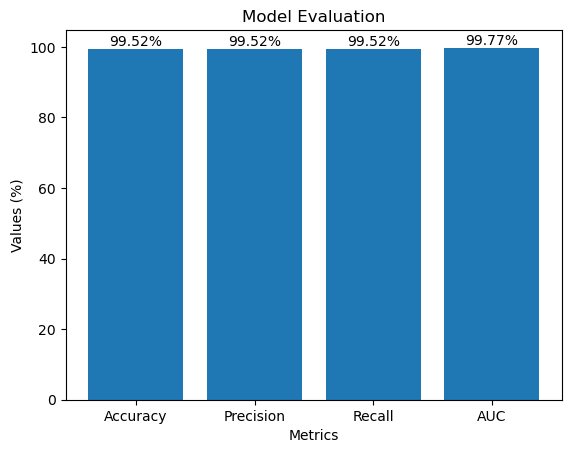
*So sánh về kết quả training và test*

**

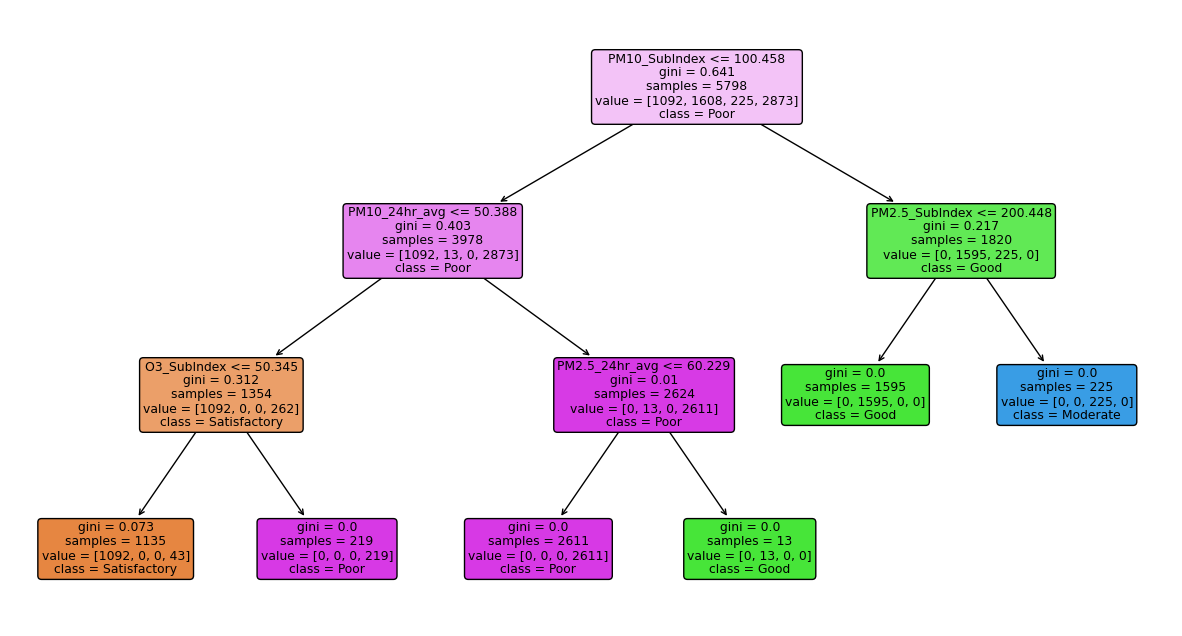
*Ma trận nhầm lẫn*

**

*Bảng so sánh dựa trên classsification\_report*

******

*Bảng đánh giá qua một vài chỉ số*

**

*Cây quyết định của mô hình*