**BÀI TOÁN PHÂN LỚP (CLASSIFICATION) DỰA TRÊN THUẬT TOÁN RỪNG NGẪU NHIÊN.**

**1, Rừng ngẫu nhiên là gì?**

* Rừng ngẫu nhiên(Random Forest) là một thuật toán học có giám sát(Supervised Learing). Đúng như tên gọi thì rừng ngẫu nhiên sử dụng cây(tree) làm nền tảng cho thuật toán.
* Rừng ngẫu nhiên là một tập hợp của các Decision Tree, mà mỗi cây được lựa chọn theo một thuật toán dựa vào rừng ngẫu nhiên.

**\*Decision Tree là gì?**

* Decision Tree là tên đại diện cho một nhóm thuật toán phát triển dựa trên cây quyết định. Ở đó, mỗi Node của cây sẽ là các thuộc tính và các nhánh là các giá trị lựa chọn của các thuộc tính đó bằng cách đi theo các giá trị thuộc tính trên cây. Cây quyết định sẽ cho ta biết giá trị dự đoán.
* Nhóm thuật toán cây quyết định có điểm mạnh là có thể sử dụng cho cả bài toán Phân loại(Classification)

**2, Đặt vấn đề.**

Giả sử bạn muốn đi trên một chuyến đi và bạn muốn đi đến một nơi mà bạn sẽ thích.

Vậy bạn sẽ làm gì để tìm một nơi mà bạn sẽ thích? Bạn có thể tìm kiếm trực tuyến, đọc các bài đánh giá trên blog và các cổng thông tin du lịch hoặc bạn cũng có thể hỏi bạn bè của mình.

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. VD: 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.

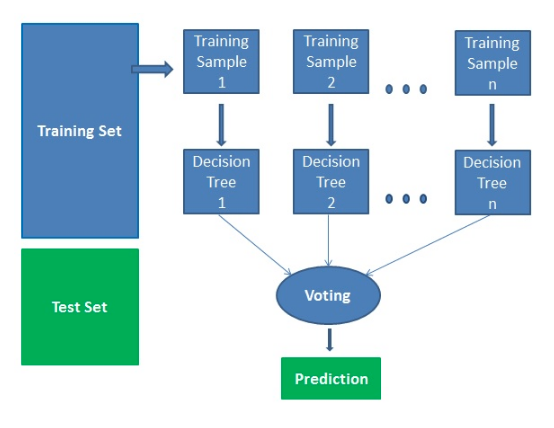
**3, Cách thức hoạt động.**

**Thuật toán:**

1. Chọn ngẫu nhiên ‘k’ features từ tập ‘m’features(k<m)
2. Từ tập ‘k’ features, tính toán ra node ‘d’ là nốt tốt nhất cho Node phân loại.
3. Chia các node con theo node tốt nhất vừa tìm được.
4. Lặp lại 1-3 cho đến khi đạt ‘k’ node.
5. Lặp lại 1-4 cho đến khi đạt ‘n’ node.

**Để biểu diễn dự đoán sử dụng RF, ta sử dụng các 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.



**4, Ưu điểm, nhược điểm của thuật toán.**

**Ưu điểm:**

* Random forests được coi là một phương pháp chính xác và mạnh mẽ vì số cây quyết định tham gia vào quá trình tính toán.
* Nó có nhiều cây hơn nên tránh vấn đề overfitting với các tập dữ liệu. Lý do chính là nó mất trung bình của tất cả các dự đoán, trong đó hủy bỏ những thành kiến.
* Thuật toán có thể được sử dụng trong cả hai vấn đề phân loại và hồi quy.
* Random forests cũng có thể xử lý các giá trị còn thiếu. Có hai cách để xử lý các giá trị này: sử dụng các giá trị trung bình để thay thế các biến liên tục và tính toán mức trung bình gần kề của các giá trị bị thiếu. Bạn có thể nhận được tầm quan trọng của tính năng tương đối, giúp chọn các tính năng đóng góp nhiều nhất cho trình phân loại.

**Nhược điểm:**

* Random forests chậm tạo dự đoán bởi vì nó có nhiều cây quyết định. Bất cứ khi nào nó đưa ra dự đoán, tất cả các cây trong rừng phải đưa ra dự đoán cho cùng một đầu vào cho trước và sau đó thực hiện bỏ phiếu trên đó.
* Tốn thời gian. Mô hình khó hiểu hơn so với cây quyết định, nơi bạn có thể dễ dàng đưa ra quyết định bằng cách đi theo đường dẫn trong cây.