**Machine learning**

ML hay máy học là một nhánh của trí tuệ nhân tạo là lĩnh vực cho phép máy tính có khả năng tự cãi thiện bản thân dựa trên dữ liệu mẫu( training dât) hoặc dựa vào kinh nghiệm ( những gì đã được học). ML có thể tự dự đoán hoặc đưa ra quyết định mà không cần được lập trình cụ thể

2 loại cụ thể là dự đoán (prediction) và phân loại( classification).

Dự đoán giá nhà , xe. Phân loại nhận diện chữ viết tay, đồ vật….

**Quy trình làm việc:**

1. Data collection- thu thập dữ liệu: để máy tính tự học cần có một bộ dữ liệu(dataset), có thể tự thu thập chúng hoặc lấy từ những bộ dữ liệu đã được công bố trước(phải từ các nguồn chính thống, có như vậy mới chính xác và máy có thể tự học và đạt hiệu quả cao)
2. Preprocessing- tiền xử lý: bước này dùng để chuẩn hóa dữ liệu, loại bỏ các thuộc tính không cần thiết, gán nhãn dữ liệu, mã hóa, trích xuất và rút gọn một số đặc trưng của dữ liệu nhưng vẫn đảm bảo được kết quả.( bước này tốn thời gian nhất tỉ lệ với lượng dữ liệu. Bước 1 và 2 thường chiếm 70% tổng thời gian thực hiện.
3. Training model- huấn luyện mô hình: bước này là bước huấn luyện cho mô hình hay chính là cho nó học trên dữ liệu mình đã thu thập và xử lý 2 bước đầu.
4. Evaluating model\_ đánh giá mô hình: sau khi xong bước 3 chúng ta cần dùng các độ đo để đánh giá mô hình, tùy vào độ do khác nhau mà mô hình cũng được đánh giá tốt hay không khác nhau. Độ chính xác của mô hình đạt trên 80% được cho là tốt.
5. Improve- cải thiện: Sau khi đánh giá mô hình đạt độ chính xác không tốt cần được train lại, sẽ lặp lại 3 bước, cho đến khi đạt được độ chính xác như kì vọng. Tổng thời gian của 3 bước cuối rơi vào 30% tông thời gian thực hiện.

**Phân loại**: có rất nhiều cách phân loại ml thông thường được phân loại thành 2 loai chính

* Supervised learning : học có giám sát
* Việc cho máy tính học trên dữ liệu đã được gán nhãn, hay nói cách khác, với mỗi đầu vào Xi, chúng ta sẽ có nhẫn Yi tương ứng.
* Unsupervised learing: học không giam sát
* Việc cho máy tính học trên những dữ liệu mà không được gán nhãn, các thuật toán ML sẽ tự tìm ra tương quan dữ liệu, mô hình hóa dữ liệu hay là làm cho máy tính có kiến thức,, hiểu về dữ liệu, từ đó chúng ta có thể phân loại các dữ liệu thành các nhóm, lớp giống nhau mà chúng ta đã được học hoặc giảm số chiều dữ liệu.

**Các khái niệm cơ bản**

* Data set là tập dữ liệu ở dạng nguyên thủy chưa qua xử lý đã được thu thập ở bước datacollection. Một dataset bao gồm nhiều data point
* Data point là điểm dữ liệu, mỗi điểm dữ liệu biểu diễn cho một quan sát có nhiều đặ trưng hay thuộc tính khác nhau được chia làm 2 loại

+ Số

+ Chuỗi

Được biểu diễn thành dòng tương ứng, mỗi dòng có thể có 1 hoặc nhiều dữ liệu ( chính là đặc trưng)

* Training data và test data: dataset thường được chia làm 2 tập này, training data dùng để huấn luyện cho mô hình, test data dùng để dự đoán và đánh giá mô hình. Có bài toán sẽ cho sẵn hai tập này thì không cần phải chia nữa, đối với bài toán chỉ cho mỗi dataset thì phải chia ra. Thường tỷ lệ giữa tập train và test là 8/2
* Featurest vector là vector đặc trưng , mỗi vector này biểu diễn cho một điểm dữ liệu trong dataset. Mỗi vector có n chiều biểu diễn đặc trưng data, mỗi đặc trưng là một chiều và phải là dữ liệu số. các mô hình chỉ có thể huấn luyện được từ các vector đặc trưng này, do đó data set cần phải chuyển về dạng một tập các vector đặc trưng.
* Model là các mô hình dung để training trên một dât theo thuật toán mô hình đó . sau đó dùng để dự đoán hoặc đưa ra các quyết định nhưng gì nó đã được học.

**Ứng dụng**

ML sử dụng trong mọi lĩnh vực trong đời sống:

* Tài chính – ngân hàng
* Sinh học
* Nông nghiệp
* Tìm kiếm, trích xuất thông tin
* Tự động hóa
* Robotics
* Hóa học
* Mạng máy tính
* Khoa học vũ trụ
* Quảng cáo
* Xử lý ngôn ngữ tự nhiên
* Thị giác máy tính

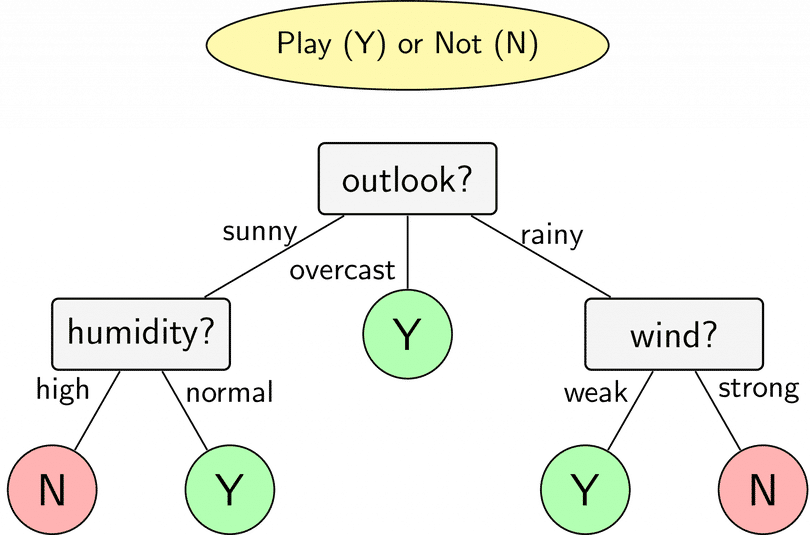
Decision tree (<https://www.youtube.com/watch?v=88rhJ3ow3Us>), <https://en.wikipedia.org/wiki/Decision_tree#Decision-tree_elements> 12/23/2023

<https://trituenhantao.io/kien-thuc/decision-tree/> 12/23/2023

Khái niệm Cây quyết định là 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 dữ liệu chưa biết..

Ví dụ:

Những đặc điểm ban đầu là:Thời tiết Độ ẩm Gió



Nếu trời nắng, độ ẩm bình thường thì khả năng các bạn nam đi chơi bóng sẽ cao. Còn nếu trời nắng, độ ẩm cao thì khả năng các bạn nam sẽ không đi chơi bóng.

Có 2 loại cây chính là

Cây phân loại: khi phân tích kết quả dự đoán thuộc về phân lớp dữ liệu. - Cây hồi quy: khi phân tích, kết quả dự đoán có thể được coi là một số thực (ví dụ: giá của một ngôi nhà, hoặc thời gian nằm viện của bệnh nhân).

**Ưu điểm** Cây quyết định là một thuật toán đơn giản và phổ biến. 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

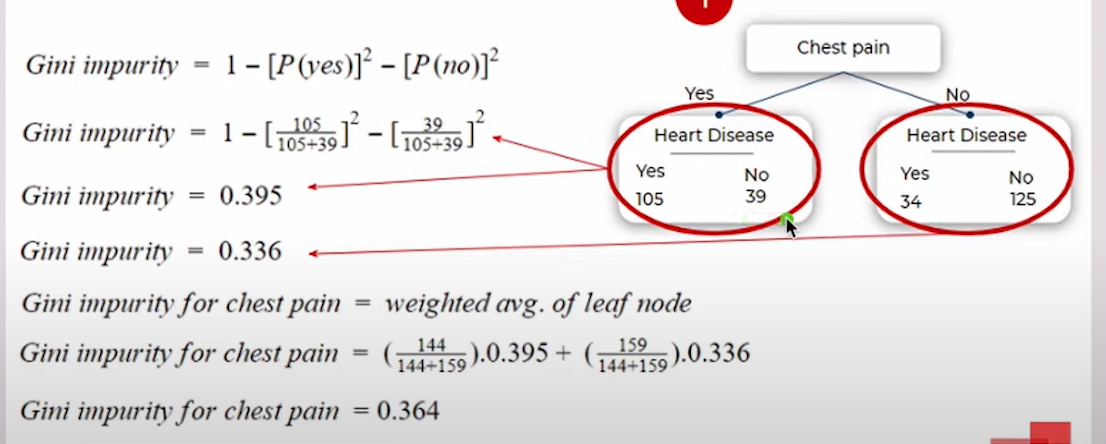
+Có thể làm việc với cả dữ liệu số và dữ liệu phân loại

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

+Có khả năng là việc với dữ liệu lớn

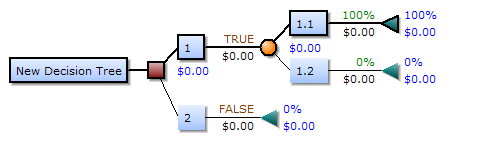
Nhược điểm mô hình cây quyết định phụ thuộc rất lớn vào dữ liệu của bạn. Thạm chí, với một sự thay đổi nhỏ trong bộ dữ liệu, cấu trúc mô hình cây quyết định có thể thay đổi hoàn toàn.

Cây quyết định hay gặp vấn đề overfitting

* Chạy hết root ( tính được cái root tốt nhất)
* 
* Chọn gini nhỏ nhật làm node root trên cùng
* Gini impurity= 1-[p(yes)^2]- [p(no)^2]

+ số( hồi quy)

**Các phần tử của cây quyết định**



Vẽ từ trái sang phải,DS chỉ có các nút cụm nhưng không có nút chìm.Nếu vẽ thủ công chúng thường phát triễn rất lớn và khó vẽ bằng tay.( Theo truyền thống, cây quyết định được tạo thủ công – như ví dụ bên cạnh cho thấy – mặc dù ngày càng có nhiều phần mềm chuyên dụng được sử dụng.)

### Quy tắc quyết định tuyến tính hóa trong đó kết quả là nội dung của nút lá và các điều kiện dọc theo đường dẫn tạo thành một kết hợp trong mệnh đề if.

### *if* condition1 *and* condition2 *and* condition3 *then* outcome