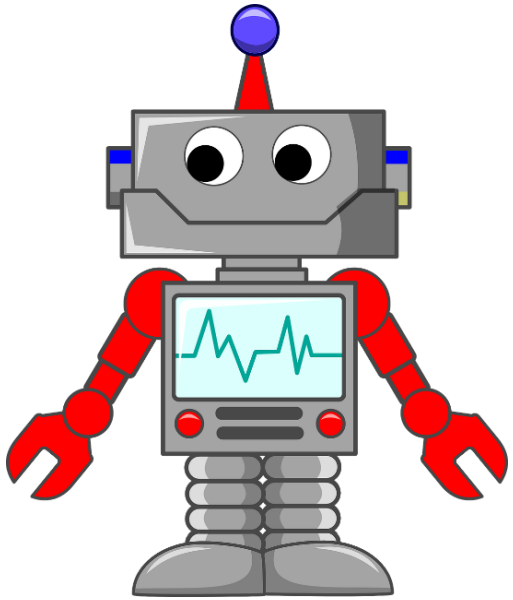
CS116 – LẬP TRÌNH PYTHON CHO MÁY HỌC

BÀI 10

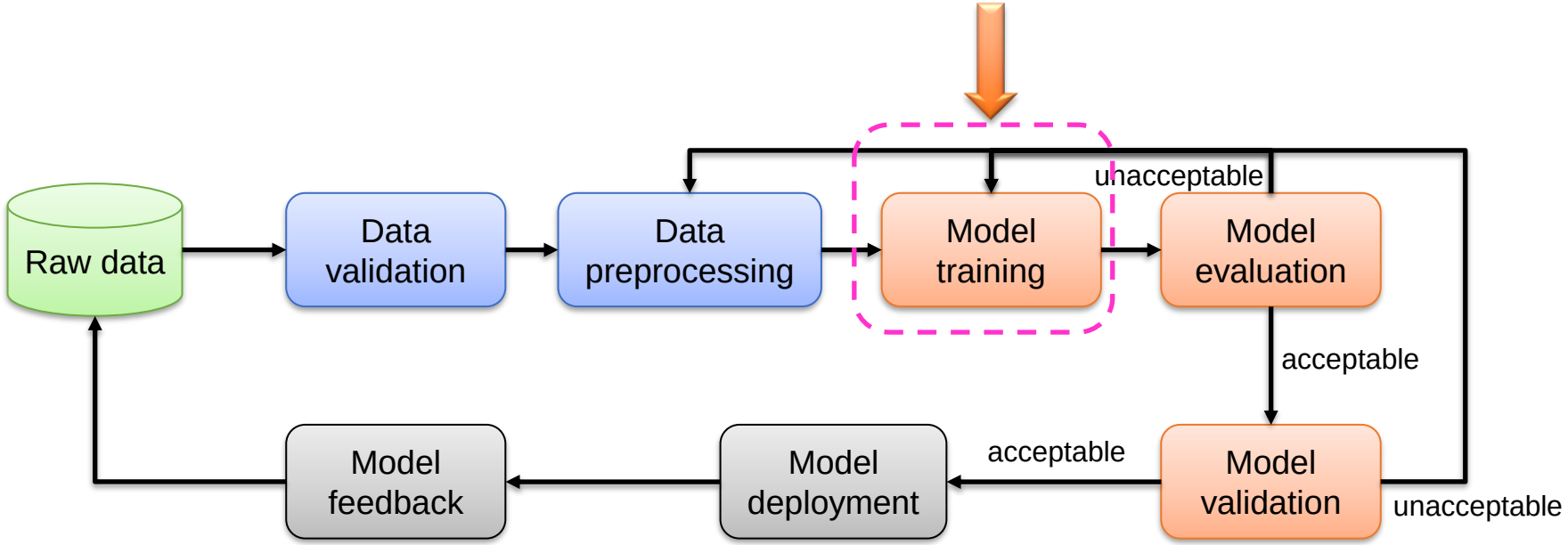
ENSEMBLE MODEL

Mô hình tổ hợp



TS. Nguyễn Vinh Tiệp

Vị trí của bài học



NỘI DUNG

1. TẠI SAO CẦN CÓ ENSEMBLE MODEL

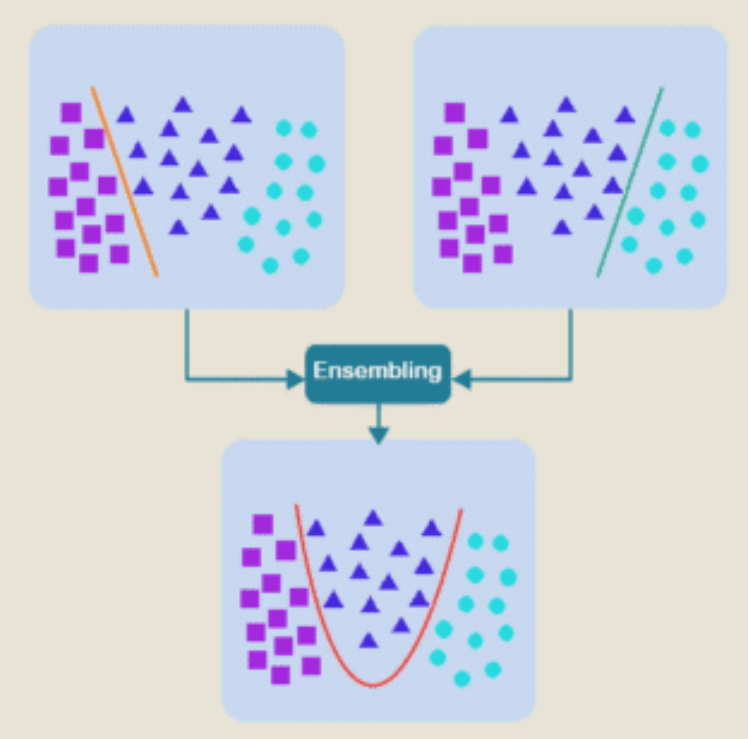
2. KỸ THUẬT CƠ BẢN: VOTING, AVERAGING, WEIGHTED AVG

3. KỸ THUẬT NÂNG CAO: STACKING, BLENDING, BAGGING, BOOSTING

Giới thiệu Ensemble Learning

1 tập con (tập dữ liệu ít hơn mà tổng quát vẫn cao)

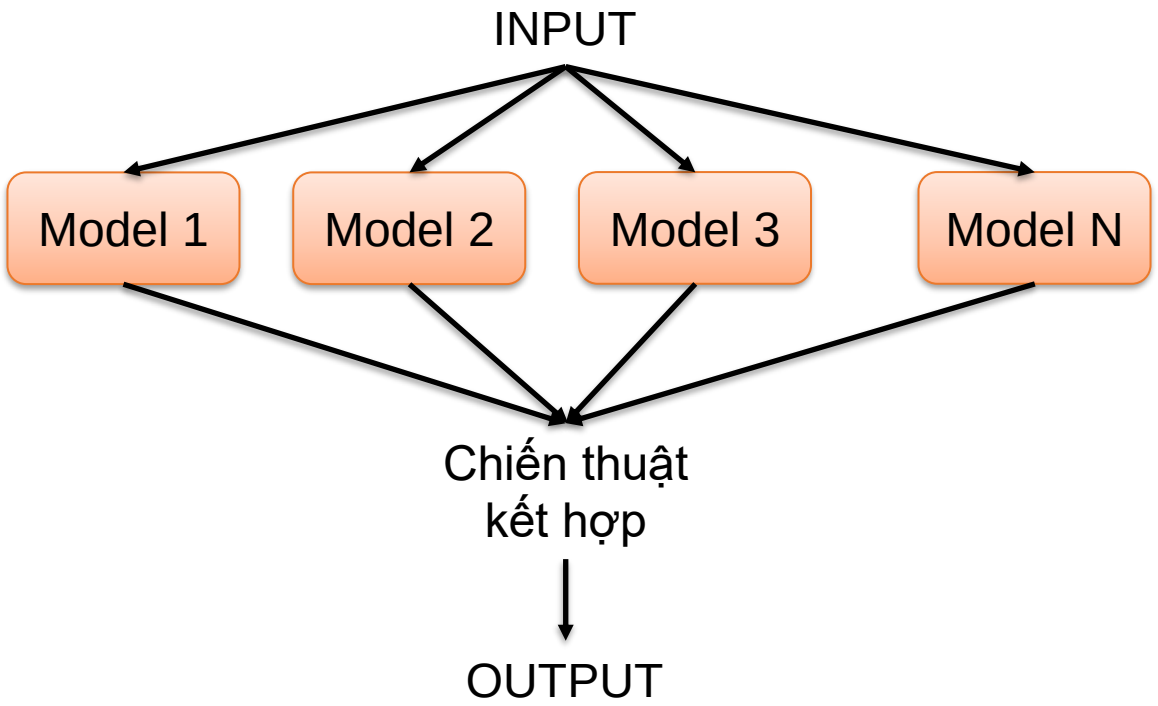
- **Mục tiêu máy học**: xây dựng mô hình có tính tổng quát hóa cao từ dữ liệu
- Có hai cách chính để cải thiện tính tổng quát hóa:
  - Cải thiện hiệu suất của một máy học (model)
  - Kết hợp nhiều mô hình và tổng hợp kết quả dự đoán ➔ Ensemble Learning



Minh họa ý tưởng của Ensemble learning

Tại sao Ensemble learning hiệu quả

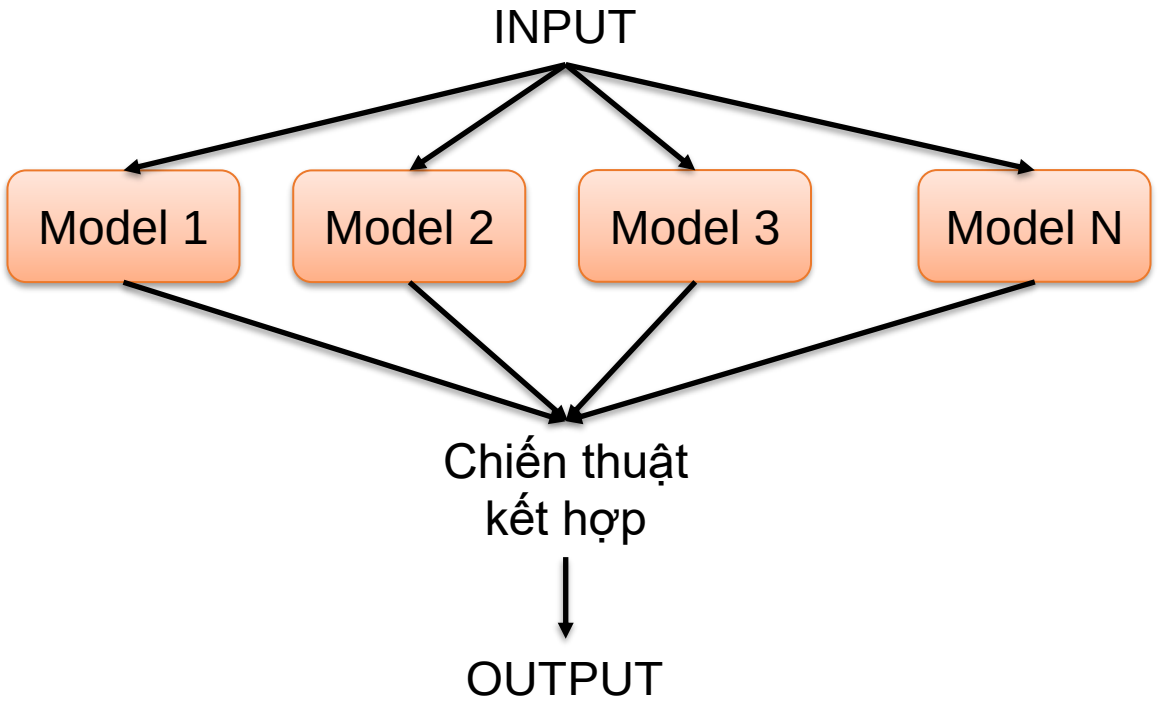
- **Vấn đề giảm Variance**:
  - Sử dụng nhiều mô hình có thể trung bình giá trị dự đoán gần với giá trị thực tế → giảm variance → tránh hiện tượng overfitting Dùng cộng trung bình
  - Thuật toán Random Forest kết hợp nhiều cây quyết định để giảm variance



Tại sao Ensemble learning hiệu quả

- **Vấn đề giảm bias**:
  - Mỗi mô hình "yếu" chỉ đoán đúng cho một số tình huống dữ liệu
  - Kết hợp nhiều mô hình "yếu" để tận dụng điểm mạnh mỗi mô hình, khắc phục những trường hợp mà từng mô hình đoán sai

kết hợp lấy các trường hợp mạnh lúc này tất cả đều được phân lớp mạnh



NỘI DUNG

1. TẠI SAO CẦN CÓ ENSEMBLE MODEL

2. KỸ THUẬT CƠ BẢN: VOTING, AVERAGING, WEIGHTED AVG

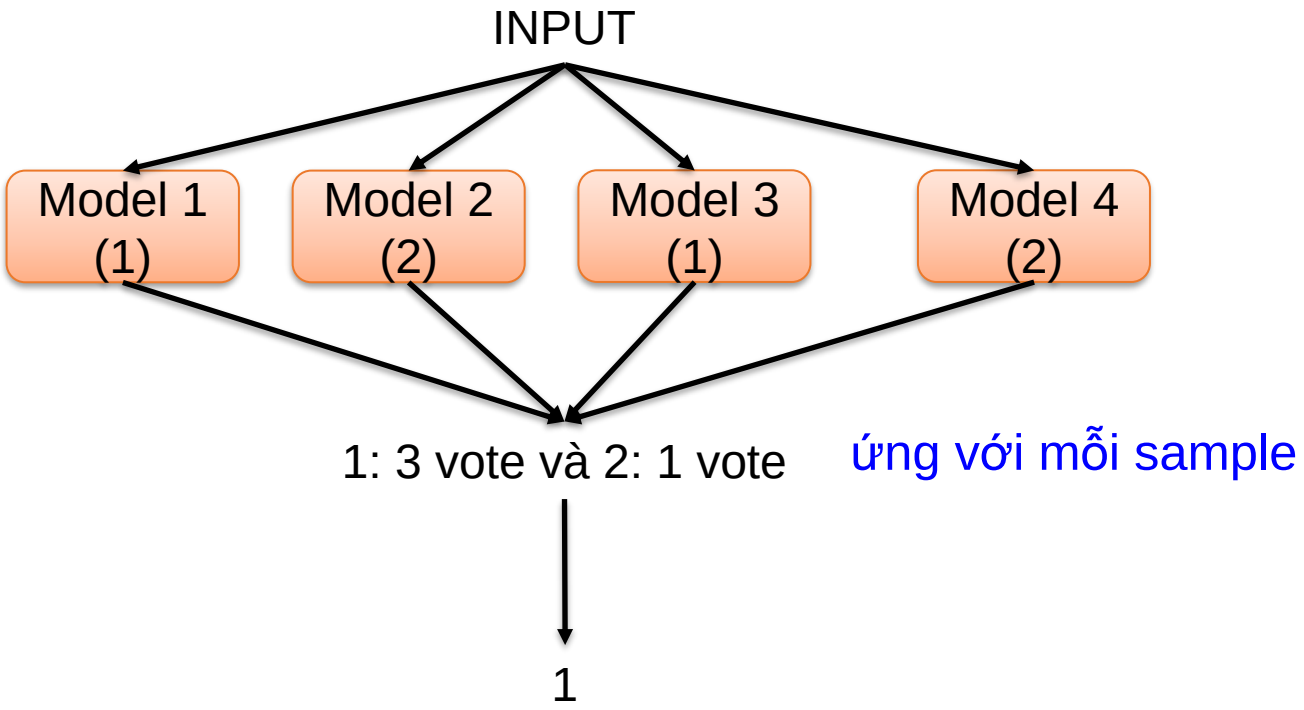
3. KỸ THUẬT NÂNG CAO: STACKING, BLENDING, BAGGING, BOOSTING

Kỹ thuật Voting

$y \in \{c_1, c_2 \cdots c_N\}$

- Thường dùng cho bài toán phân loại

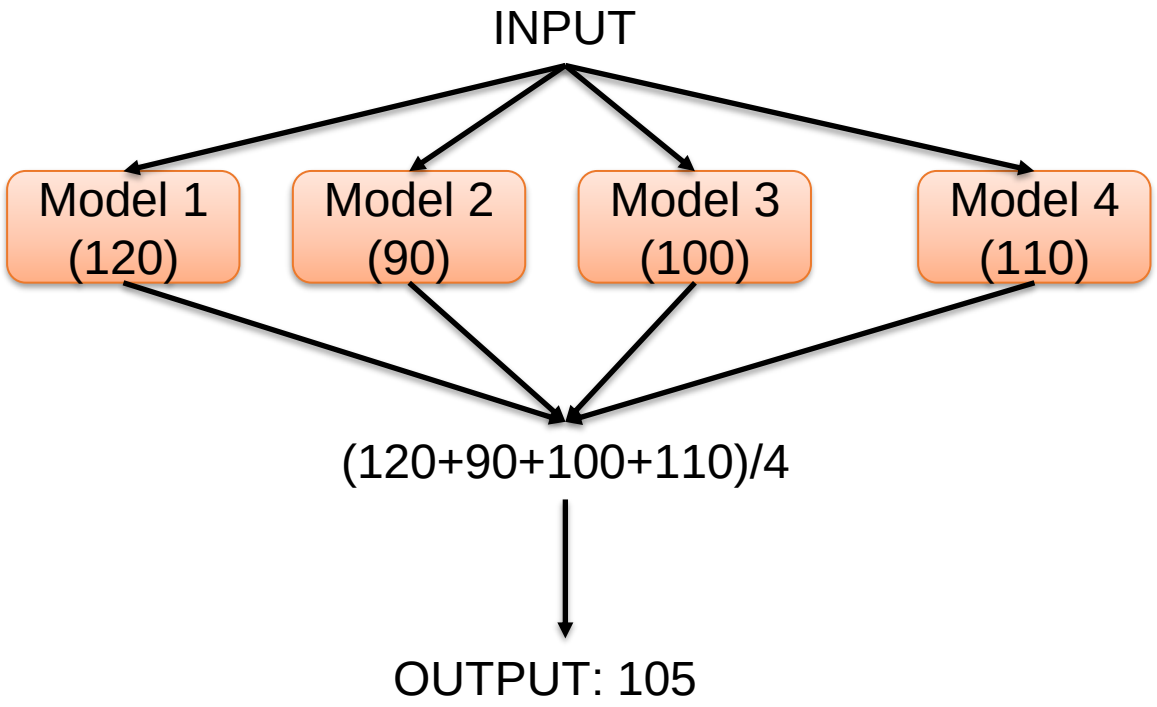
- Mỗi mô hình như một “cử tri”, quyết định cuối cùng thuộc về số đông



Kỹ thuật Averaging

- Thường dùng cho bài toán hồi quy

- Tính trung bình cộng kết quả của từng mô hình để tổng hợp

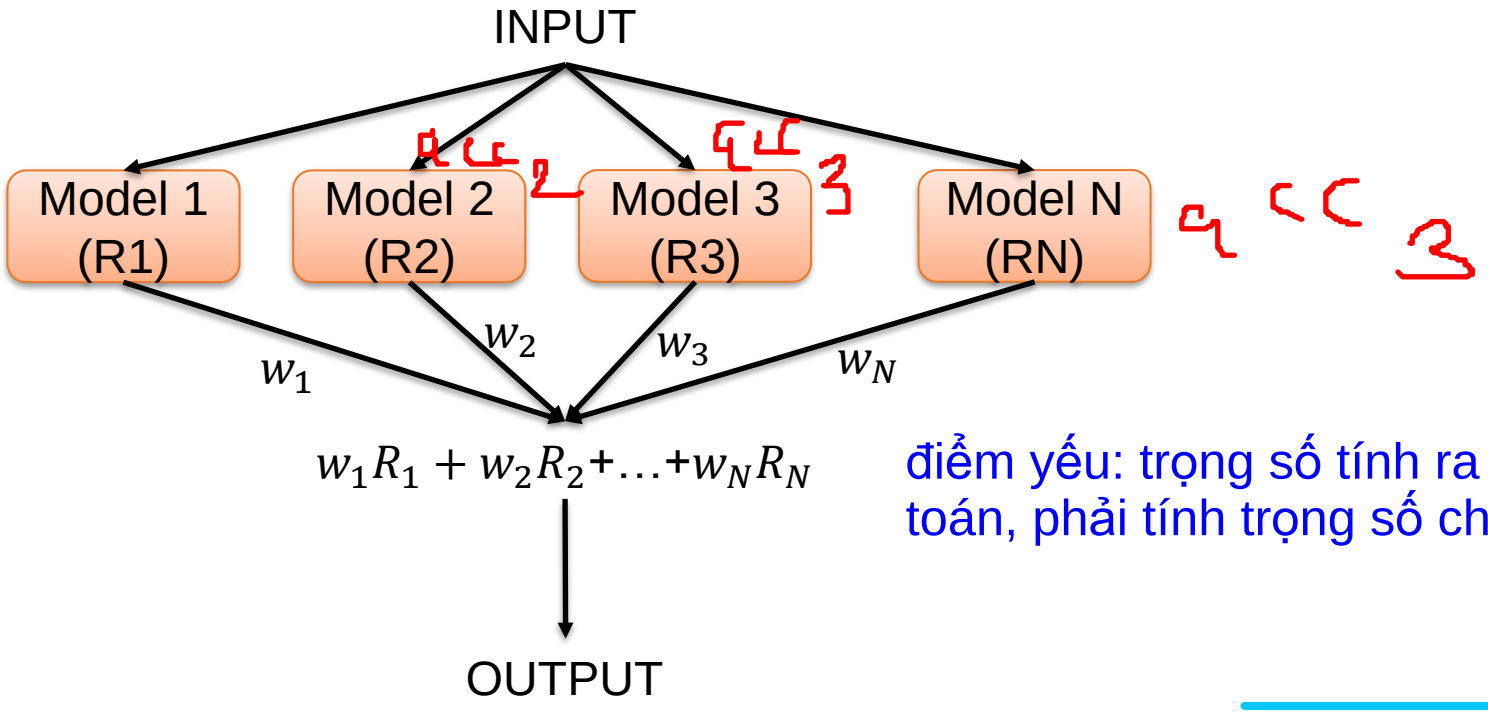


Kỹ thuật Weighted Averaging

- Mỗi mô hình có hiệu quả / trọng số khác nhau nên có trọng số khác nhau
- Trọng số có thể được tính từ độ chính xác trên tập train/validation

có mô hình có ưu điểm khác nhau => lấy trọng số cho công băng

dự đoán trên tập validation
=> acc1



điểm yếu: trọng số tính ra tốn chi phí tính toán, phải tính trọng số cho phù hợp

NỘI DUNG

1. TẠI SAO CẦN CÓ ENSEMBLE MODEL

2. KỸ THUẬT CƠ BẢN: VOTING, AVERAGING

trộn dữ liệu gốc với kết quả dự đoán

3. KỸ THUẬT NÂNG CAO: STACKING, BLENDING, BAGGING,

BOOSTING

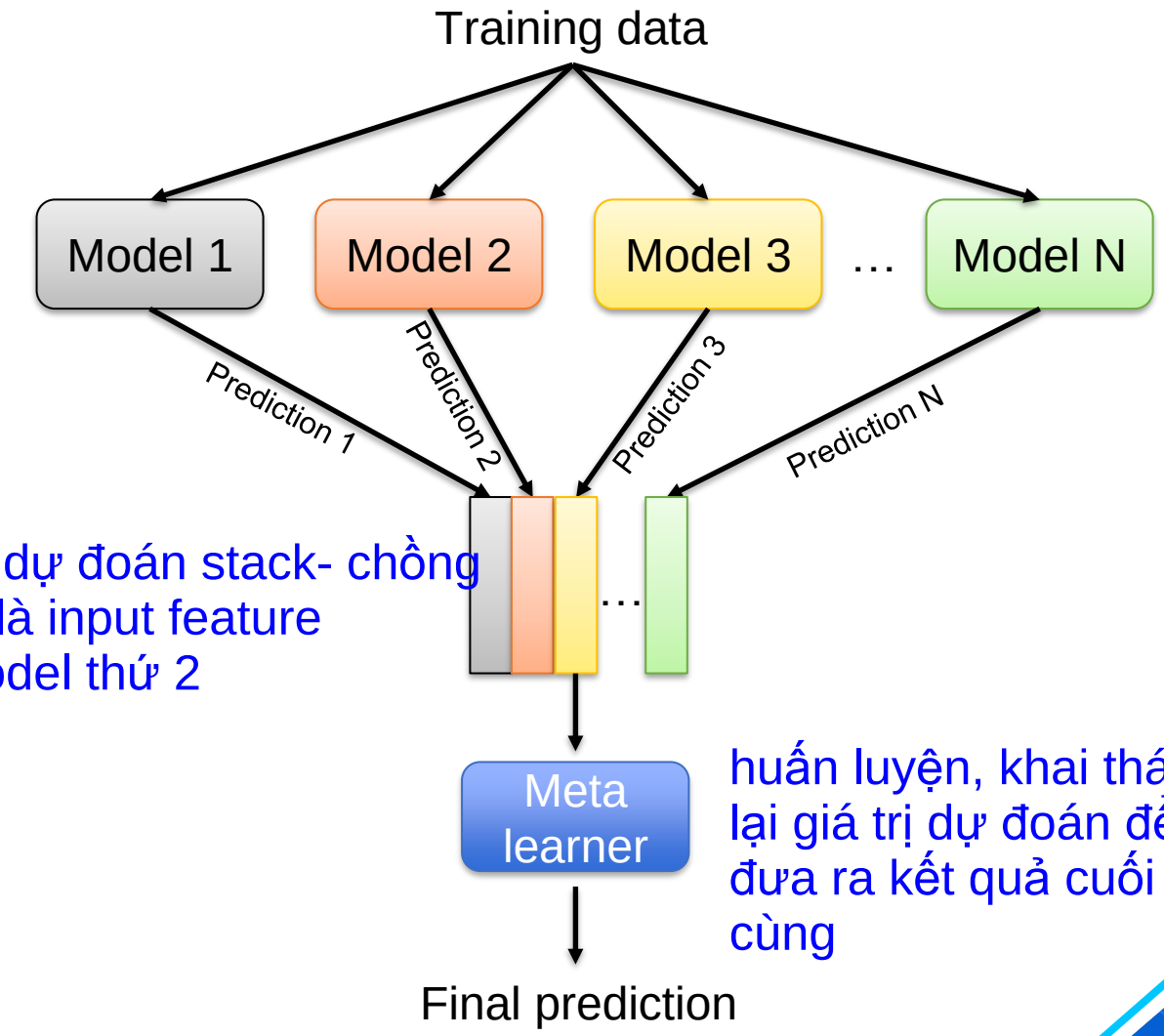
chồng lớp kết quả model

học song song

học tuần tự cho mô hình trước sau học tuần tự để kế thừa

Kỹ thuật Stacking

- Sử dụng kết quả dự đoán của tập train làm đặc trưng để huấn luyện mô hình tổng hợp (meta learner)



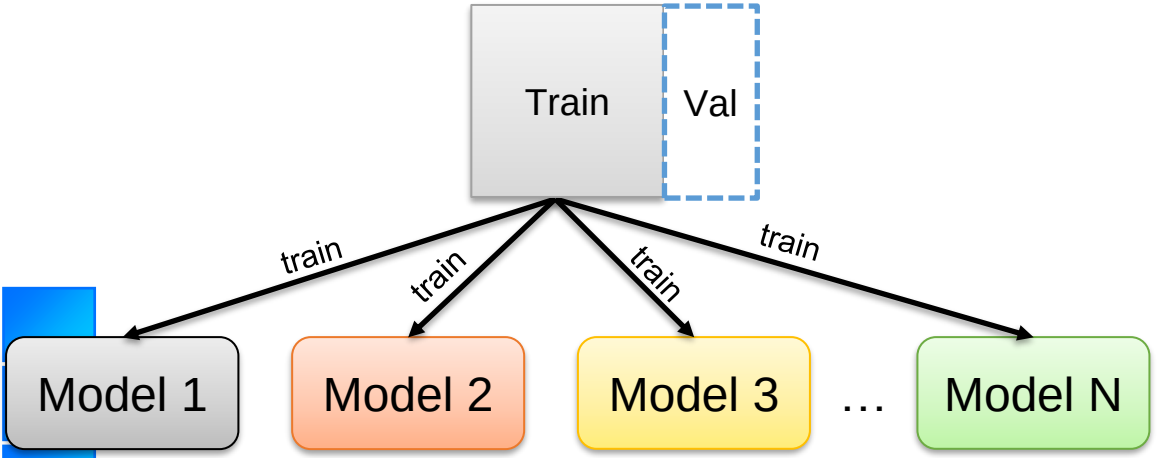
tất cả giá trị dự đoán stack- chồng lên nhau=> là input feature cho tầng model thứ 2

huấn luyện, khai thác lại giá trị dự đoán để đưa ra kết quả cuối cùng

Kỹ thuật Blending

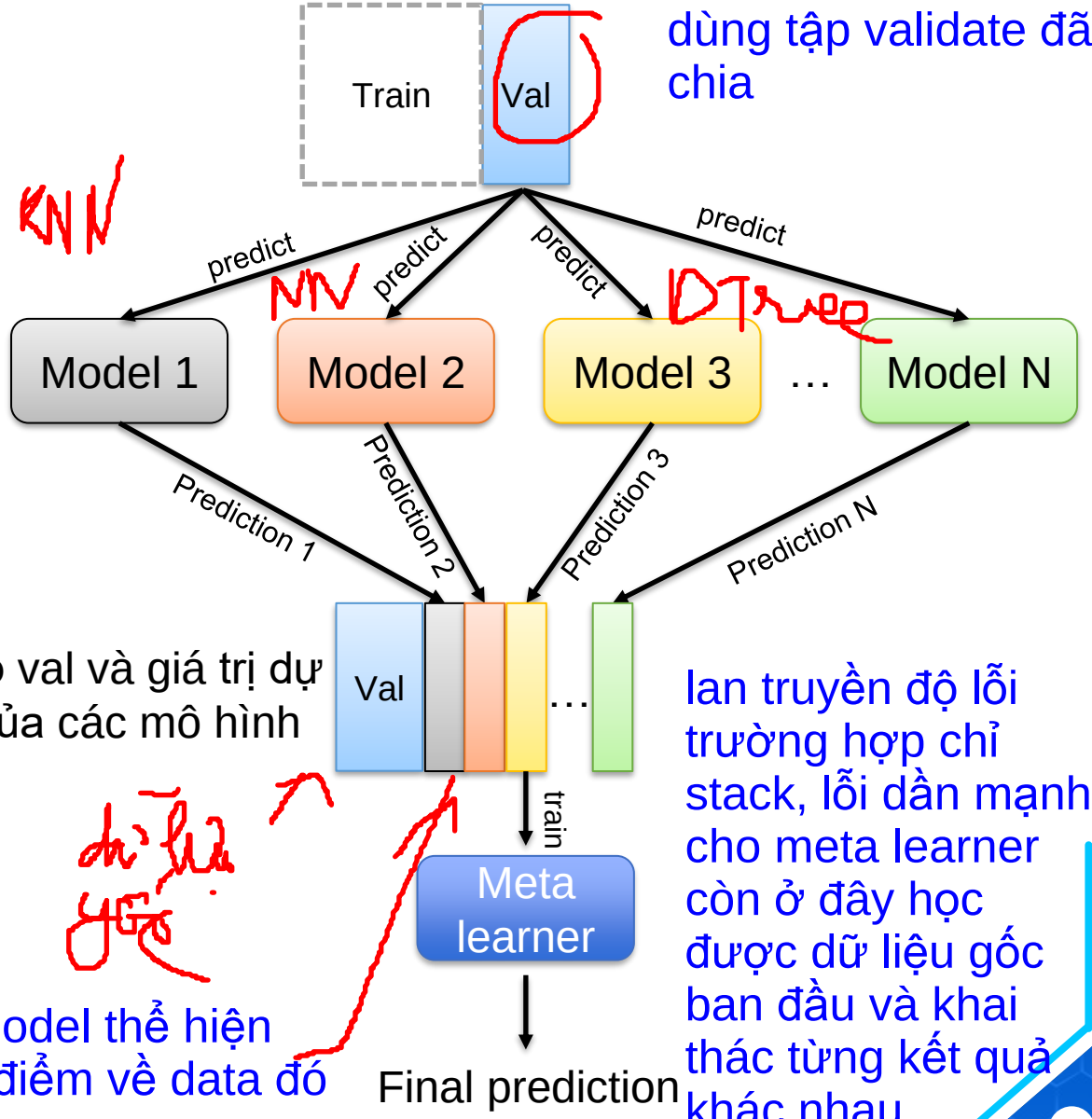
- Sử dụng (đặc trưng + kết quả dự đoán của tập validation) làm đặc trưng huấn luyện mô hình tổng hợp

Bước 1



huấn luyện để nó giải quyết được trên tập train

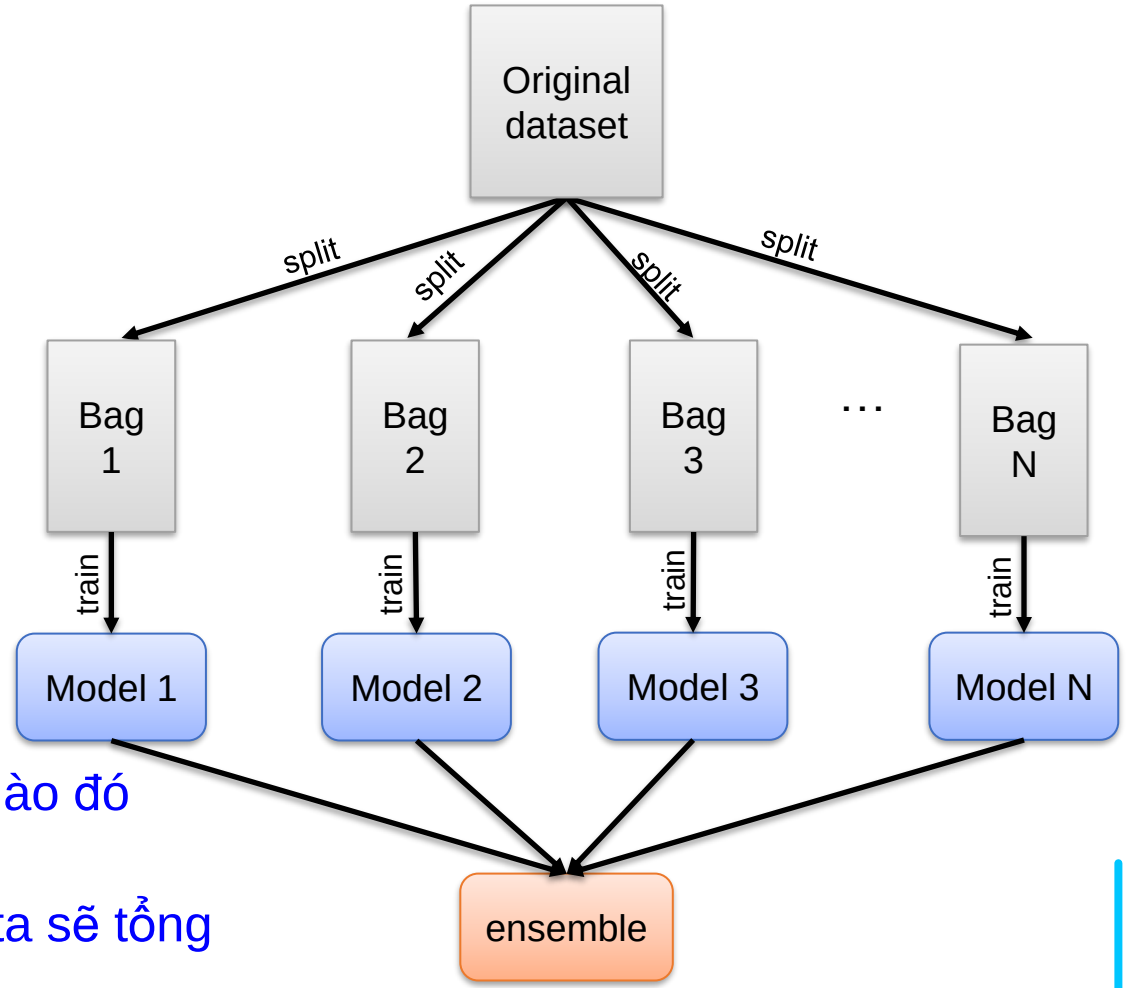
bước 2



Kỹ thuật Bagging

- Khác với 2 thuật toán trước, Bagging sử dụng cùng một thuật toán cho tất cả mô hình con cùng 1 model
- Bagging huấn luyện độc lập các mô hình con, trên các tập con của dataset

sau đó có thể dùng voting, avr, wavr



có thể giúp tổng quát không phụ thuộc 1 tập dữ liệu nào đó

model 1 rất tốt trên bag1, model 2 rất tốt trên bag2,...
nhưng yếu trên bag chưa được train, khi ensemble ta sẽ tổng hợp được những điểm chạy tốt, dự đoán tốt

hiệu quả vì chia dataset thì ta có tính tổng quát hơn, vì không quá phụ thuộc vào 1 mẫu dữ liệu nào đó, ví dụ 1 model train trên toàn bộ dataset thì có thể chỉ train bias, 1 vài điểm mạnh, còn cách này toàn diện điểm mạnh tình huống hơn

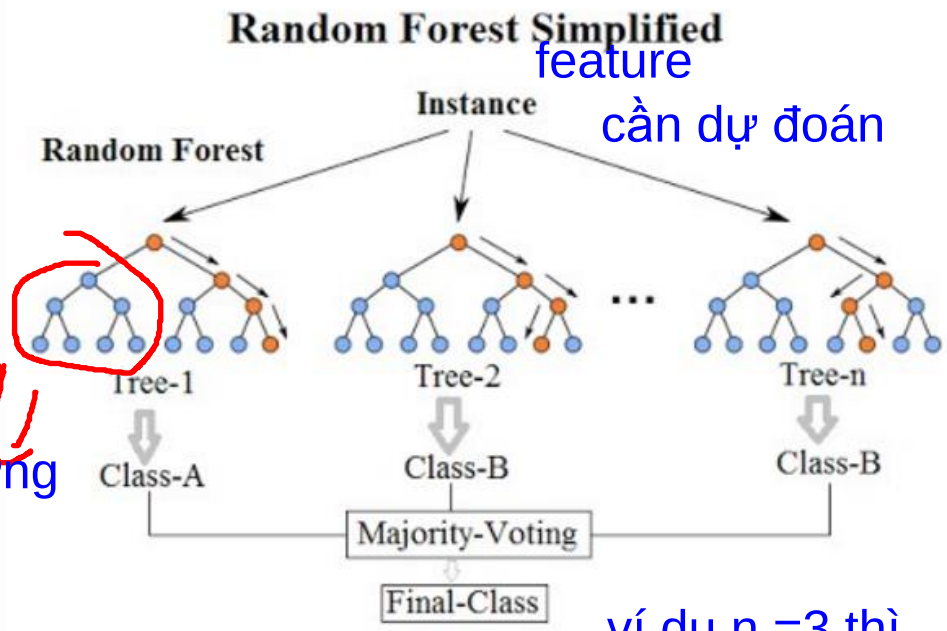
Kỹ thuật Bagging

decision tree là model thành phần

- Điển hình của kỹ thuật này là Random Forest

regressor hay classifier

- Ưu điểm:
  - **Hiệu quả**: chính xác, tổng quát hóa cao
  - **Tiện lợi**: có thể thực hiện được trên số đặc trưng lớn mà không cần phân tích đặc trưng
  - **Linh hoạt**: dùng cho cả hồi quy và phân lớp
  - Có thể song song hóa thuật toán
  - **Bền vững**: ít bị ảnh hưởng bởi outlier, ít khả năng overfitting



feature cần dự đoán

mỗi core CPU xử lý 1 cây => độc lập

chỉ bị ảnh hưởng cục bộ

chỉ overfit trên 1 bag

ví dụ n =3 thì kết quả = B

khi có rất nhiều đặc trưng tại 1 cây, tại 1 node làm việc cho 1 feature đưa ra quyết định cuối đến node lá rải đều ra các cây, phân tán feature, không quá bias

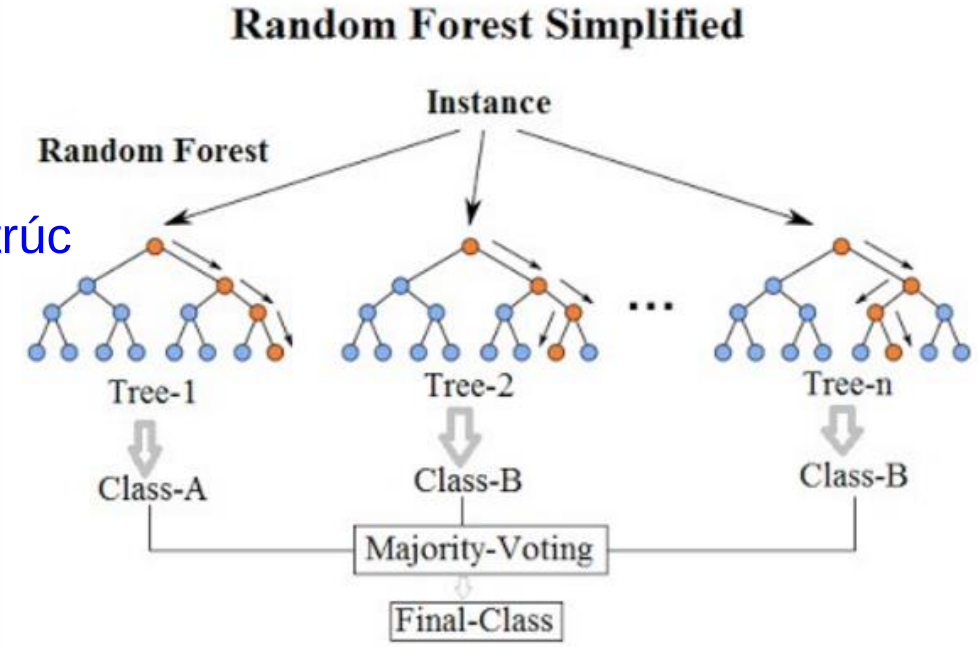
phối hợp toàn bộ feature

Kỹ thuật Bagging

tổng hợp feature rải ra rất nhiều => khó cảm nhận tại sao chia cây như vậy, feature tương tác nhau khó giải thích

- Một số khuyết điểm:
  - Khó giải thích mô hình
  - Độ phức tạp tính toán cao do xây từng cấu trúc cây
  - Bias với dữ liệu không cân bằng

mẫu dữ liệu có nhãn quá thiên lệch

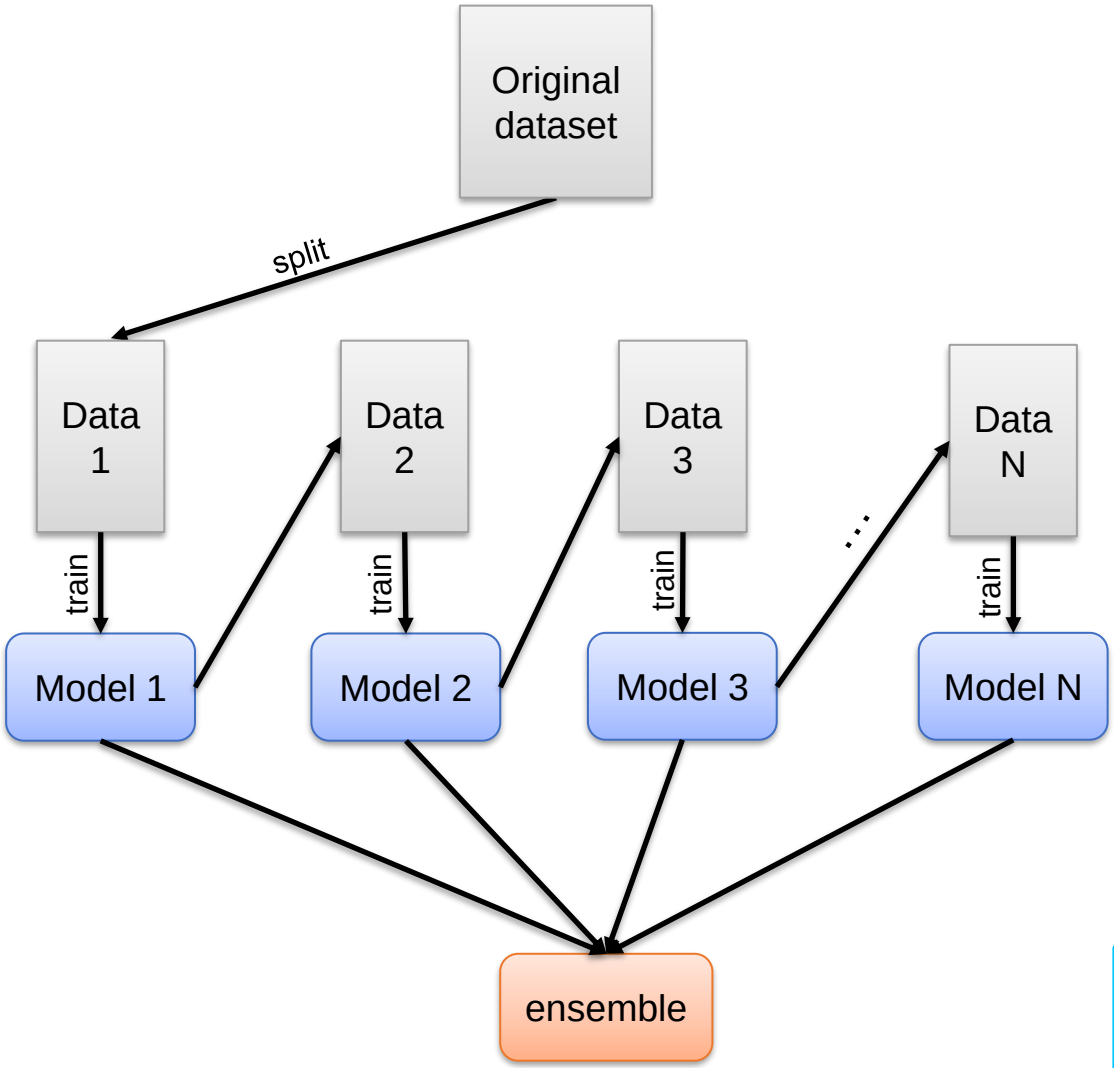


- Các mô hình điển hình: Random Forest, Bagged CART

A, B , C random nên xác suất dữ liệu A bị cao do ban đầu A đã có quá nhiều dữ liệu, bias ngay từ dataset

Kỹ thuật Boosting

- Boosting huấn luyện một cách tuần tự: mô hình sau được train dựa theo kết quả của mô hình trước đó để cố gắng sửa các lỗi sai còn lại



Kỹ thuật Boosting – Gradient Boost

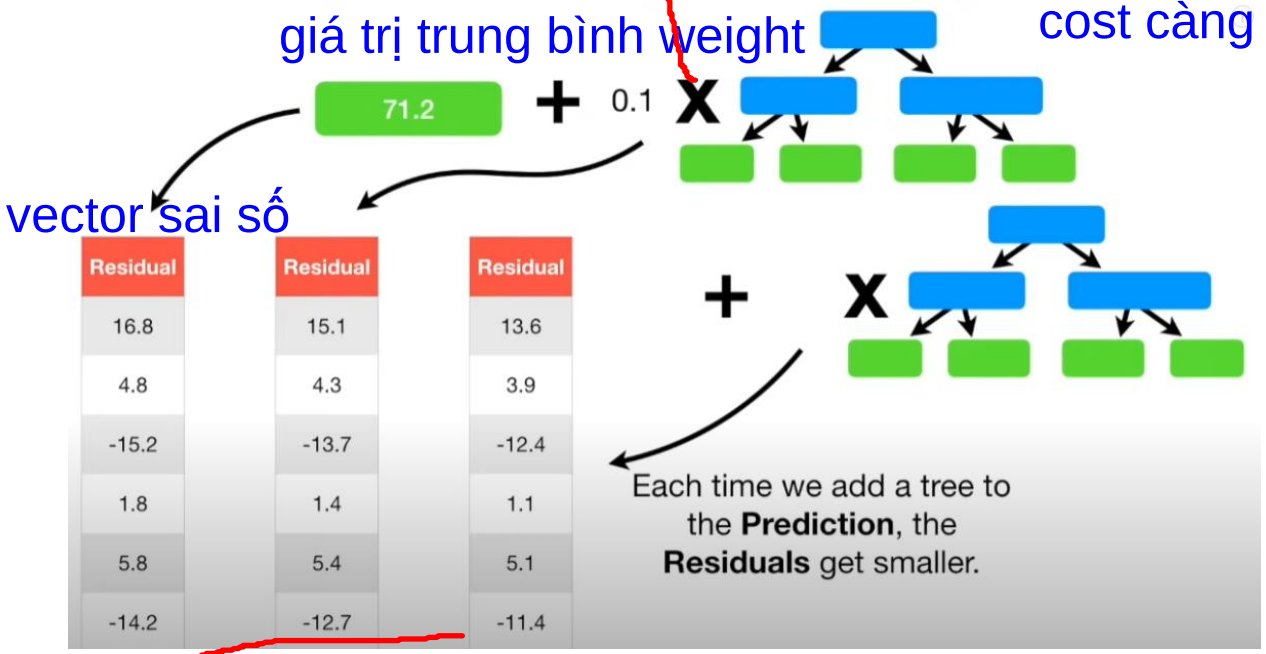
- **Ý tưởng**: xây dựng chuỗi cây quyết định liên tiếp, cây sau làm giảm sai số dự đoán của các cây trước

learning rate, càng nhỏ thì số cây càng dài, càng lớn

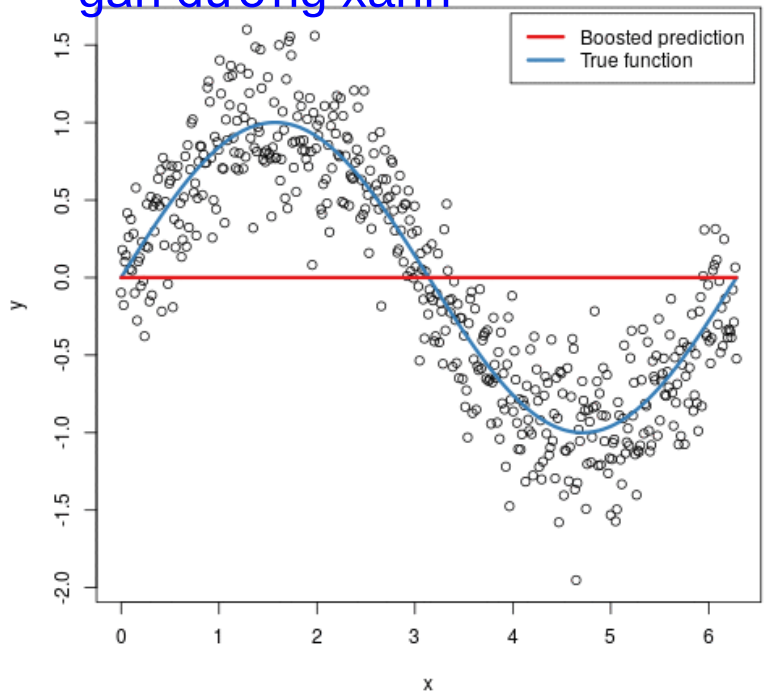
càng chạy đường đỏ càng mịn hơn gần đường xanh

giá trị trung bình weight

cost càng lớn



Nguồn Youtube: https://www.youtube.com/watch?v=3CC4N4z3GJc

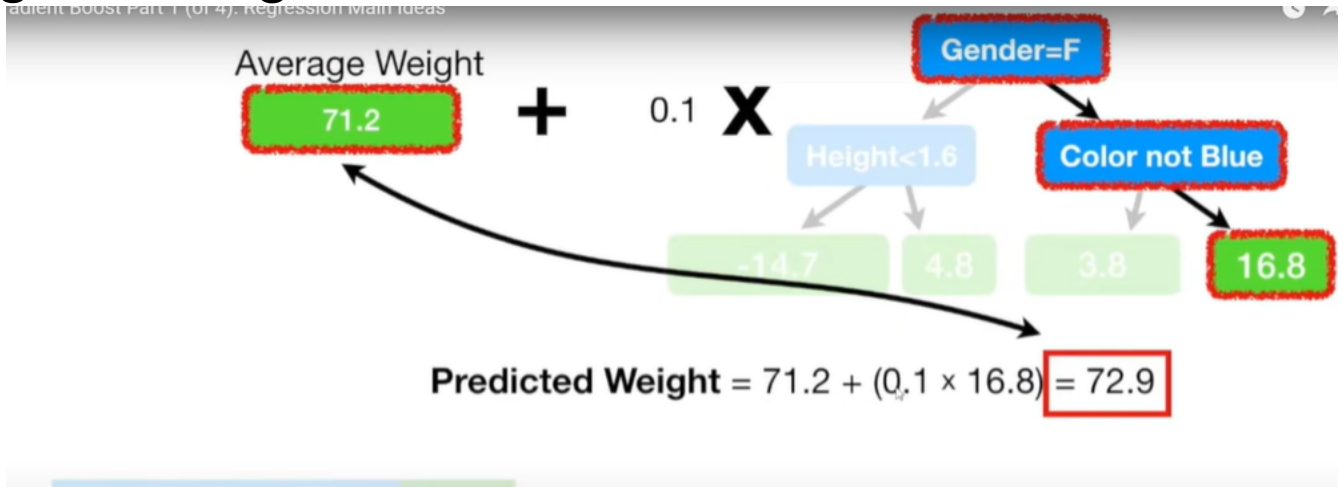


Nguồn: https://uc-r.github.io/gbm_regression

độ lỗi mới, cây mới, lấy khoảng 10% giá trị dự đoán

Kỹ thuật Boosting

- Một số thuật toán Boosting nổi tiếng:
  - AdaBoost
  - GradientBoost
  - XGBoost
  - LightGBM
  - CatBoost



- Đây đều là các thuật toán đạt giải cao trong các cuộc thi của Kaggle

Tổng kết

- Ensemble learning là kỹ thuật quan trọng để mô hình có tính tổng quát cao

- Bagging và Boosting là hai kỹ thuật nâng cao có tính hiệu quả cao

  các cuộc thi kaggle thường sử dụng

- Trong quá trình sử dụng cần chọn các siêu tham số cho phù hợp bằng phương pháp tinh chỉnh tham số

BÀI QUIZ VÀ HỎI ĐÁP