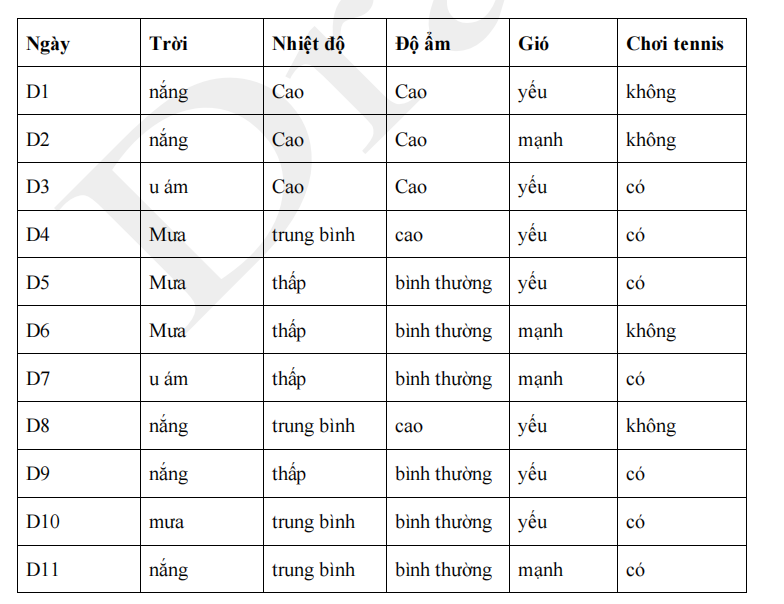
# Decesion Tree.

Node

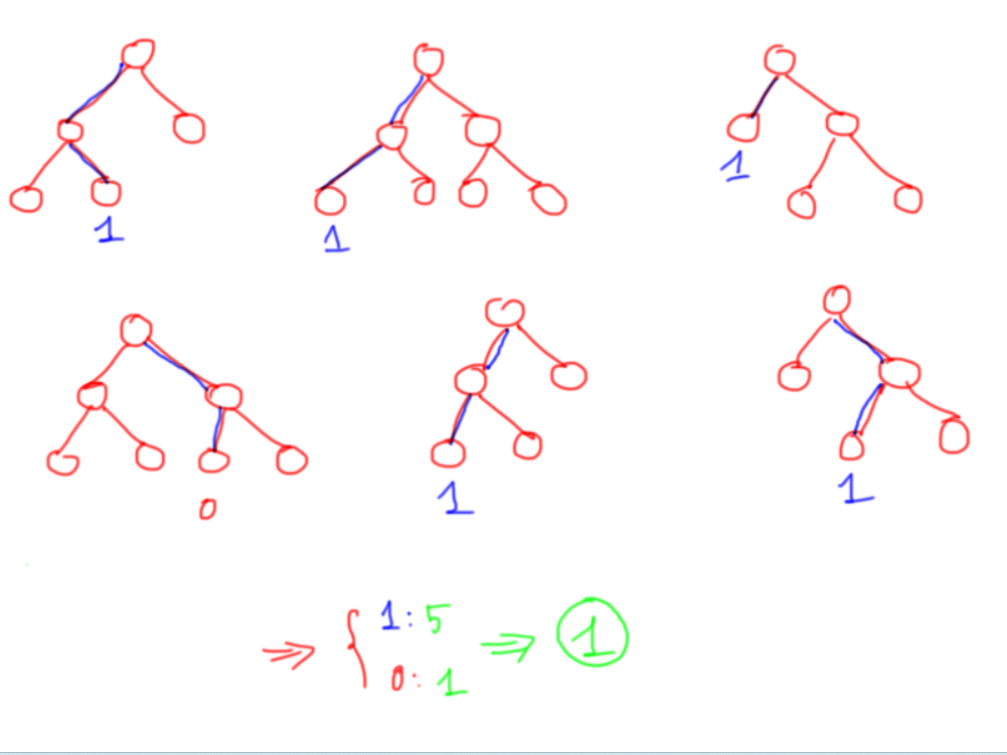
Cạnh

# 



# Random Forest

1. Phát triển từ Decesion Tree.
2. Tổng hợp kết quả từ nhiều cây khác nhau -> kết quả phổ biến nhất

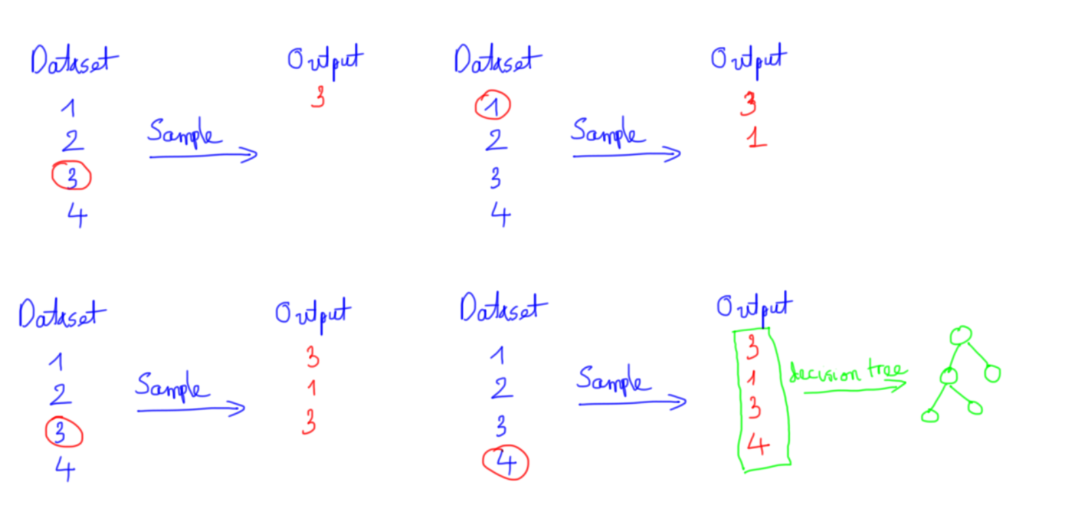


1. Các cây được lấy như sau:

Bằng cách Sampling dữ liệu, chọn 1 dữ liệu bất kỳ trong bộ dữ liệu nhưng không vứt nó đi, tương tự như thế ta được n điểm dữ liệu có thể trùng lặp nhau, thuộc tính cũng có thể giảm đi.

Với mỗi bộ dữ liệu random được thì chúng ta xây cây quyết định như bình thường thôi

1. Với cây quyết định ấy, vì cây được xây sẵn trên toàn bộ dữ liệu, thuộc tính tập train, vì vậy mô hình chắc chắn sẽ đúng 100% trên tập train nêú cây xây tốt, vì vậy cho nên dẫn đến vấn đề độ chính xác của mô hình trên tập Val ,test sẽ rất thấp vì cây gần như là được fix cứng rồi, thay đổi dữ liệu thì nó sẽ không còn đúng nữa ( overfitting )
2. Để giải quyết được nhược điểm này, người ta sinh ra Random Forest , với nhiều cây, mỗi cây dữ liệu không quá đầy đủ về bộ cũng như thuộc tính , không bị overfiting nhưng lại bị Underfitting hoặc bias, => tập hợp nhiều cây lại thì sẽ bù trừ các khuyết điểm này và dẫn đến một mô hình hoàn hảo hơn.



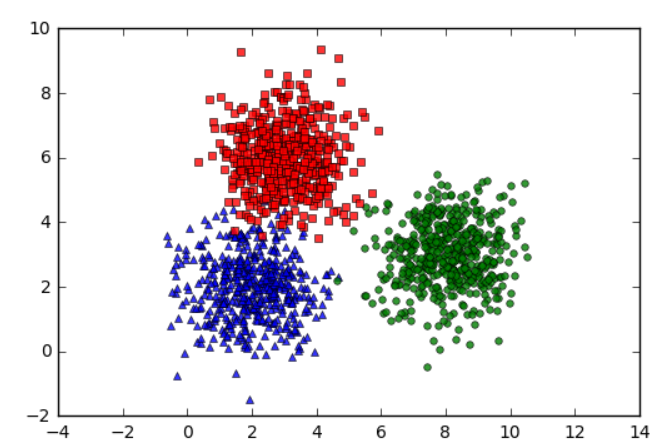
Kỹ thuật chia dữ liệu với RandomForest:

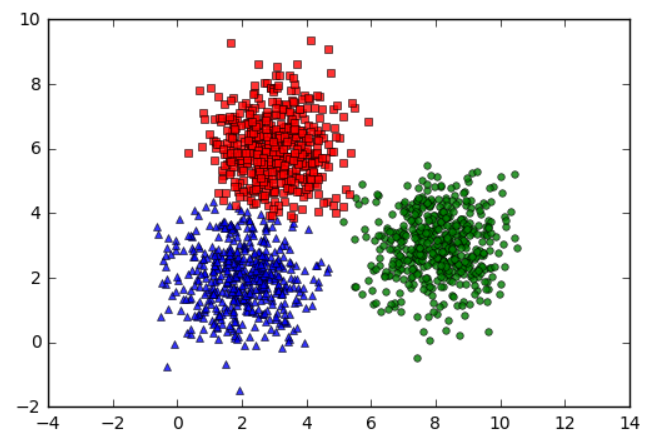
- CrossValidation : chia thành k bộ dữ liệu, mỗi cây sẽ có K-1 bộ dữ liệu để huấn luyện mô hình và 1 bộ để kiểm tra, lặp lại như thế ta được k-1 cây, từ đó ta lấy trung bình kết quả cacs cây

Kỹ thuật Sampling Bosstrap:

Chọn ngẫu nhiên bộ dữ liệu và có thể trùng nhau -> sẽ có những bộ chưa được chọn -> dùng để đánh giá mô hình -> khi không có nhiều dữ liệu để chia -> tiết kiệm mô hình -> hiệu suất cao hơn với bộ dữ liệu ít

1. Mean Clusstering :





# Phân cụm quang phổ : Special Clustering. ( Graph)

