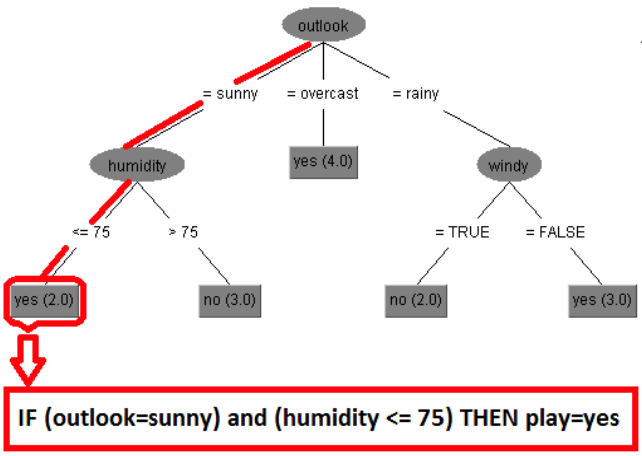
Quá trình tìm k láng giềng của $\boldsymbol{x}$ thường sử dụng khoảng cách (distance) hay độ tương tự (similarity).

3.3 Cây quyết định (Decision Trees - DT)

Cây quyết định đề xuất bởi (Breiman *et al.*, 1984; Quinlan, 1993) là mô hình máy học tự động sử dụng rất nhiều trong khai mỏ dữ liệu (Wu and Kumar, 2009) do tính đơn giản và hiệu quả. Hình 3 minh họa một ví dụ của cây quyết định thu được bằng cách học từ tập dữ liệu, để dự đoán chơi Golf (ŷ = yes / no?) từ các biến (thời tiết, nhiệt độ, độ ẩm, gió). Mô hình rất dễ hiểu bởi vì chúng ta có thể rút trích luật quyết định tương ứng với nút lá có dạng IF-THEN được tạo ra từ việc thực hiện AND trên các điều kiện theo đường dẫn từ nút gốc đến nút lá. Các luật quyết định dễ hiểu với người sử dụng.



outlook	temp.	hum.	windy	Play
sunny	85	85	false	No
sunny	80	90	true	No
overcast	83	78	false	Yes
rain	70	96	false	Yes
rain	68	80	false	Yes
rain	65	70	true	No
. . .	. . .	. . .	. . .	. . .
. . .	. . .	. . .	. . .	. . .

Hình 3: Cây quyết định học từ dữ liệu cho phép dự báo chơi Golf

Xét tập dữ liệu bao gồm $\boldsymbol{m}$ phần tử $\boldsymbol{x_1, x_2, \ldots, x_m}$ trong không gian $\boldsymbol{n}$ chiều, có giá trị tương ứng của biến phụ thuộc là $\boldsymbol{y_1, y_2, \ldots, y_m}$. Giải thuật học từ dữ liệu là quá trình xây dựng cây bắt đầu từ nút gốc đến nút lá. Đây là giải thuật đệ quy phân hoạch tập dữ liệu theo các biến độc lập thành các phân vùng chữ nhật rời nhau mà ở đó các phần tử dữ liệu $\boldsymbol{x_i}$, $\boldsymbol{x_j}$, …, $\boldsymbol{x_k}$ của cùng phân vùng (nút lá) có các $\boldsymbol{y_i}$, $\boldsymbol{y_j}$, …, $\boldsymbol{y_k}$ là thuần khiết:

- Giống nhau trong vấn đề phân lớp,
- Tương tự nhau trong vấn đề hồi quy.

Giải thuật học mô hình cây quyết định từ dữ liệu gồm 2 bước lớn: xây dựng cây, cắt nhánh để tránh học vẹt. Quá trình xây dựng cây được làm như sau:

- Bắt đầu từ nút gốc, tất cả các dữ liệu học ở nút gốc,
- Nếu các phần tử dữ liệu tại 1 nút là thuần khiết thì nút đang xét được cho là nút lá, giá trị dự báo của nút lá cho vấn đề phân lớp với bình chọn số đông trong các giá trị $\{\boldsymbol{y_i, \ldots, y_k}\}$, cho vấn đề hồi quy với giá trị trung bình của các $\{\boldsymbol{y_i, \ldots, y_k}\}$.
- Nếu dữ liệu ở nút quá hỗn loạn (các giá trị $\{\boldsymbol{y_i, \ldots, y_k}\}$ rất khác nhau) thì nút được cho là nút trong, tiến hành phân hoạch dữ liệu một cách đệ quy bằng việc chọn 1 biến để thực hiện phân hoạch tốt nhất có thể.

Một biến được cho là tốt được sử dụng để phân hoạch dữ liệu sao cho kết quả thu được cây nhỏ nhất. Việc lựa chọn này dựa vào các heuristics: chọn biến sinh ra các nút thuần khiết nhất. Hiện nay có 2 giải thuật học cây quyết định tiêu biểu là C4.5 của (Quinlan, 1993), CART của (Breiman *et al.*, 1984).

Để đánh giá và chọn biến khi phân hoạch dữ liệu, Quinlan đề nghị sử dụng độ lợi thông tin (chọn biến có độ lợi thông tin lớn nhất) và tỉ số độ lợi dựa trên hàm entropy của Shannon. Độ lợi thông tin của một biến được tính bằng: độ đo hỗn loạn trước khi phân hoạch trừ cho sau khi phân hoạch. Giả sử $\boldsymbol{p_c}$ là xác suất mà phần tử trong dữ liệu $\boldsymbol{D}$ thuộc lớp $\boldsymbol{y_c}$ (c =1 , C), độ đo hỗn loạn thông tin trước khi phân hoạch được tính theo công thức entropy (4) như sau:

$$Info(D) = -\sum_{c=1}^{C} p_c \log_2 p_c \qquad (4)$$