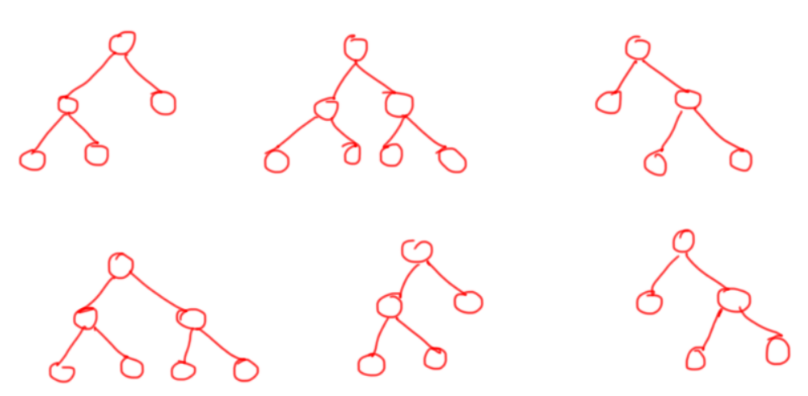
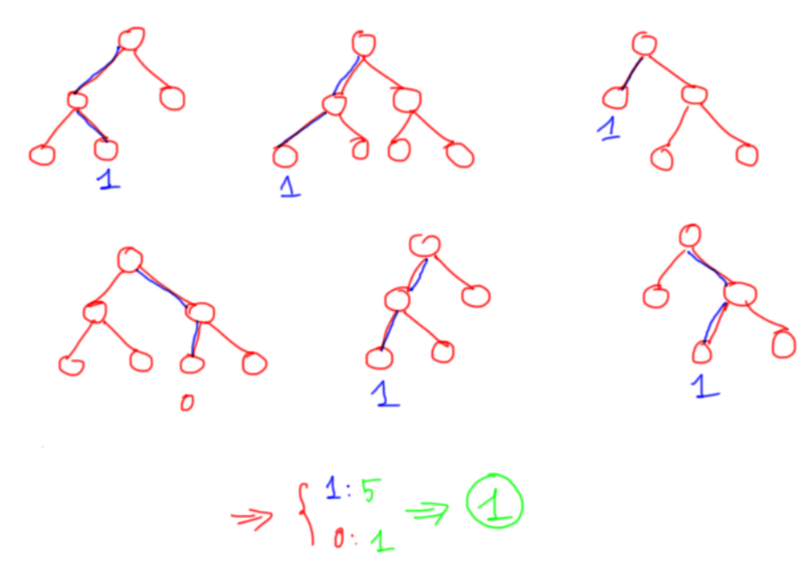
1. Khái niệm

* Random Forest(Rừng ngẫu nhiên): là một thuật toán bao gồm nhiều cây quyết định, mỗi cây được xây dựng dùng thuật toán Decision Tree trên tập dữ liệu khác nhau và dùng tập thuộc tính khác nhau. Sau đó kết quả dự đoán của thuật toán Random Forest sẽ được tổng hợp từ các cây quyết định để đưa ra kết quả chuẩn xác nhất.
* Decision trees(Cây quyết định): bắt đầu bằng một câu hỏi cơ bản, chẳng hạn như "Tôi có nên đi chơi không?". Từ đó, bạn có thể đưa ra một loạt câu hỏi khác để xác định câu trả lời, chẳng hạn như “Đường có tắc không?” hay “Trời có mưa không?”. Mỗi câu hỏi giúp một cá nhân đi đến quyết định cuối cùng, quyết định này sẽ được biểu thị bằng nút lá.
* Do quá trình xây dựng mỗi cây quyết định đều có yếu tố ngẫu nhiên (random) nên kết quả là các cây quyết định trong thuật toán Random Forest có thể khác nhau.

Ví dụ : xây dựng nhiều cây quyết định, các cây quyết định khác nhau



Sau đó ở bước dự đoán, với một dữ liệu mới, thì ở mỗi cây quyết định sẽ đi từ trên xuống theo các node điều kiện để được các dự đoán.

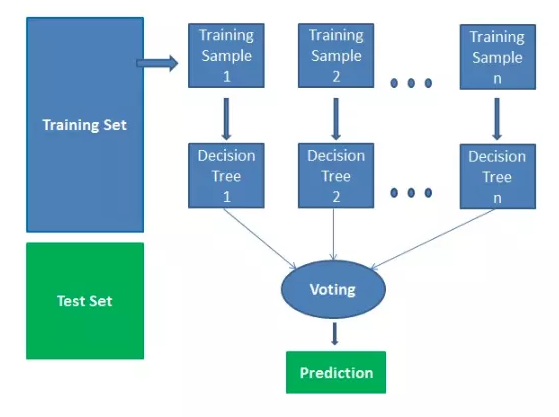


Với ví dụ như trên, thuật toán Random Forest có 6 cây quyết định, 5 cây dự đoán 1 và 1 cây dự đoán 0, do đó sẽ cho ra dự đoán cuối cùng là 1.

## Thuật toán Random Forest

* Thuật toán hoạt động theo bốn bước:

1. Chọn các mẫu ngẫu nhiên từ tập dữ liệu đã cho.
2. Thiết lập cây quyết định cho từng mẫu và nhận kết quả dự đoán từ mỗi quyết định cây.
3. Hãy bỏ phiếu cho mỗi kết quả dự đoán.
4. Chọn kết quả được dự đoán nhiều nhất là dự đoán cuối cùng.



Các thuật toán Random Forest(Rừng ngẫu nhiên) có ba tham số chính cần được thiết lập trước khi huấn luyện, bao gồm kích thước nút, số lượng cây và số lượng tính năng được lấy mẫu.

1. Ưu điểm và nhược điểm
2. Ưu điểm

* Do mỗi cây quyết định trong thuật toán Random Forest không dùng tất cả dữ liệu training, cũng như không dùng tất cả các thuộc tính của dữ liệu để xây dựng cây nên mỗi cây có thể sẽ dự đoán không tốt, độ chính xác thấp. Tuy nhiên, kết quả cuối cùng của thuật toán Random Forest lại tổng hợp từ nhiều cây quyết định, thế nên thông tin từ các cây sẽ bổ sung thông tin cho nhau, dẫn đến mô hình có độ chính xác cao và phương sai thấp, hay mô hình có kết quả dự đoán tốt.

1. Nhược điểm

* Quy trình tốn thời gian: Vì các thuật toán Random Forest có thể xử lý các tập dữ liệu lớn nên chúng có thể đưa ra các dự đoán chính xác hơn nhưng có thể xử lý dữ liệu chậm do chúng đang tính toán dữ liệu cho từng cây quyết định riêng lẻ.
* Yêu cầu nhiều tài nguyên hơn: Vì các khu rừng ngẫu nhiên xử lý các tập dữ liệu lớn hơn nên chúng sẽ yêu cầu nhiều tài nguyên hơn để lưu trữ dữ liệu đó.
* Phức tạp hơn: Dự đoán của một cây quyết định đơn lẻ dễ diễn giải hơn khi so sánh với một rừng cây quyết định.

1. Ứng dụng

* Thuật toán Random Forest đã được áp dụng trong một số ngành, cho phép họ đưa ra các quyết định kinh doanh tốt hơn.
* Một số trường hợp khác sử dụng:

+ Chăm sóc sức khỏe: Các ứng dụng trong sinh học điện toán, cho phép các bác sĩ giải quyết các vấn đề như phân loại biểu hiện gen, khám phá sinh học. Có thể đưa ra ước tính xung quanh phản ứng của thuốc đối với các loại thuốc cụ thể.

+ Thương mại điện tử: Được sử dụng cho các công cụ đề xuất cho mục đích buôn bán có lợi cho cả hai bên.

+ Tài chính: Được sử dụng để đánh giá những khách hàng có rủi ro tín dụng cao, để phát hiện gian lận và các vấn đề về giá quyền chọn.