Y/c: Phân tích tìm hiểu thuận toán Random Forest

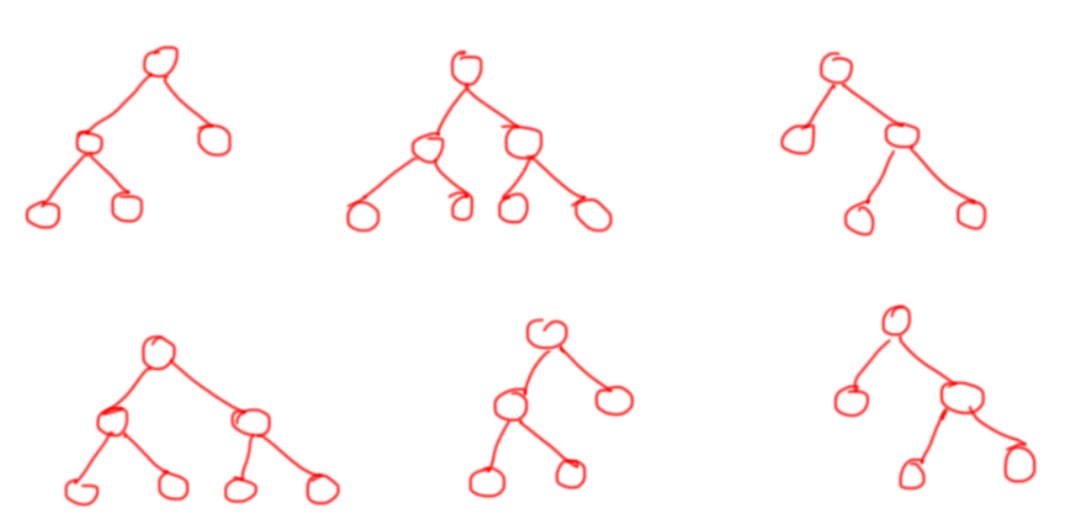
Lê Thị Thùy Dương \_ 52000321

**Random Forest Algorithm**

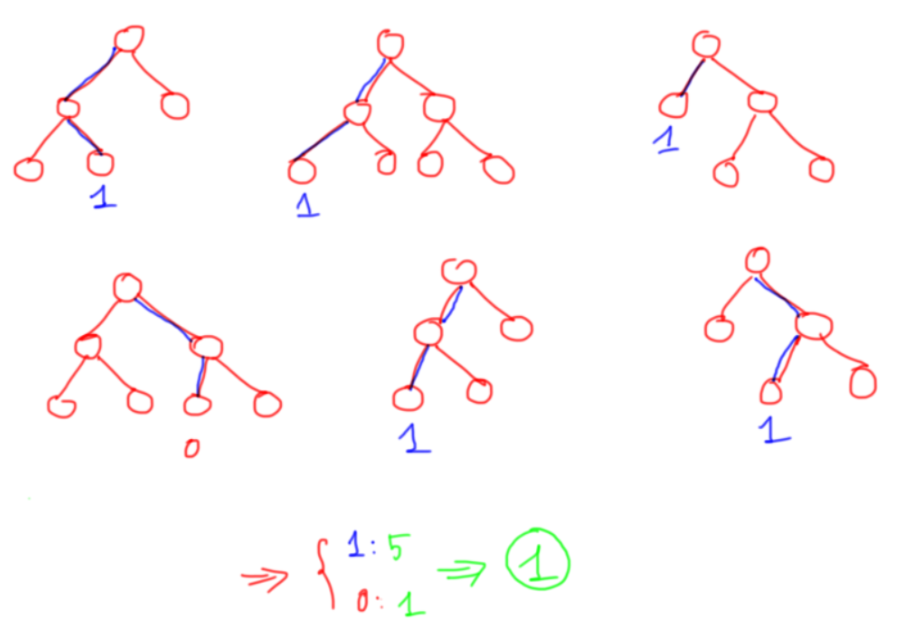
1. **Giới thiệu thuật toán:**

Random là ngẫu nhiên, Forest là rừng, nên ở thuật toán Random Forest sẽ xây dựng nhiều cây quyết định bằng thuật toán Decision Tree, tuy nhiên mỗi cây quyết định sẽ khác nhau (có yếu tố random). Sau đó kết quả dự đoán được tổng hợp từ các cây quyết định.

Ở bước huấn luyện thì xây dựng nhiều cây quyết định, các cây quyết định có thể khác nhau:



Sau đó ở bước dự đoán, với một dữ liệu mới, thì ở mỗi cây quyết định mình sẽ đi từ trên xuống theo các node điều kiện để được các dự đoán, sau đó kết quả cuối cùng được tổng hợp từ kết quả của các cây quyết định.



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 mình sẽ vote là cho ra dự đoán cuối cùng là 1.

1. **Xây dựng thuật toán:**

Bước 1: Lấy Mẫu Bootstrap

Giả sử chúng ta có tập dữ liệu huấn luyện gồm 100 mẫu. Chúng ta tạo ra 3 tập dữ liệu con (bootstrap samples) bằng cách lấy mẫu ngẫu nhiên với lặp lại. Mỗi tập dữ liệu con có thể bao gồm mẫu trùng lặp.

* Tập Dữ Liệu Con 1: [2, 5, 8, 2, 3, 7, 1, 9, 4, 6, ...]
* Tập Dữ Liệu Con 2: [1, 3, 5, 7, 2, 4, 8, 3, 6, 9, ...]
* Tập Dữ Liệu Con 3: [4, 6, 1, 8, 2, 9, 7, 5, 3, 1, ...]

Bước 2: Xây Dựng Cây Quyết Định

Với mỗi tập dữ liệu con, chúng ta xây dựng một cây quyết định. Mỗi cây được xây dựng thông qua việc chọn ngẫu nhiên một số tính năng tại mỗi nút và sử dụng thuật toán cây quyết định.

Bước 3: Dự Đoán từng Cây

Với mỗi cây, chúng ta dự đoán kết quả trên tập kiểm tra hoặc dữ liệu mới sử dụng cây đó. Mỗi cây sẽ có một dự đoán riêng.

* Cây 1 Dự Đoán: [0, 1, 0, 1, 0, 1, 1, 0, 1, 0, ...]
* Cây 2 Dự Đoán: [1, 0, 1, 0, 1, 0, 0, 1, 0, 1, ...]
* Cây 3 Dự Đoán: [0, 1, 1, 0, 0, 1, 1, 1, 0, 1, ...]

Bước 4: Kết Hợp Dự Đoán

Kết hợp dự đoán từ tất cả các cây để đưa ra dự đoán cuối cùng. Có thể sử dụng phương pháp đa số hoặc trung bình.

Kết Hợp Dự Đoán: [0, 1, 1, 0, 0, 1, 1, 1, 1, 1, ...]

Bước 5: Đánh Giá Hiệu Suất

So sánh kết quả dự đoán cuối cùng với nhãn thực tế trên tập kiểm tra để đánh giá hiệu suất của Random Forest.

Bước 6: Quan Trọng của Tính Năng

Thuật toán sẽ cung cấp thông tin về quan trọng của từng tính năng trong quá trình dự đoán.

Bước 7: Tinh Chỉnh Tham Số

Nếu cần, thực hiện tinh chỉnh các tham số như số lượng cây, độ sâu, và số lượng tính năng được chọn để tối ưu hóa hiệu suất.

Ví dụ trên chỉ là một tình huống giả định. Trong thực tế, Random Forest có thể được sử dụng cho nhiều loại dữ liệu và vấn đề khác nhau. Việc điều chỉnh tham số và hiểu rõ về quy trình làm việc của thuật toán giúp tối ưu hóa hiệu suất của mô hình.

1. **Ví dụ thực tế để hiểu thuật toán**

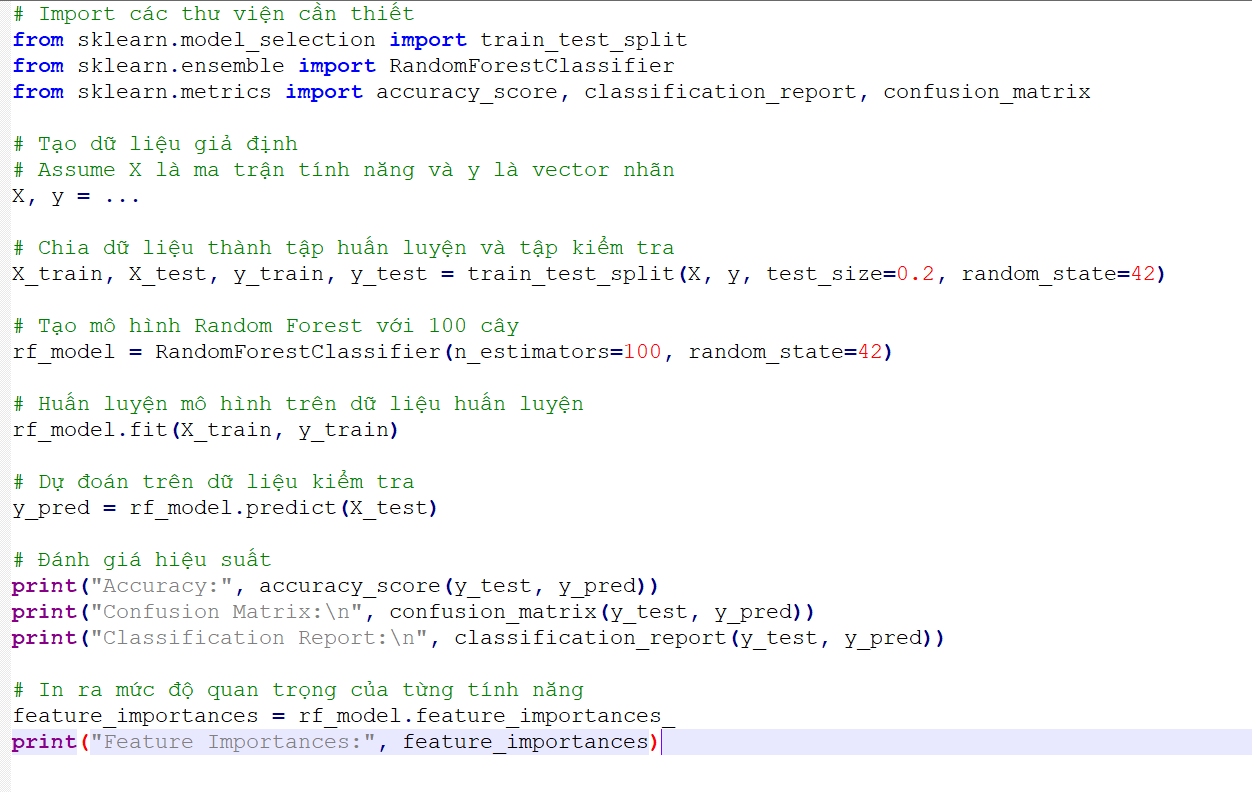
Giả sử bạn đã quyết định hỏi bạn bè và nói chuyện với họ về trải nghiệm du lịch trong quá khứ của họ đến những nơi khác nhau. Bạn sẽ nhận được một số khuyến nghị từ tất cả các bạn. Bây giờ bạn phải tạo danh sách các địa điểm được đề xuất. Sau đó, bạn yêu cầu họ bỏ phiếu (hoặc chọn địa điểm tốt nhất cho chuyến đi) từ danh sách các địa điểm được đề xuất bạn đã thực hiện. Địa điểm có số phiếu bầu cao nhất sẽ là lựa chọn cuối cùng của bạn cho chuyến đi.

Trong quá trình quyết định ở trên, có hai phần. Trước tiên, hãy hỏi bạn bè về trải nghiệm du lịch cá nhân của họ và nhận được đề xuất từ nhiều nơi họ đã ghé thăm. Điều này cũng giống như sử dụng thuật toán cây quyết định. Ở đây, mỗi người trong số các bạn chọn những nơi mà họ đã ghé thăm cho đến nay.Phần thứ hai, sau khi thu thập tất cả các khuyến nghị, là thủ tục bỏ phiếu để chọn địa điểm tốt nhất trong danh sách các khuyến nghị. Toàn bộ quá trình nhận được khuyến nghị từ bạn bè và bỏ phiếu cho họ để tìm ra nơi tốt nhất được gọi là thuật toán rừng ngẫu nhiên.

Về mặt kỹ thuật, nó là một phương pháp tổng hợp (dựa trên cách tiếp cận phân chia và chinh phục) của các cây quyết định được tạo ra trên một tập dữ liệu được chia ngẫu nhiên. Bộ sưu tập phân loại cây quyết định này còn được gọi là rừng. Cây quyết định riêng lẻ được tạo ra bằng cách sử dụng chỉ báo chọn thuộc tính như tăng thông tin, tỷ lệ tăng và chỉ số Gini cho từng thuộc tính. Mỗi cây phụ thuộc vào một mẫu ngẫu nhiên độc lập. Trong bài toán phân loại, mỗi phiếu bầu chọn và. lớp phổ biến nhất được chọn là kết quả cuối cùng Trong trường hợp hồi quy, mức trung bình của tất cả các kết quả đầu ra của cây được coi là kết quả cuối cùng. Nó đơn giản và mạnh mẽ hơn so với các thuật toán phân loại phi tuyến tính khác.

1. **Mã nguồn mẫu**

Dưới đây là một đoạn mã nguồn Python sử dụng thư viện scikit-learn để triển khai một mô hình Random Forest trên một tập dữ liệu giả định:



Hãy chắc chắn rằng đã có tập dữ liệu thực tế hoặc đã thay thế phần X, y = ... bằng dữ liệu thực tế. Mã nguồn trên sử dụng mô hình Random Forest để phân loại và đánh giá hiệu suất trên tập kiểm tra. Cũng có thể điều chỉnh các tham số của mô hình để tối ưu hóa hiệu suất trên dữ liệu cụ thể.