Ta sẽ dùng một ví dụ để giải thích về cây quyết định:

David là quản lý của một câu lạc bộ đánh golf nổi tiếng. Anh ta đang có rắc rối chuyện các thành viên đến hay không đến. Có ngày ai cũng muốn chơi golf nhưng số nhân viên câu lạc bộ lại không đủ phục vụ. Có hôm, không hiểu vì lý do gì mà chẳng ai đến chơi, và câu lạc bộ lại thừa nhân viên.

Mục tiêu của David là tối ưu hóa số nhân viên phục vụ mỗi ngày bằng cách dựa theo thông tin dự báo thời tiết để đoán xem khi nào người ta sẽ đến chơi golf. Để thực hiện điều đó, anh cần hiểu được tại sao khách hàng quyết định chơi và tìm hiểu xem có cách giải thích nào cho việc đó hay không.

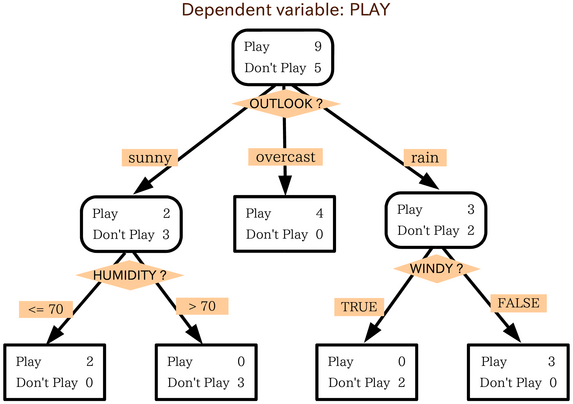
Vậy là trong hai tuần, anh ta thu thập thông tin về:

Trời (*outlook*) (nắng (*sunny*),

Và tất nhiên là số người đến chơi golf vào hôm đó. David thu được một bộ dữ liệu gồm 14 dòng và 5 cột.

[](https://vi.wikipedia.org/wiki/T%E1%BA%ADp_tin:Golf_dataset.png)

Sau đó, để giải quyết bài toán của David, người ta đã đưa ra một mô hình cây quyết định.

[](https://vi.wikipedia.org/wiki/T%E1%BA%ADp_tin:Decision_tree_model.png)

Nhóm người chơi golf khi trời nắng, nhóm chơi khi trời nhiều mây, và nhóm chơi khi trời mưa.

Kết luận thứ nhất: nếu trời nhiều mây, người ta luôn luôn chơi golf. Và có một số người ham mê đến mức chơi golf cả khi trời mưa.

Tiếp theo, ta lại chia nhóm trời nắng thành hai nhóm con. Ta thấy rằng khách hàng không muốn chơi golf nếu độ ẩm lên quá 70%.

Cuối cùng, ta chia nhóm trời mưa thành hai và thấy rằng khách hàng sẽ không chơi golf nếu trời nhiều gió.

Và đây là lời giải ngắn gọn cho bài toán mô tả bởi cây phân loại. David cho phần lớn nhân viên nghỉ vào những ngày trời nắng và ẩm, hoặc những ngày mưa gió. Vì hầu như sẽ chẳng có ai chơi golf trong những ngày đó. Vào những hôm khác, khi nhiều người sẽ đến chơi golf, anh ta có thể thuê thêm nhân viên thời vụ để phụ giúp công việc.

Kết luận là cây quyết định giúp ta biến một biểu diễn dữ liệu phức tạp thành một cấu trúc đơn giản hơn rất nhiều.

**Gini impurity**

Dùng trong thuật toán CART (Classification and Regression Trees). Nó dựa vào việc bình phương các xác suất thành viên cho mỗi thể loại đích trong nút. Giá trị của nó tiến đến cực tiểu (bằng 0) khi mọi trường hợp trong nút rơi vào một thể loại đích duy nhất.

Giả sử *y* nhận các giá trị trong {1, 2,..., *m*} và gọi *f(i,j)* là tần suất của giá trị *j* trong nút *i*. Nghĩa là *f(i,j)* là tỷ lệ các bản ghi với *y*=*j* được xếp vào nhóm *i*.