HUST

ĐẠI HỌC BÁCH KHOA HÀ NỘI

HANOI UNIVERSITY OF SCIENCE AND TECHNOLOGY

ONE LOVE. ONE FUTURE.

ĐẠI HỌC
BÁCH KHOA HÀ NỘI
HANOI UNIVERSITY
OF SCIENCE AND TECHNOLOGY

Ensemble Learning

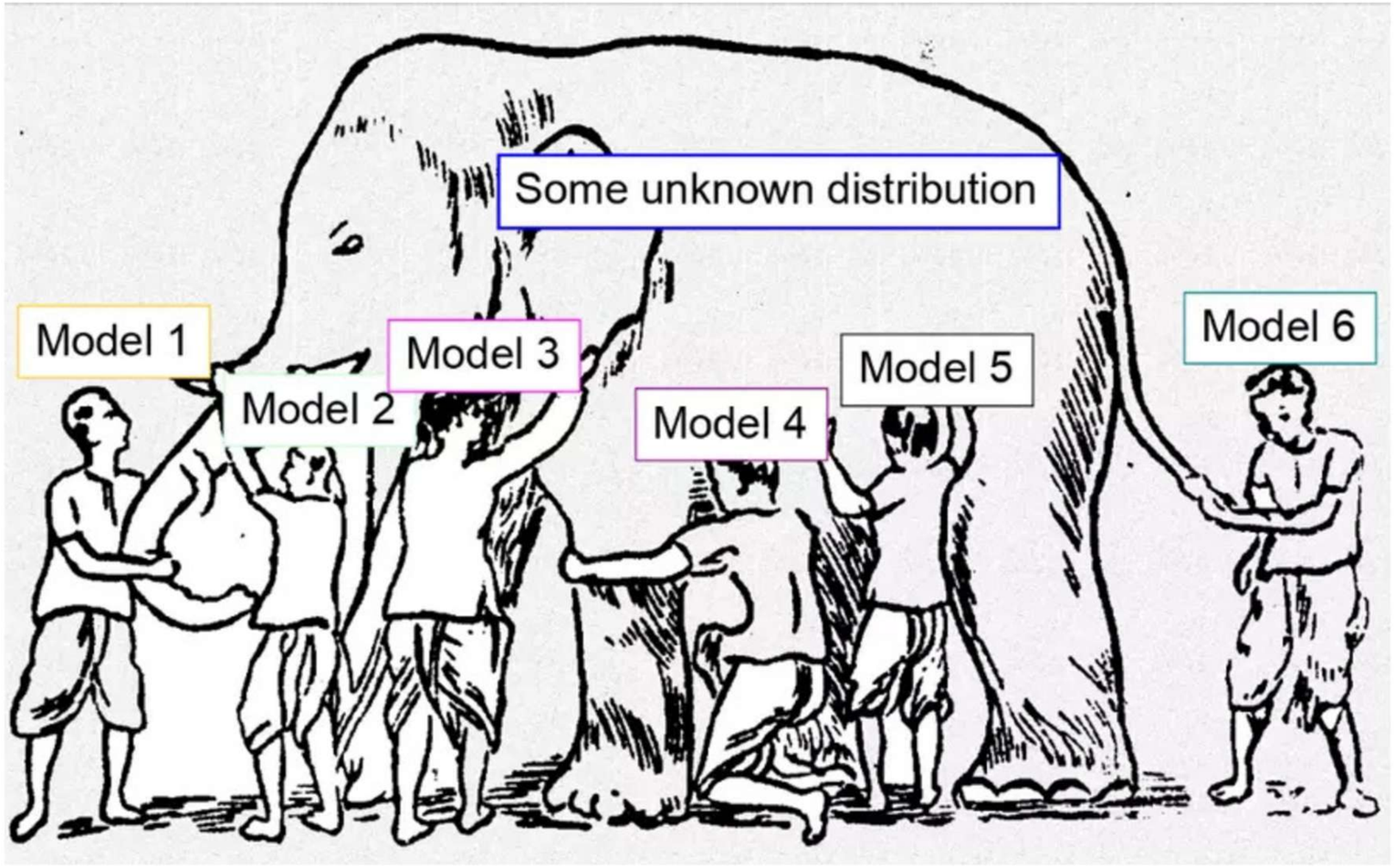
ONE LOVE. ONE FUTURE.

Nội dung chính

- Ensemble Learning
  - Bagging
  - Boosting
  - Stacking
- Ứng dụng trong bài toán thực tế

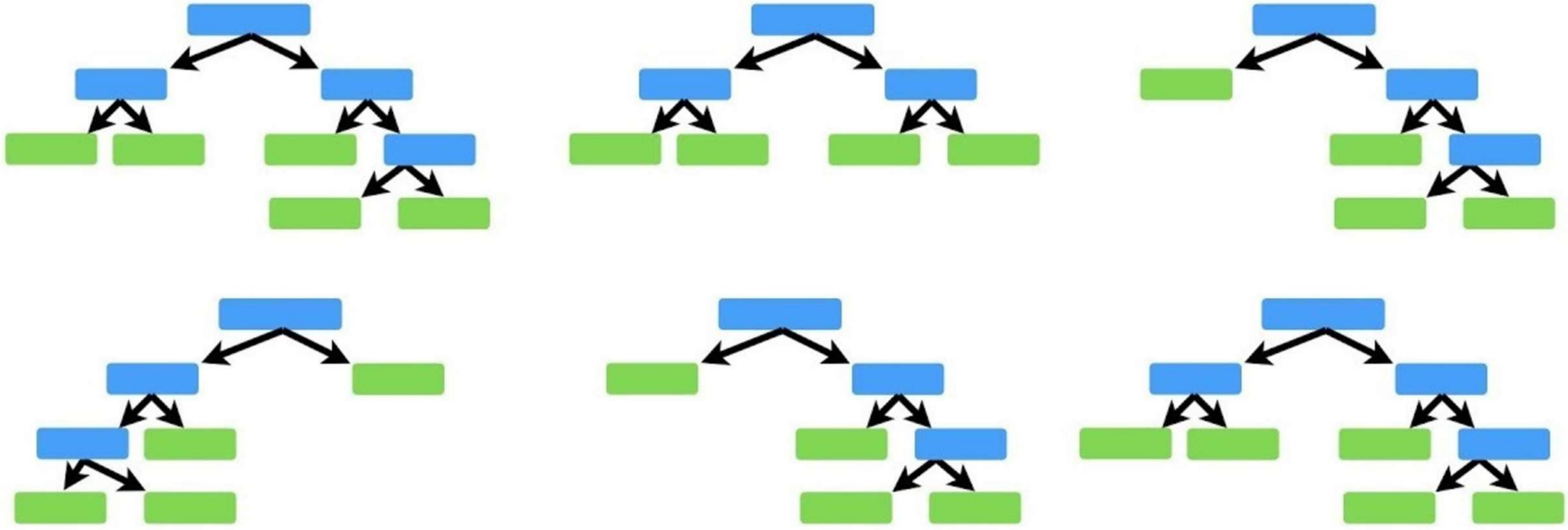
Ensemble Learning

- Thay vì cố gắng xây dựng một mô hình tốt duy nhất, mục tiêu trở thành xây dựng một họ các mô hình yếu hơn một chút, nhưng khi kết hợp các mô hình lại, sẽ thu được một mô hình còn vượt trội hơn cả



Bagging

- Bagging là lấy một lượng lớn các model cùng loại trên những tập con khác nhau của dữ liệu và huấn luyện chúng độc lập, song song với nhau. Kết quả cuối cùng sẽ là tổng hợp của các mô hình

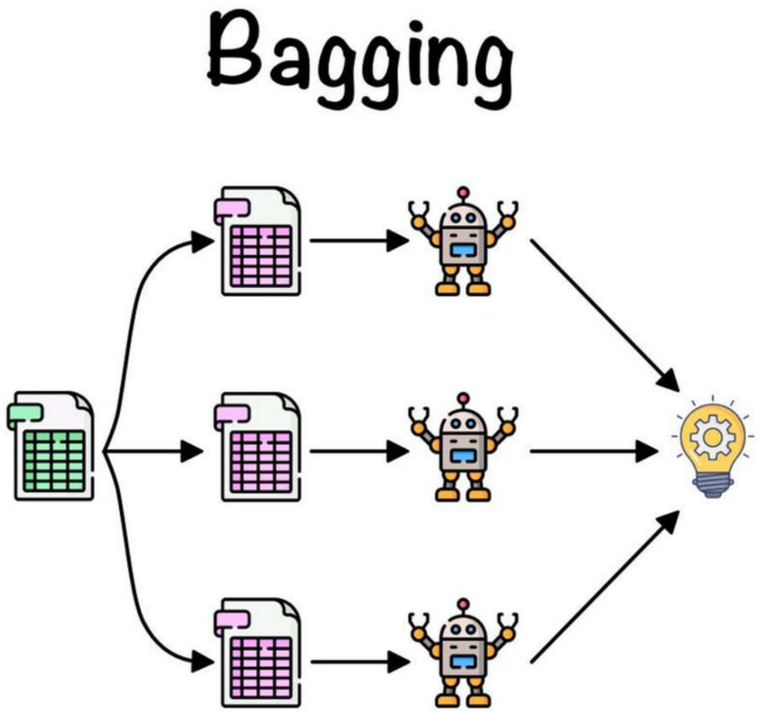


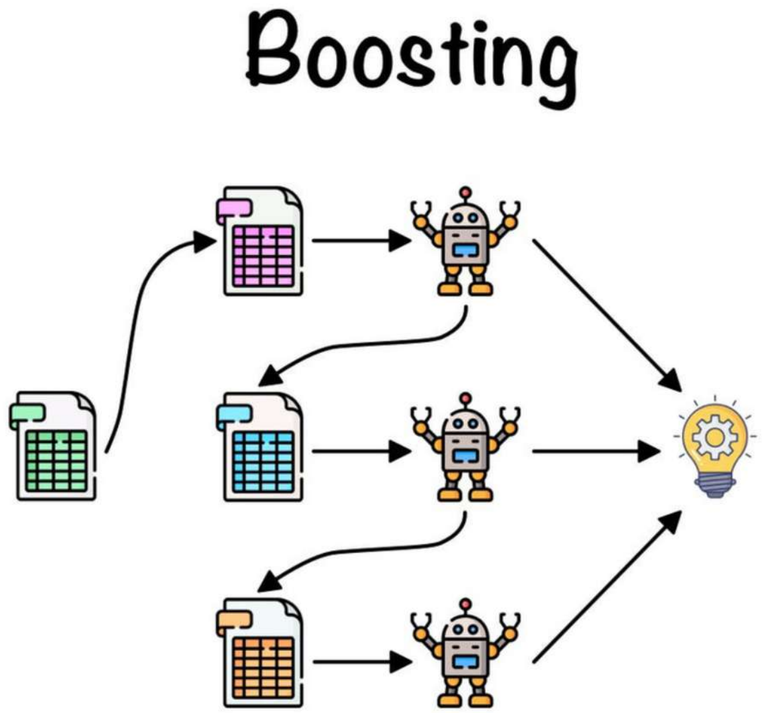
Bagging

- Ưu điểm của Bagging:
  - Dùng trong cả bài toán phân loại lẫn hồi quy
  - Giảm độ nhạy với nhiễu, chống overfitting
  - Các mô hình được huấn luyện song song, tận dụng tốt phần cứng
- Nhược điểm của Bagging:
  - Chi phí tính toán và bộ nhớ lớn hơn một mô hình riêng lẻ
  - Khó giải thích
  - Các mô hình học độc lập với nhau, không học được gì từ nhau

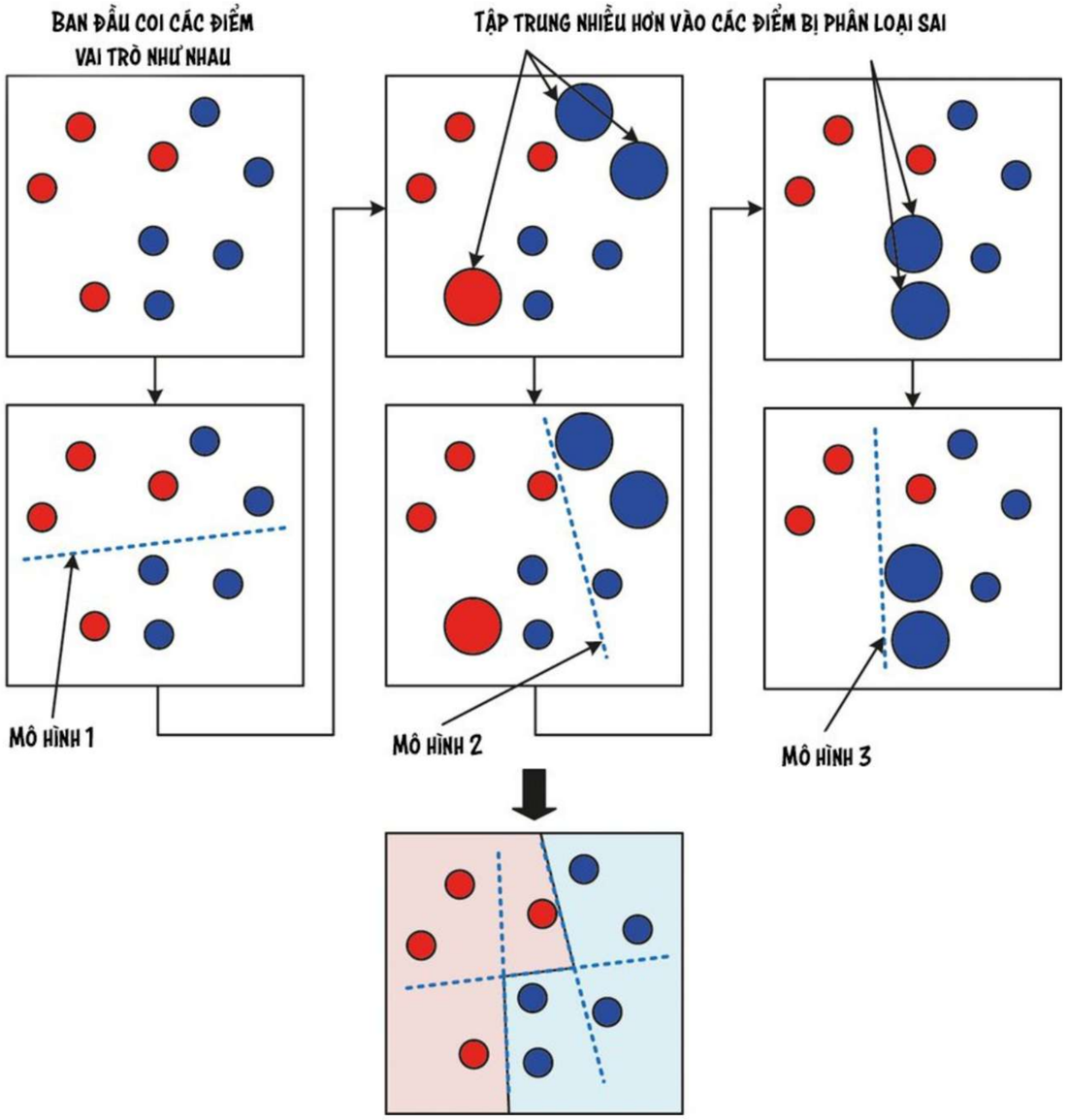
Boosting

- Boosting sẽ tạo ra một loạt các model yếu, học bổ sung lẫn nhau, các model sau sẽ cố gắng học để hạn chế lỗi lầm của các model trước





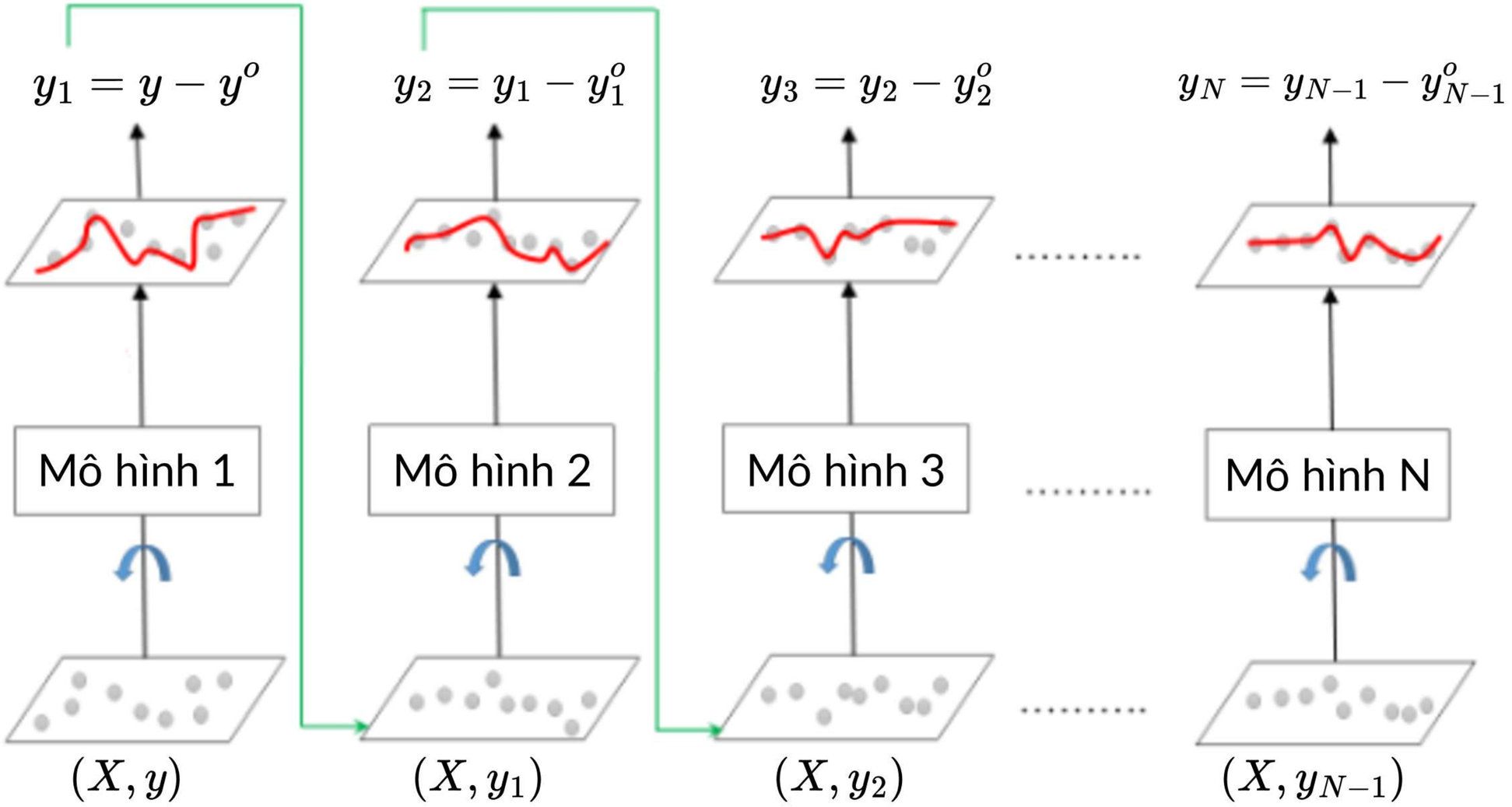
AdaBoost


BAN ĐẦU COI CÁC ĐIỂM
VAI TRÒ NHƯ NHAU
TẬP TRUNG NHIỀU HƠN VÀO CÁC ĐIỂM BỊ PHÂN LOẠI SAI
MÔ HÌNH 1
MÔ HÌNH 2
MÔ HÌNH 3

AdaBoost

- Bước 1: Khởi tạo trọng số mỗi điểm bằng nhau $w_i^1 = \frac{1}{n}; i = 1, \ldots, n$
- Bước 2:
  - Mô hình thành phần $h_t(x)$ tối thiểu hóa hàm mất mát là tổng trọng số các điểm phân loại sai $E_t = \sum_{y_i \neq h_t(x_i)} w_i^t$
  - Gán trọng số cho mô hình thành phần $\alpha_t = \frac{1}{2} ln\left(\frac{1 - E_t}{E_t}\right)$
  - Cập nhật mô hình cuối cùng $H_t(x) = H_{t-1}(x) + \eta \alpha_t h_t(x)$
  - Cập nhật trọng số mỗi điểm $w_i^{t+1} = w_i^t e^{-y_i \alpha_t h_t(x_i)}$
- Bước 3: Mô hình cuối cùng thu được từ tổng các mô hình thành phần

Gradient Boosting

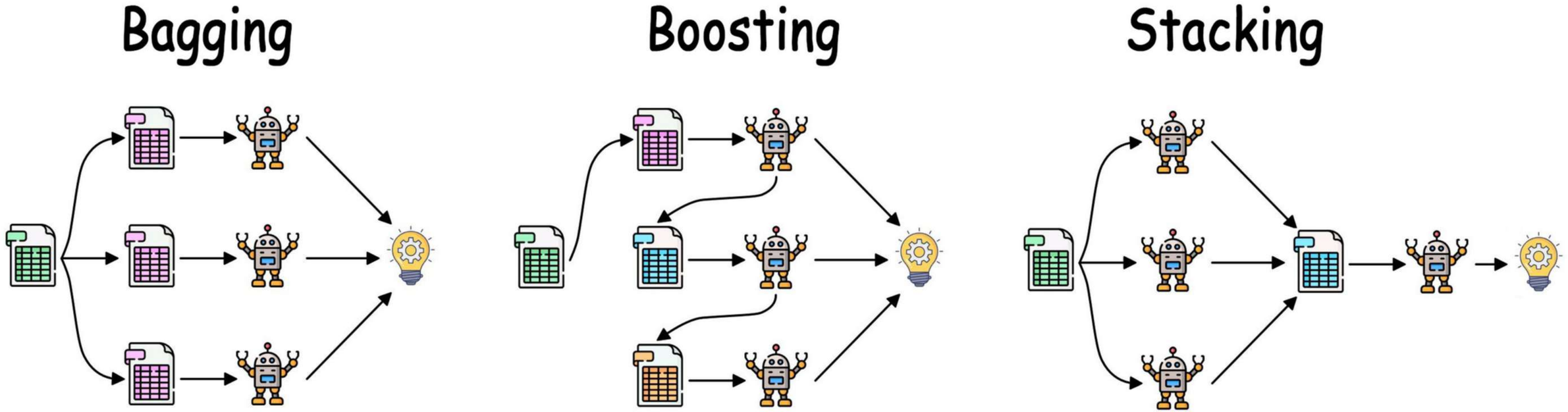


Boosting

- Các thuật toán Boosting hiện đại còn có XGBoost, CatBoost và LightGBM, phổ biến trong nhiều bài toán
- Ưu điểm của Boosting:
  - Độ chính xác tốt, là baseline của nhiều bài toán
  - Tự động tập trung vào sai số
  - Có nhiều biến thể tùy theo hàm mất mát mong muốn
- Nhược điểm của Boosting:
  - Dễ bị overfit do học kỹ sai số
  - Tốn thời gian huấn luyện do phải huấn luyện tuần tự
  - Khó giải thích

Stacking

- Stacking là lấy một lượng các model khác loại huấn luyện trên cả tập huấn luyện và sử dụng duy nhất một mô hình để đưa ra dự đoán tốt nhất từ các dự đoán của các mô hình khác



Stacking

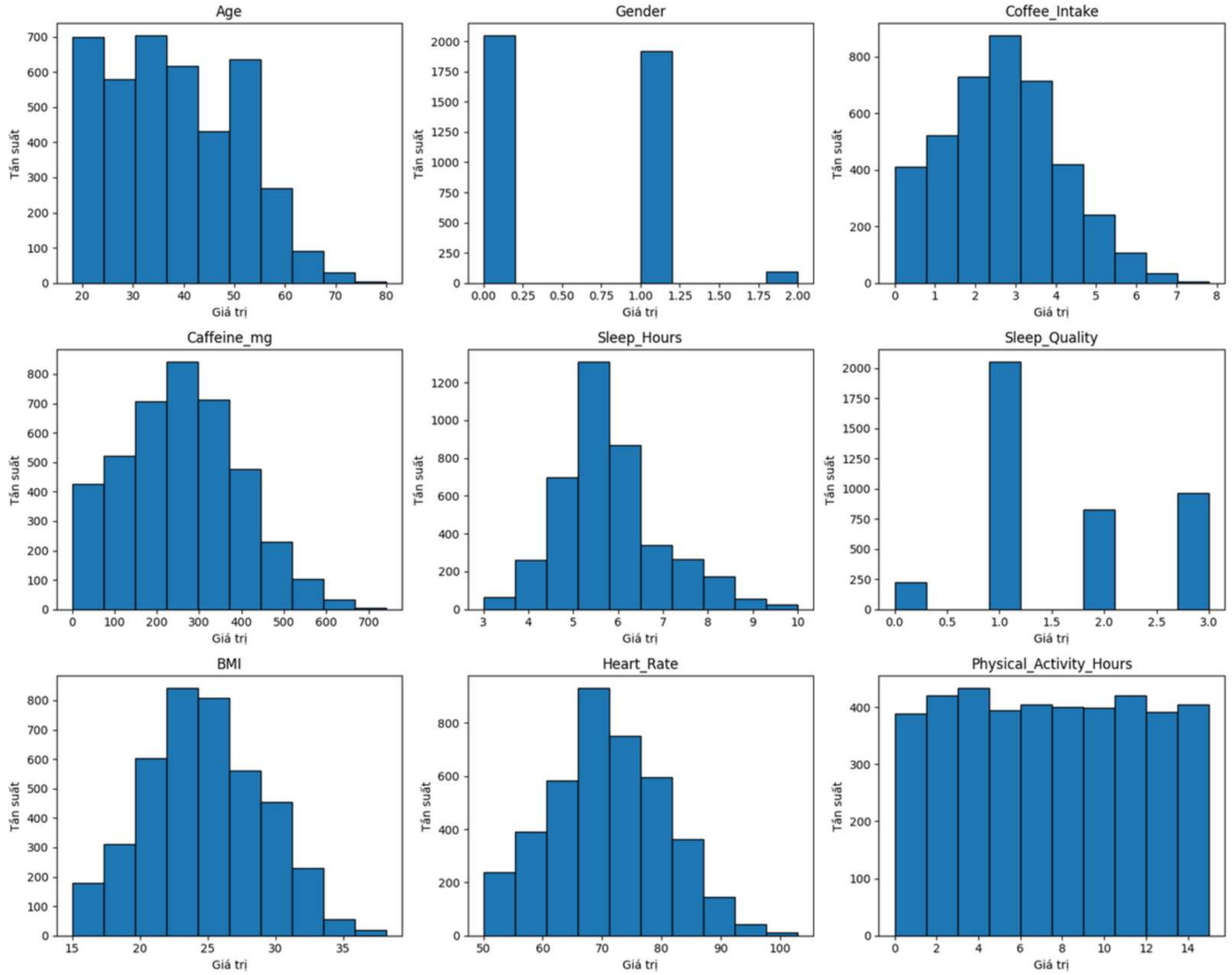
- Một vài biến thể
  - Out-of-fold Stacking: Stacking kết hợp Cross-Validation để giảm khả năng bị overfit
  - Multi-level Stacking: Thay vì chỉ 2 tầng thì nâng thành nhiều tầng (Level-0 đến Level-N)
  - Blending: Chỉ tạo một validation set cố định thay vì dùng Cross-Validation

Stacking

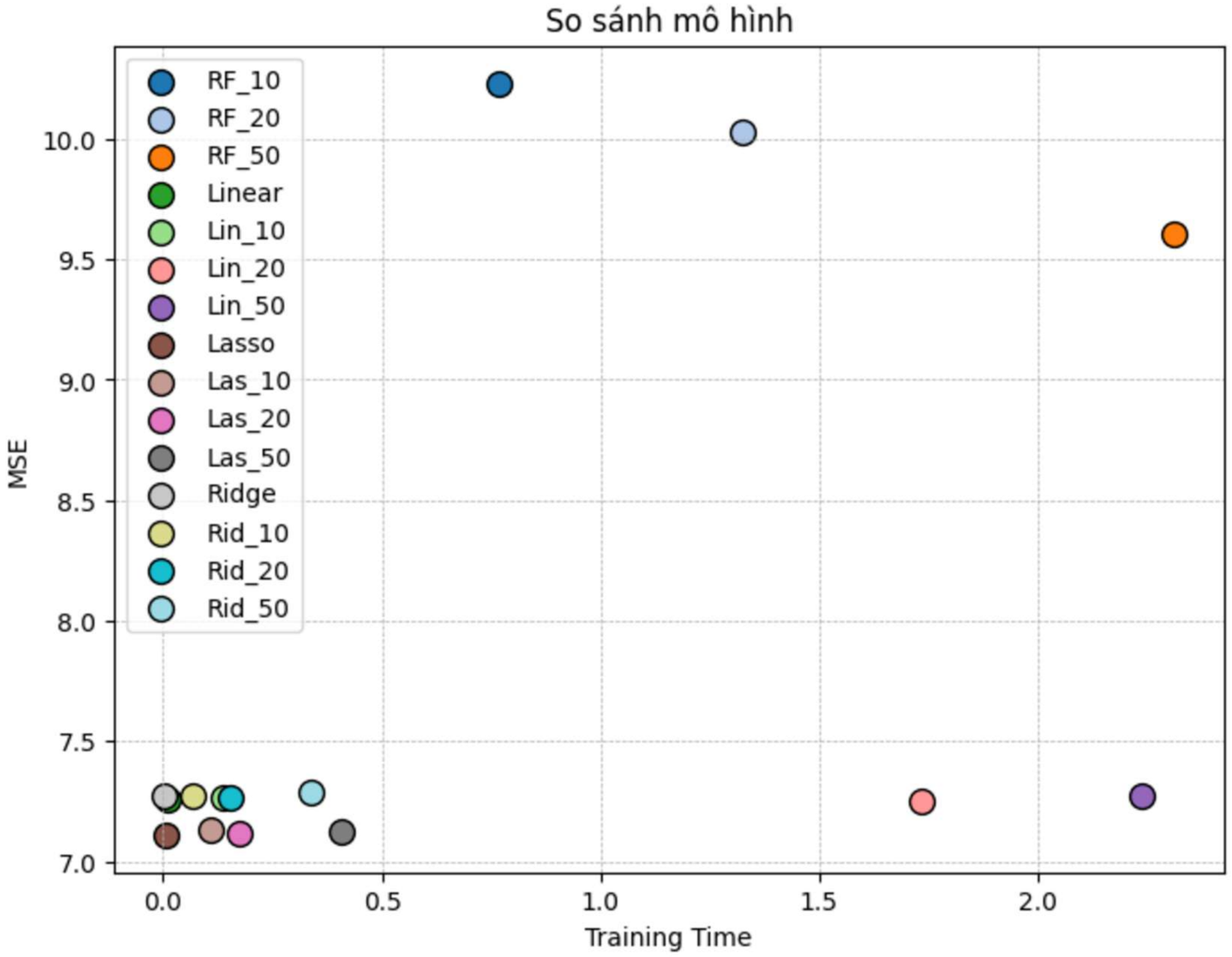
- Ưu điểm của Stacking:
  - Tận dụng tốt điểm mạnh mỗi mô hình, vượt trội hơn Bagging và Boosting trong nhiều trường hợp
  - Linh hoạt trong việc chọn mô hình thành phần
  - Có thể mở rộng nhiều tầng
- Nhược điểm của Stacking:
  - Có nguy cơ overfit cao
  - Chi phí huấn luyện cao
  - Khó diễn giải
  - Phụ thuộc nhiều vào việc lựa chọn mô hình thành phần

Ứng dụng trong bài toán thực tế

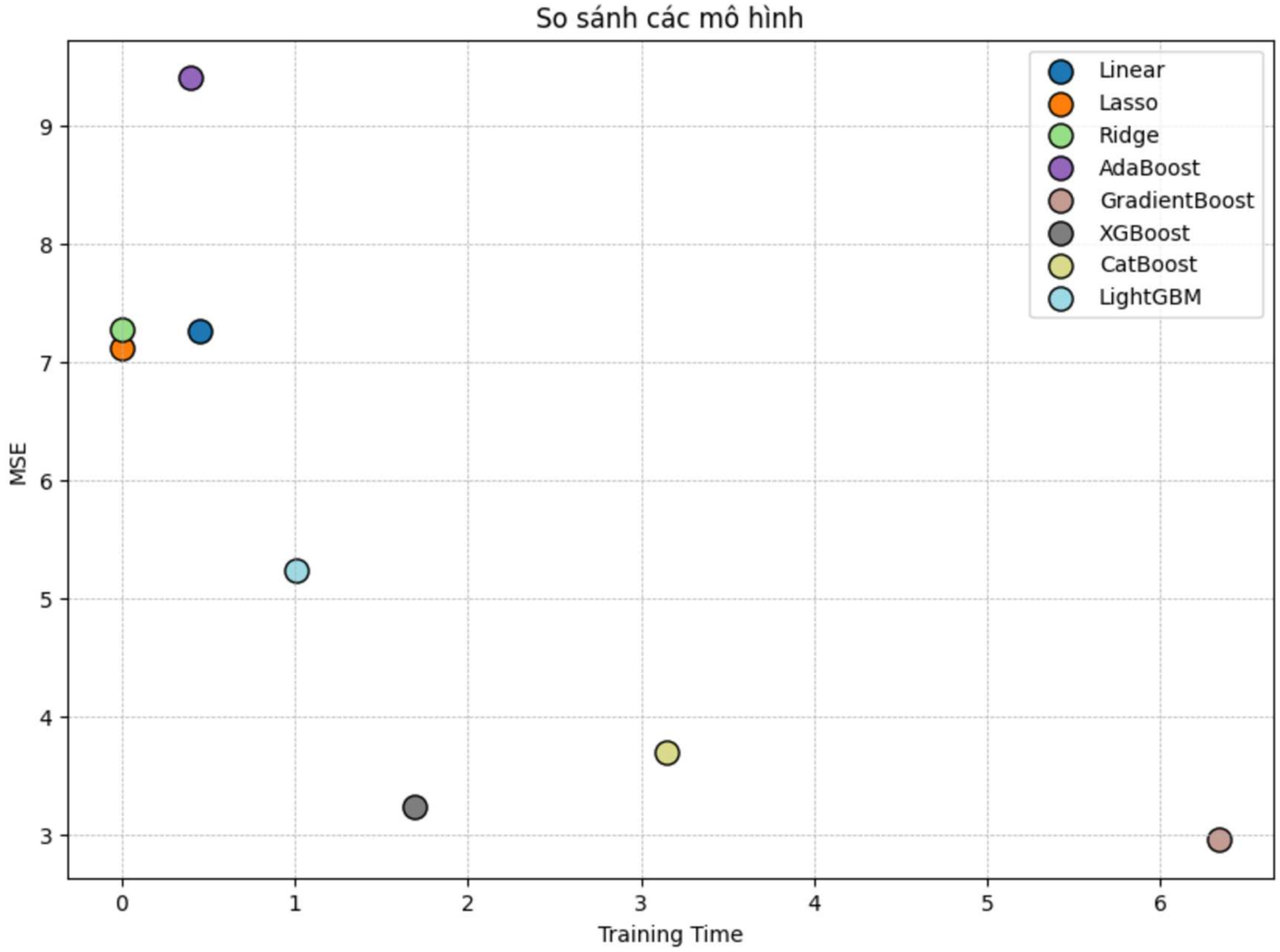
- Bộ dữ liệu: Global Coffee Health Dataset
  - 4000 mẫu
  - Gồm các đặc trưng: tuổi, giới tính người uống, lượng cà phê uống, nhịp tim,...



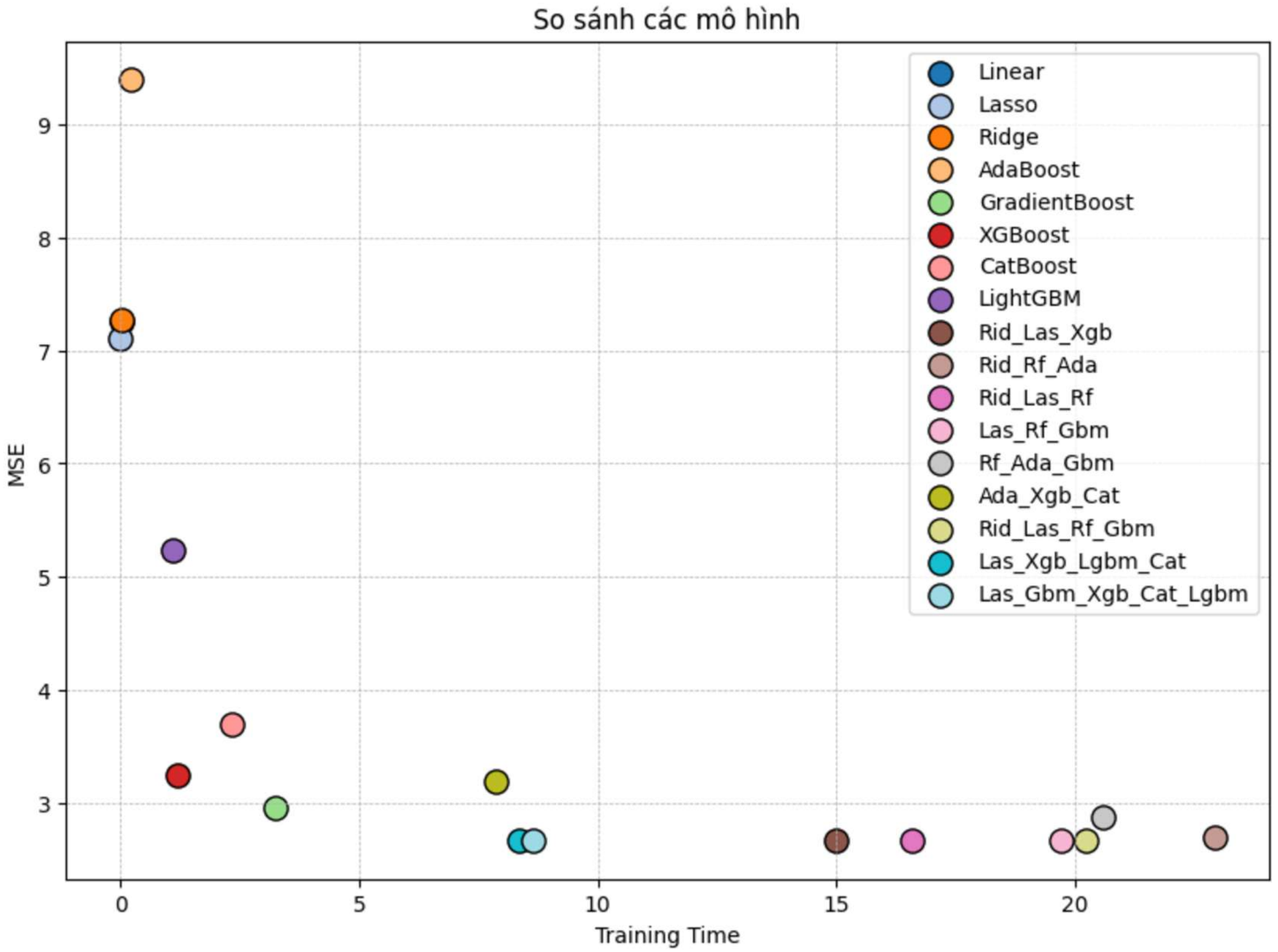
Ứng dụng trong bài toán thực tế



Ứng dụng trong bài toán thực tế



Ứng dụng trong bài toán thực tế



HUST

hust.edu.vn fb.com/dhbkhn

THANK YOU !