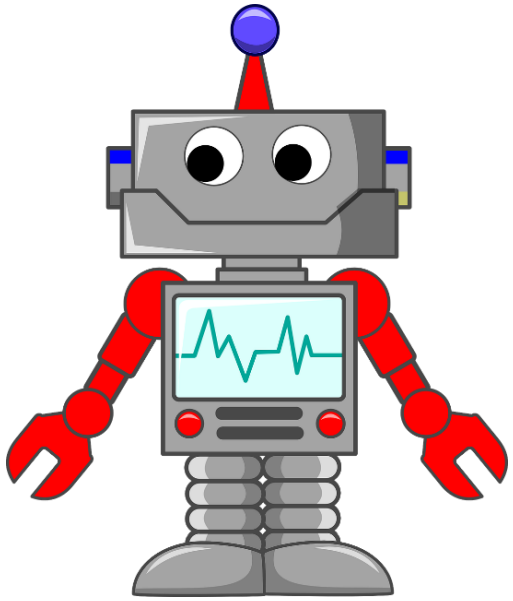
CS116 – LẬP TRÌNH PYTHON CHO MÁY HỌC

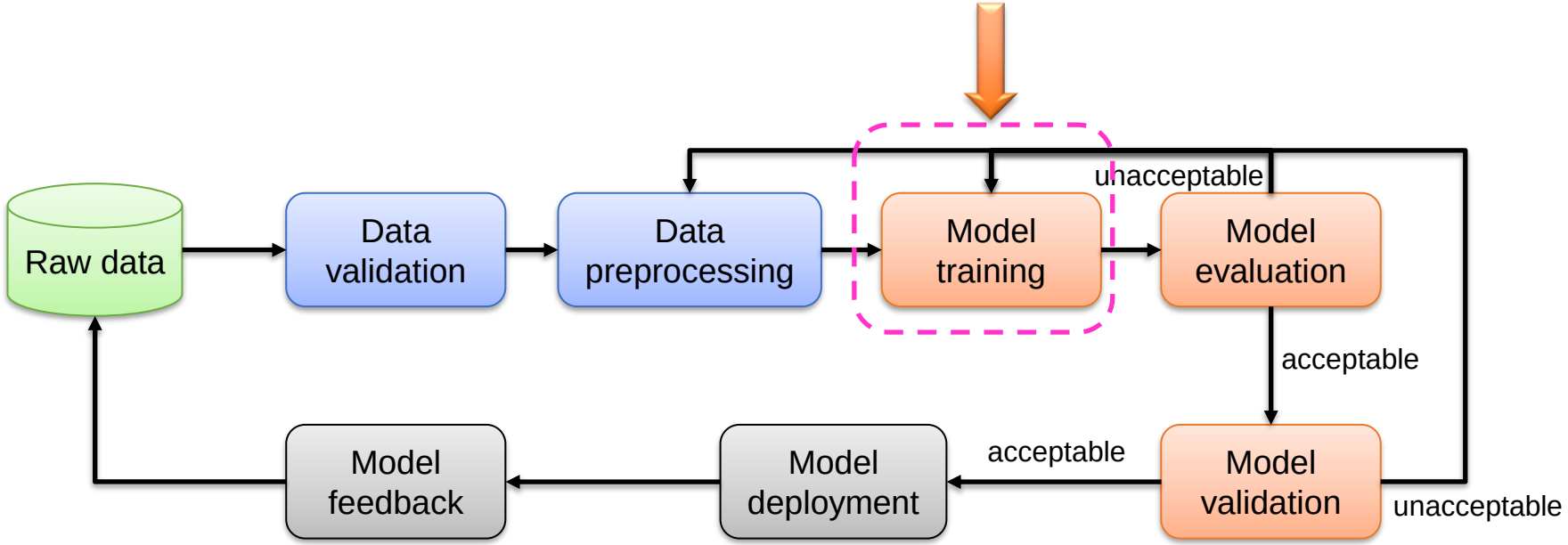
BÀI 10

ENSEMBLE MODEL



TS. Nguyễn Vinh Tiệp

Vị trí của bài học



NỘI DUNG

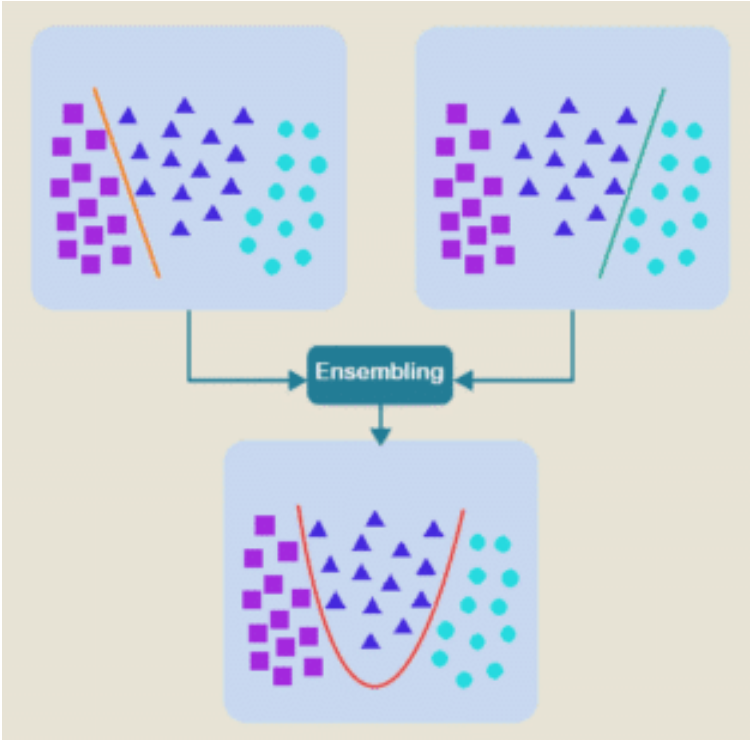
1. TẠI SAO CẦN CÓ ENSEMBLE MODEL

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

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

Giới thiệu Ensemble Learning

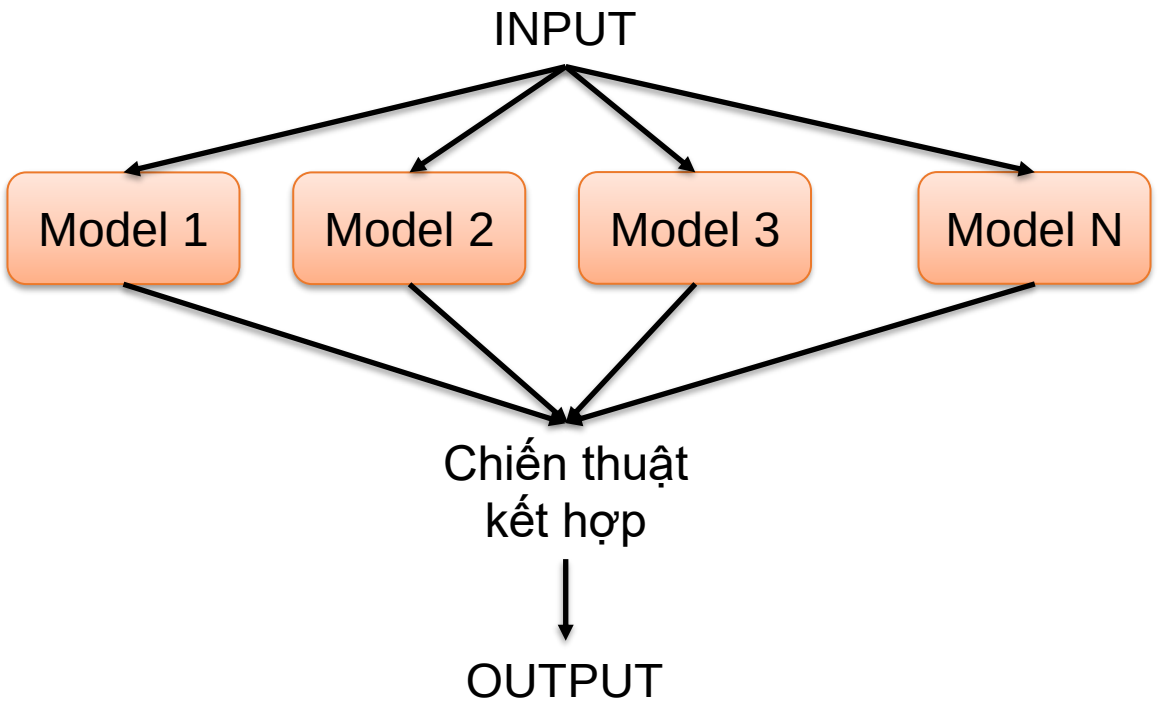
- **Mục tiêu máy học**: xây dựng mô hình có tính tổng quát hóa cao từ dữ liệu
- Có hai cách chính để cải thiện tính tổng quát hóa:
  - Cải thiện hiệu suất của một máy học (model)
  - Kết hợp nhiều mô hình và tổng hợp kết quả dự đoán ➔ Ensemble Learning



Minh họa ý tưởng của Ensemble learning

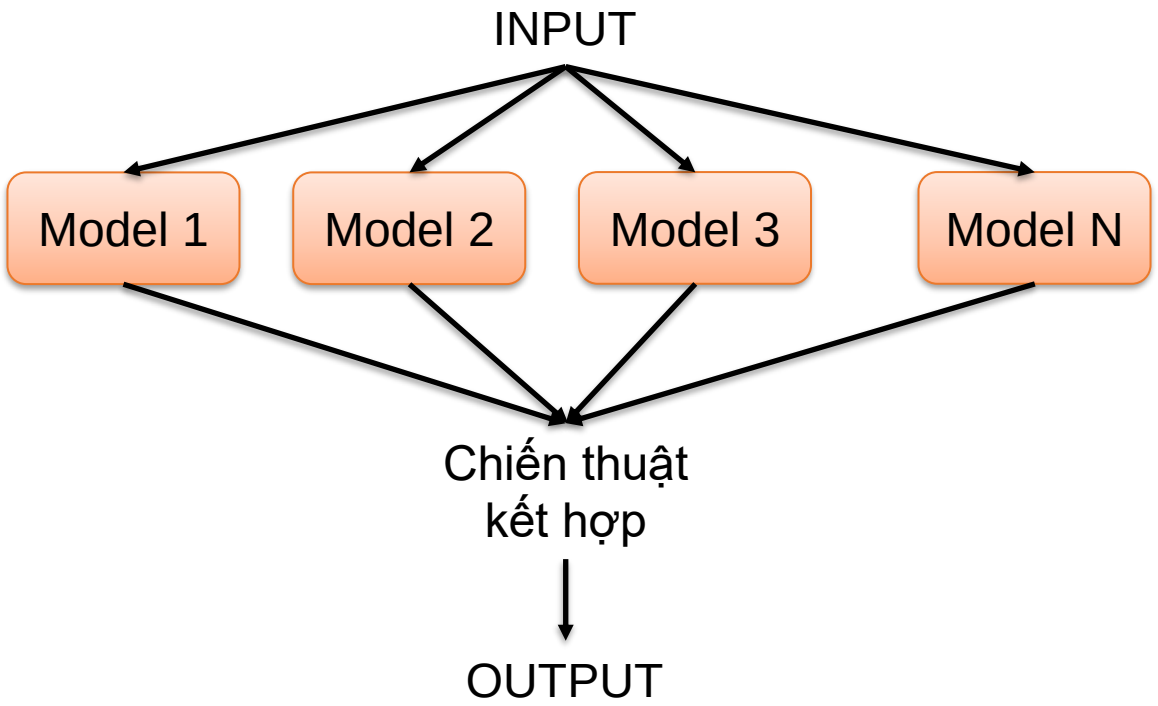
Tại sao Ensemble learning hiệu quả

- **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
  - 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



NỘI DUNG

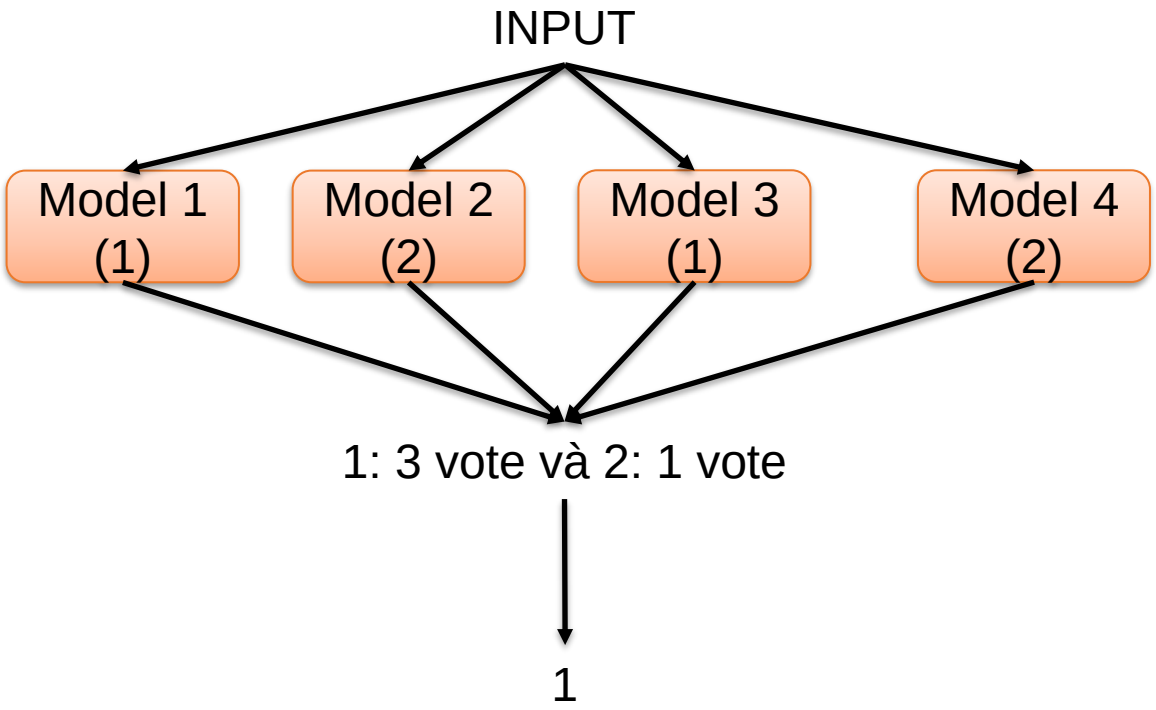
1. TẠI SAO CẦN CÓ ENSEMBLE MODEL

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

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

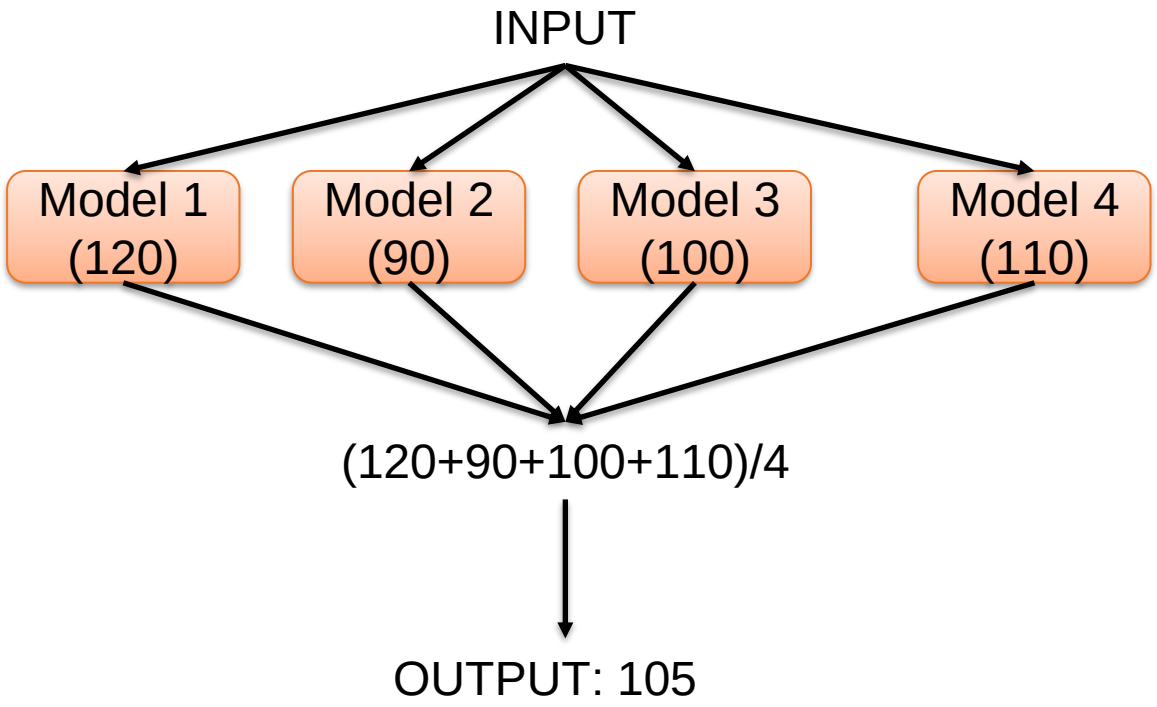
Kỹ thuật Voting

- Thường dùng cho bài toán phân loại
- Mỗi mô hình như một “cử tri”, quyết định cuối cùng thuộc về số đông



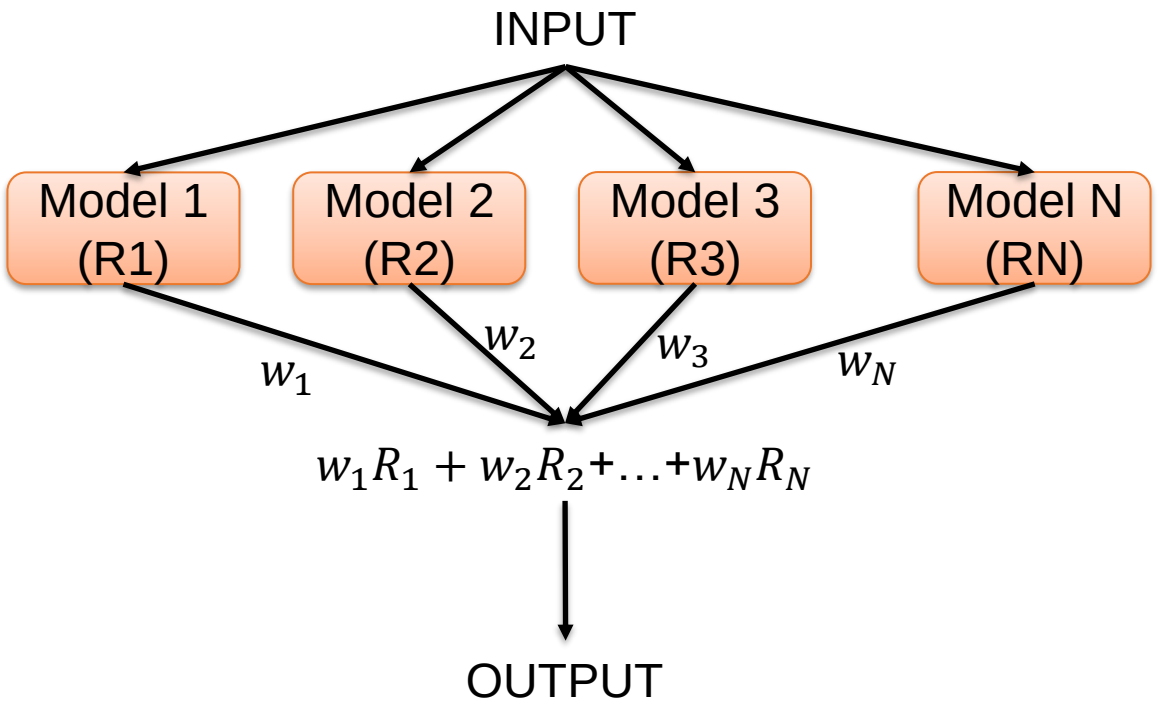
Kỹ thuật Averaging

- 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



NỘI DUNG

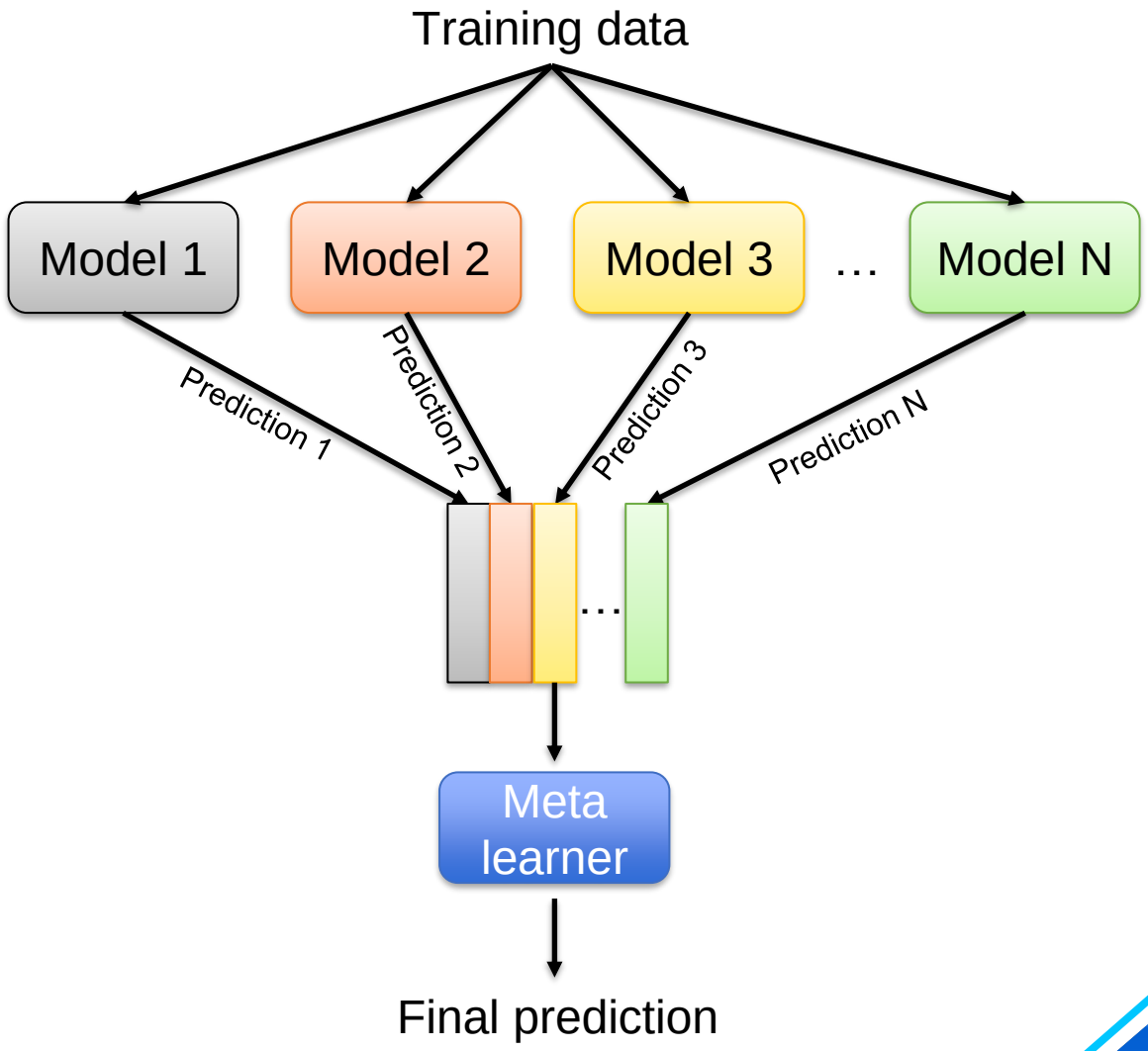
1. TẠI SAO CẦN CÓ ENSEMBLE MODEL

2. KỸ THUẬT CƠ BẢN: VOTING, AVERAGING

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

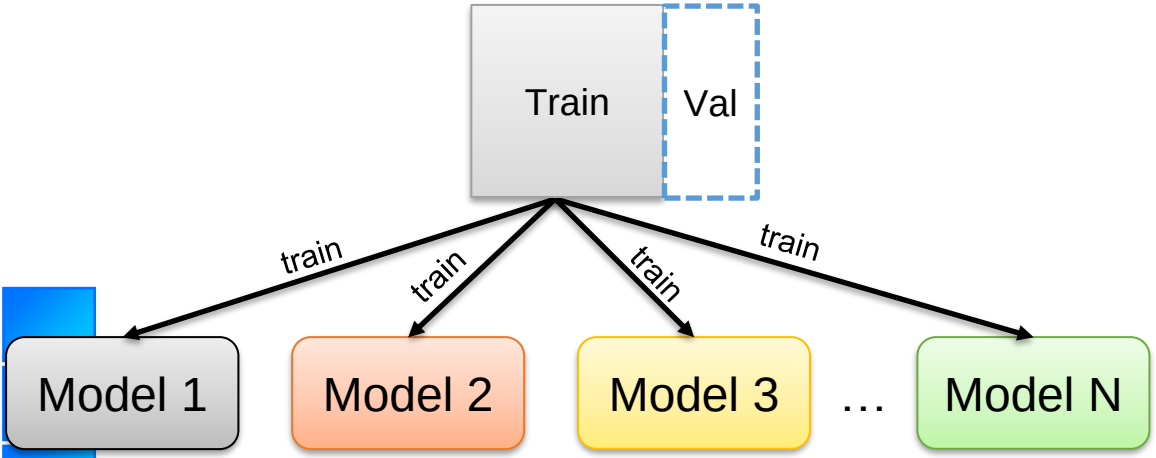
Kỹ thuật Stacking

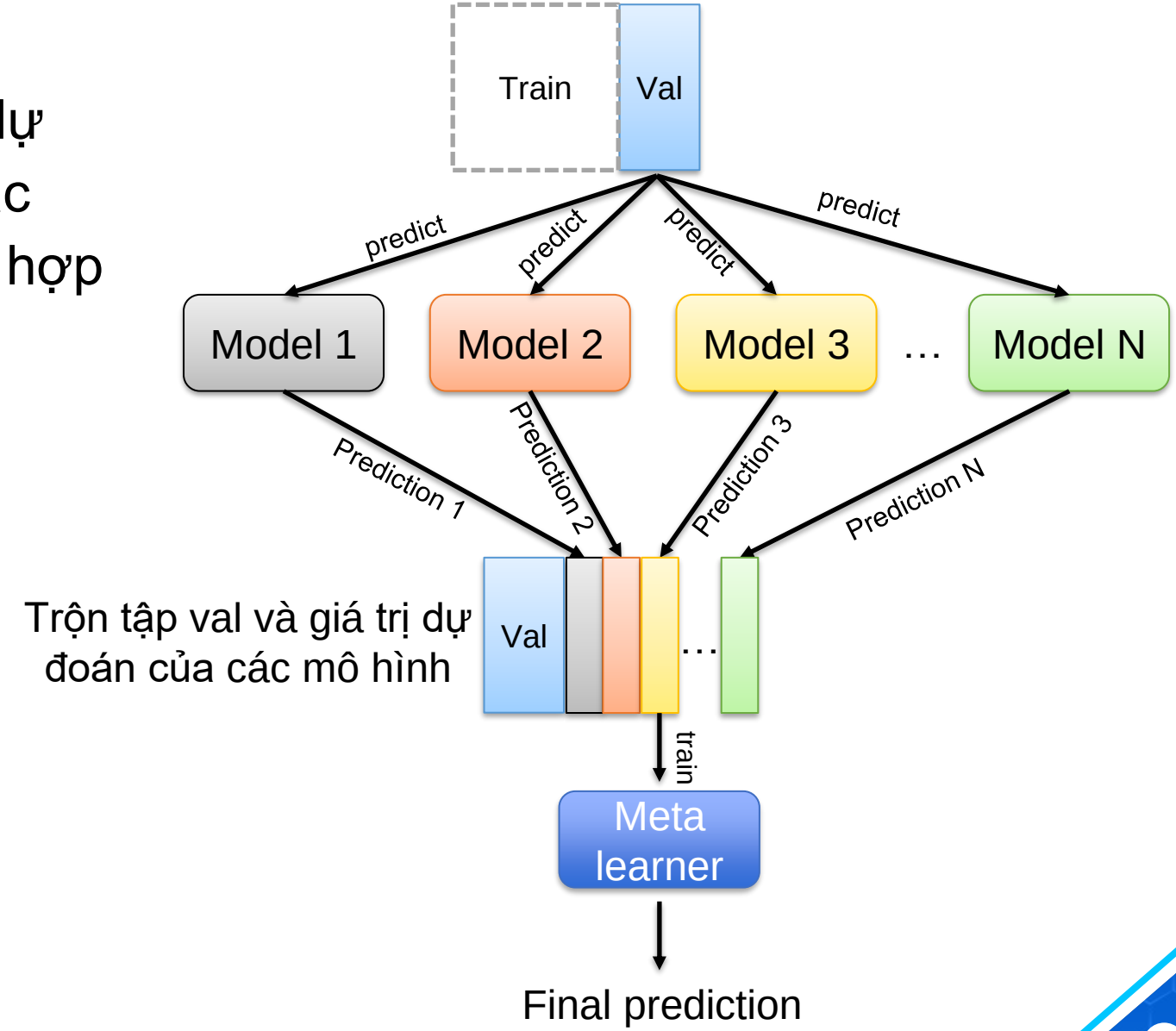
- Sử dụng kết quả dự đoán của tập train làm đặc trưng để huấn luyện mô hình tổng hợp (meta learner)



Kỹ thuật Blending

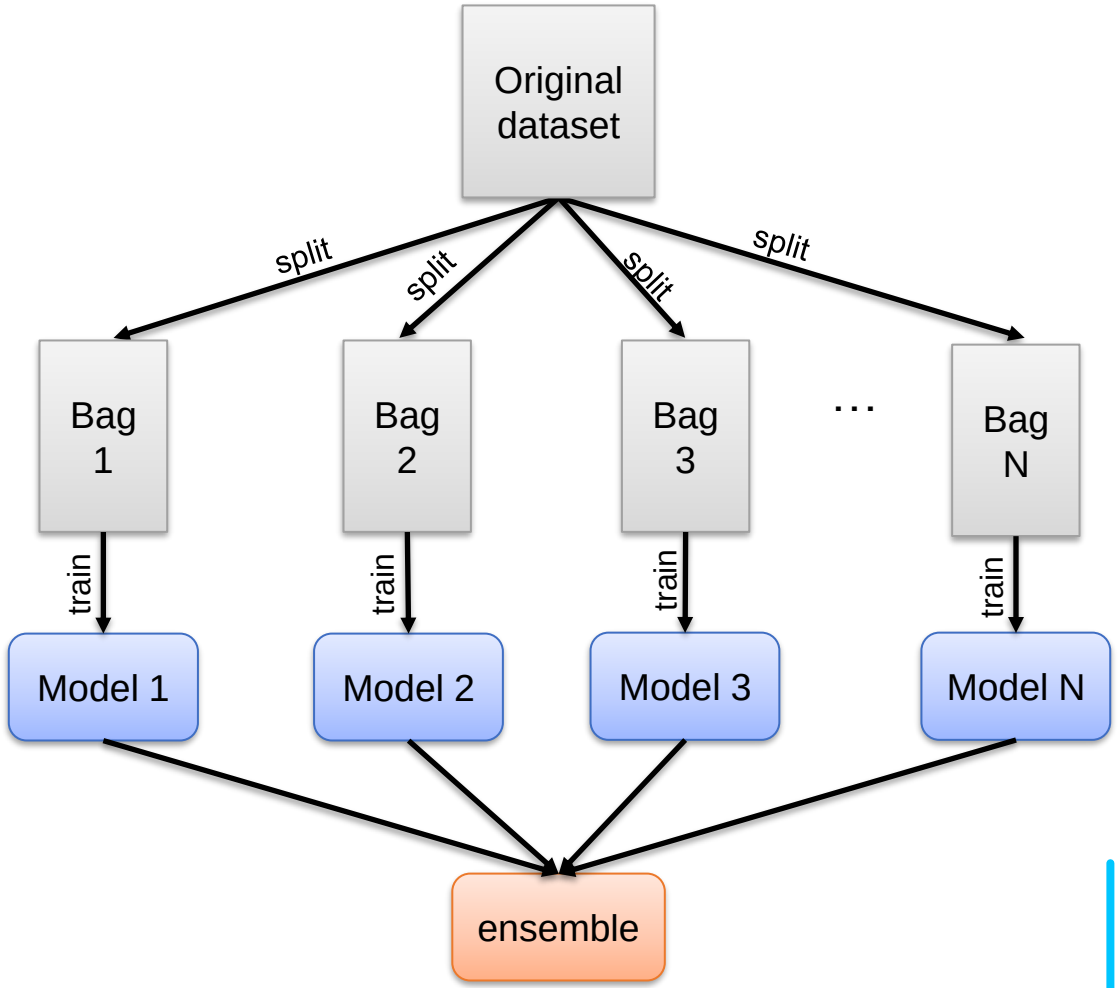
- Sử dụng (đặc trưng + kết quả dự đoán của tập validation) làm đặc trưng huấn luyện mô hình tổng hợp





Kỹ thuật Bagging

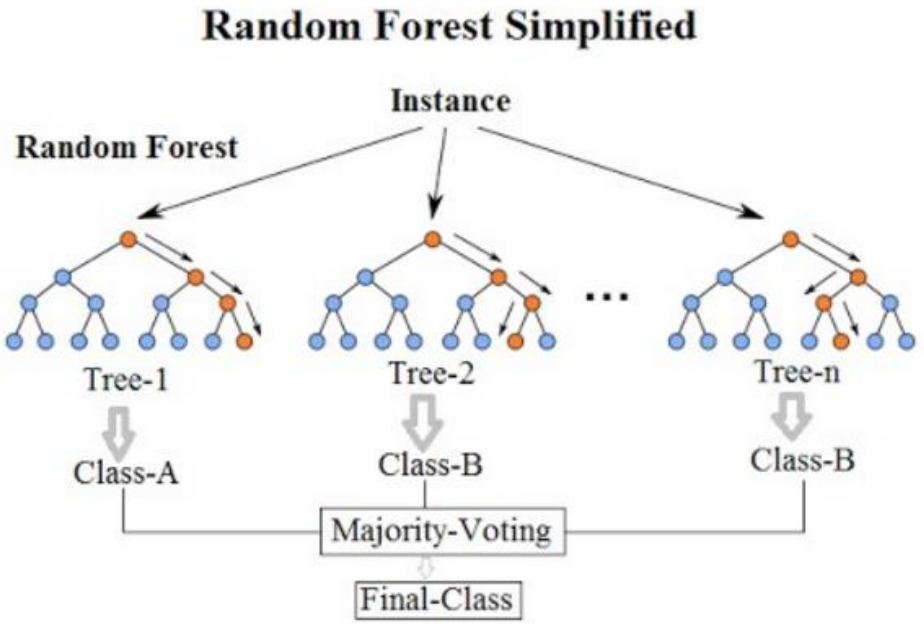
- Khác với 2 thuật toán trước, Bagging sử dụng cùng một thuật toán cho tất cả mô hình con
- Bagging huấn luyện độc lập các mô hình con, trên các tập con của dataset



Kỹ thuật Bagging

- Điển hình của kỹ thuật này là Random Forest

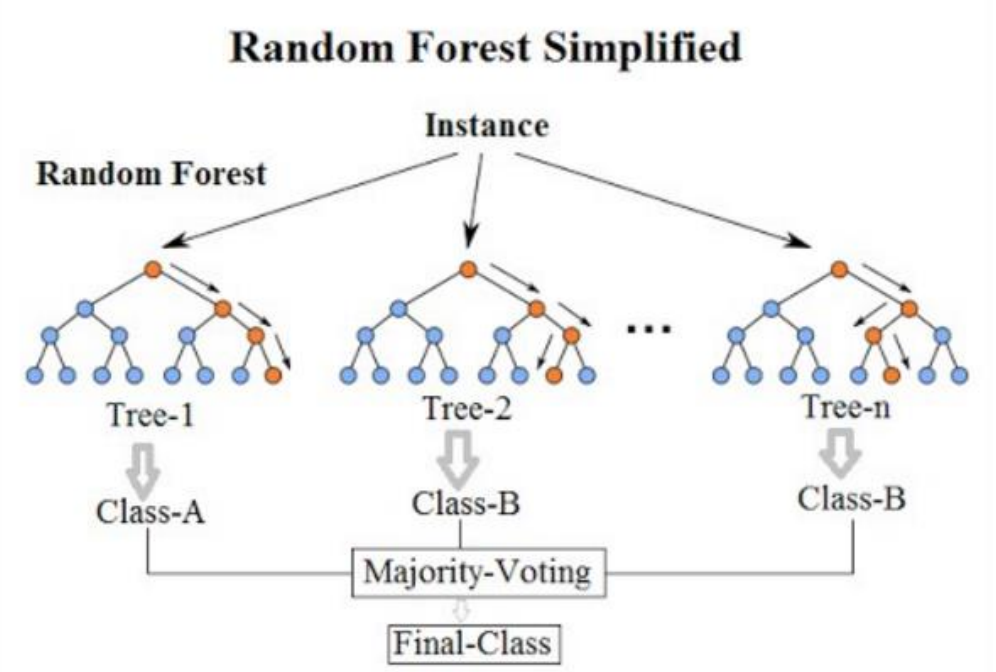
- Ưu điểm:
  - **Hiệu quả**: chính xác, tổng quát hóa cao
  - **Tiện lợi**: có thể thực hiện được trên số đặc trưng lớn mà không cần phân tích đặc trưng
  - **Linh hoạt**: dùng cho cả hồi quy và phân lớp
  - Có thể song song hóa thuật toán
  - **Bền vững**: ít bị ảnh hưởng bởi outlier, ít khả năng overfitting



Kỹ thuật Bagging

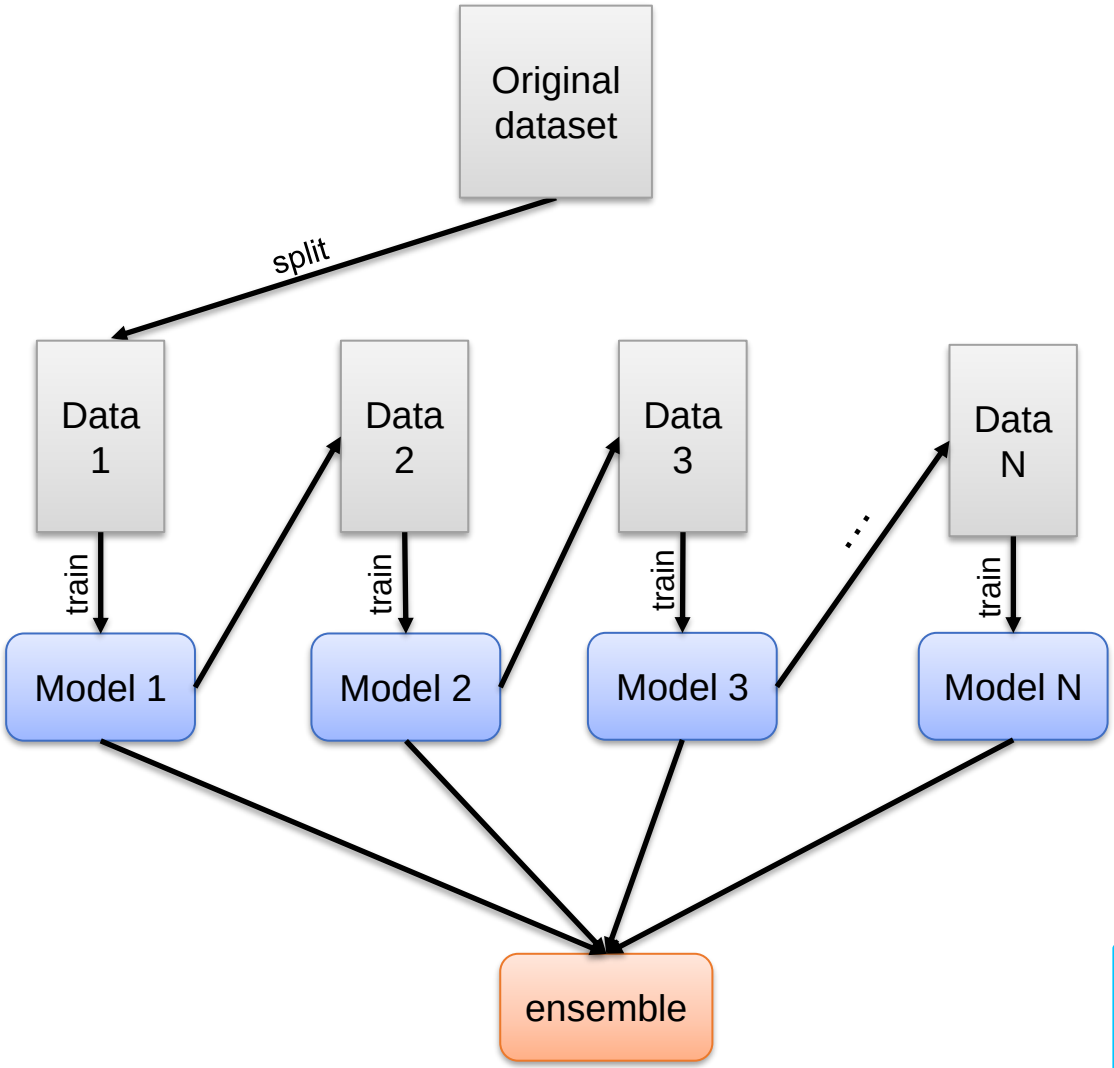
- Một số khuyết điểm:
  - Khó giải thích mô hình
  - Độ phức tạp tính toán cao
  - Bias với dữ liệu không cân bằng

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



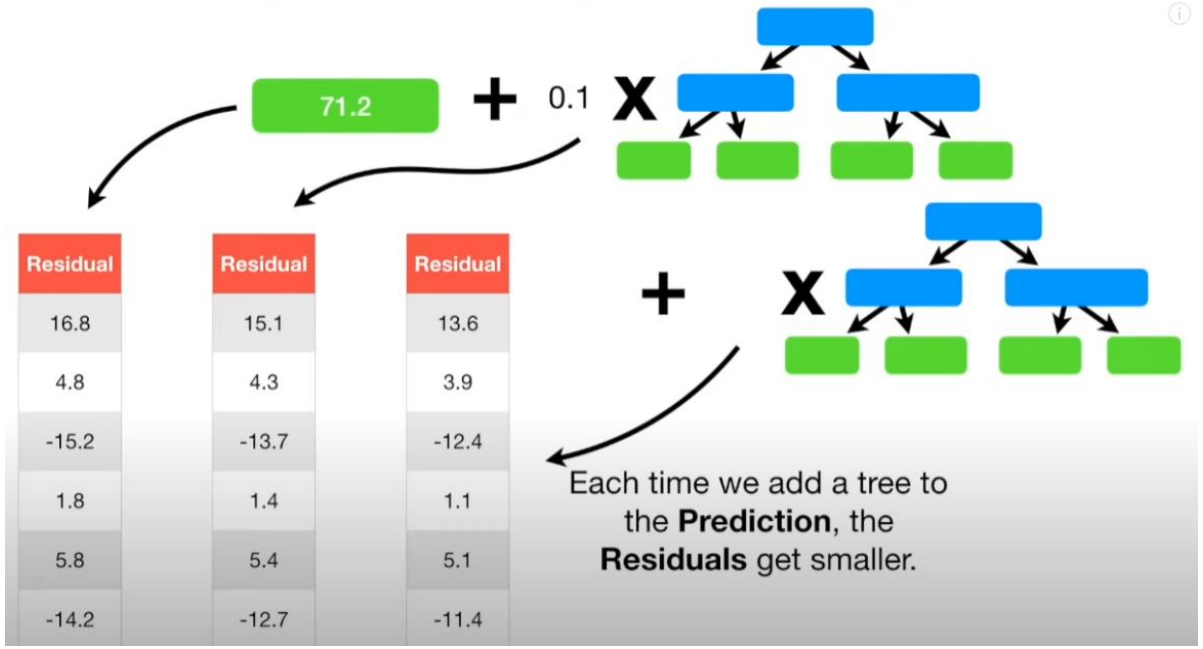
Kỹ thuật Boosting

- 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

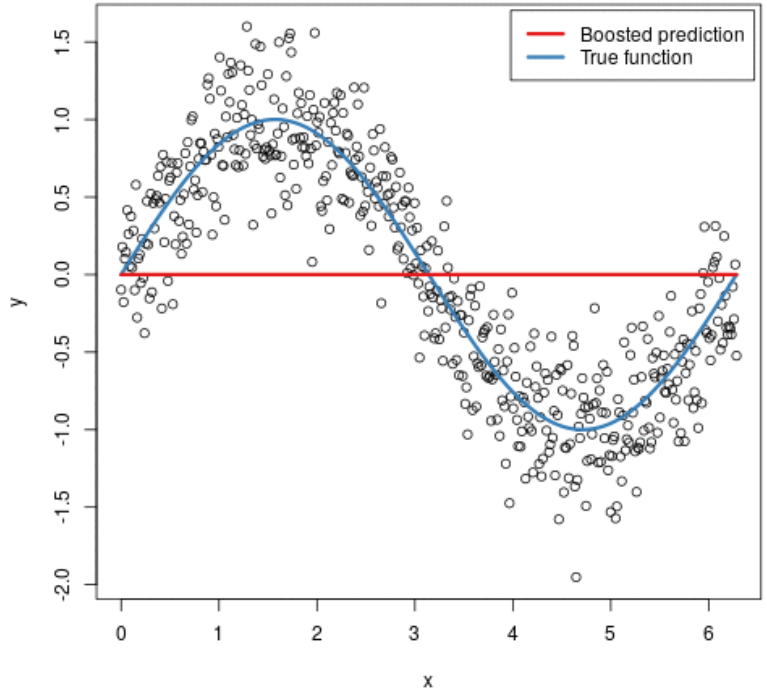


Kỹ thuật Boosting – Gradient Boost

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



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



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

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
- 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