



BÀI TẬP LỚN

Đề tài:

DỰ ĐOÁN RỦI RO TÍN DỤNG

Sinh viên thực hiện: PHẠM THỊ LINH CHI

Mã sinh viên: 12423005

Giảng viên hướng dẫn: PGS. TS. NGUYỄN VĂN HẬU

NỘI DUNG

- ▶ 1. Lý do chọn đề tài
- ▶ 2. Giới thiệu dataset
- ▶ 3. Tiền xử lý dữ liệu
- ▶ 4. Các mô hình học máy áp dụng
- ▶ 5. Kết quả
- ▶ 6. Kết luận

NỘI DUNG

- ▶ 1. Lý do chọn đề tài
- ▶ 2. Giới thiệu dataset
- ▶ 3. Tiền xử lý dữ liệu
- ▶ 4. Các mô hình học máy áp dụng
- ▶ 5. Kết quả
- ▶ 6. Kết luận

I. LÝ DO CHỌN ĐỀ TÀI

Lý do chọn đề tài:

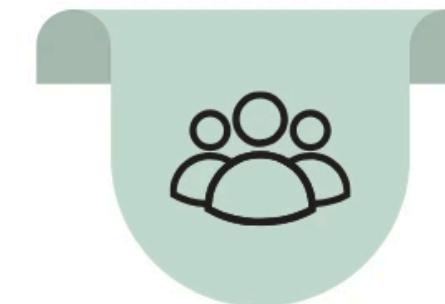
- Rủi ro tín dụng chiếm tỷ trọng lớn trong tổng rủi ro của các ngân hàng thương mại.
- Bỏ sót khách hàng rủi ro gây thiệt hại lớn và làm gia tăng nợ xấu
- Học máy giúp tự động hóa đánh giá rủi ro, nhanh và nhạy hơn phương pháp truyền thống



I. LÝ DO CHỌN ĐỀ TÀI



Nguyên nhân dẫn đến rủi ro tín dụng



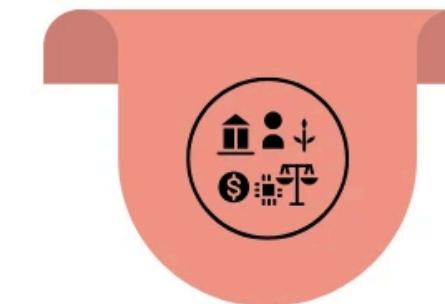
TỪ PHÍA KHÁCH HÀNG

- Khả năng tài chính yếu kém
- Suy giảm uy tín
- Hành vi cố ý không trả nợ



TỪ PHÍA NGÂN HÀNG

- Đánh giá tín dụng không chính xác
- Tập trung tín dụng quá mức
- Quản trị yếu kém
- Chính sách tín dụng không phù hợp



YẾU TỐ BÊN NGOÀI

- Biến động kinh tế vĩ mô
- Chính sách pháp luật thay đổi
- Rủi ro thị trường

NỘI DUNG

- ▶ 1. Lý do chọn đề tài
- ▶ **2. Giới thiệu dataset**
- ▶ 3. Tiền xử lý dữ liệu
- ▶ 4. Các mô hình học máy áp dụng
- ▶ 5. Kết quả
- ▶ 6. Kết luận

II. GIỚI THIỆU DATASET

Dataset: credit_risk_dataset.csv

person_age	person_income	person_home_ownership	person_emp_length	loan_intent	loan_grade	loan_amnt	loan_int_rate	loan_status	loan_percent_income	cb_person_default_on_file	cb_person_cred_hist_length
22	59000	RENT	123	PERSONAL	D	35000	1.602	1	59	Y	3
21	9600	OWN	5	EDUCATION	B	1000	1.114	0	1	N	2
25	9600	MORTGAGE	1	MEDICAL	C	5500	1.287	1	57	N	3
23	65500	RENT	4	MEDICAL	C	35000	1.523	1	53	N	2
24	54400	RENT	8	MEDICAL	C	35000	1.427	1	55	Y	4
21	9900	OWN	2	VENTURE	A	2500	714	1	25	N	2

II. GIỚI THIỆU DATASET

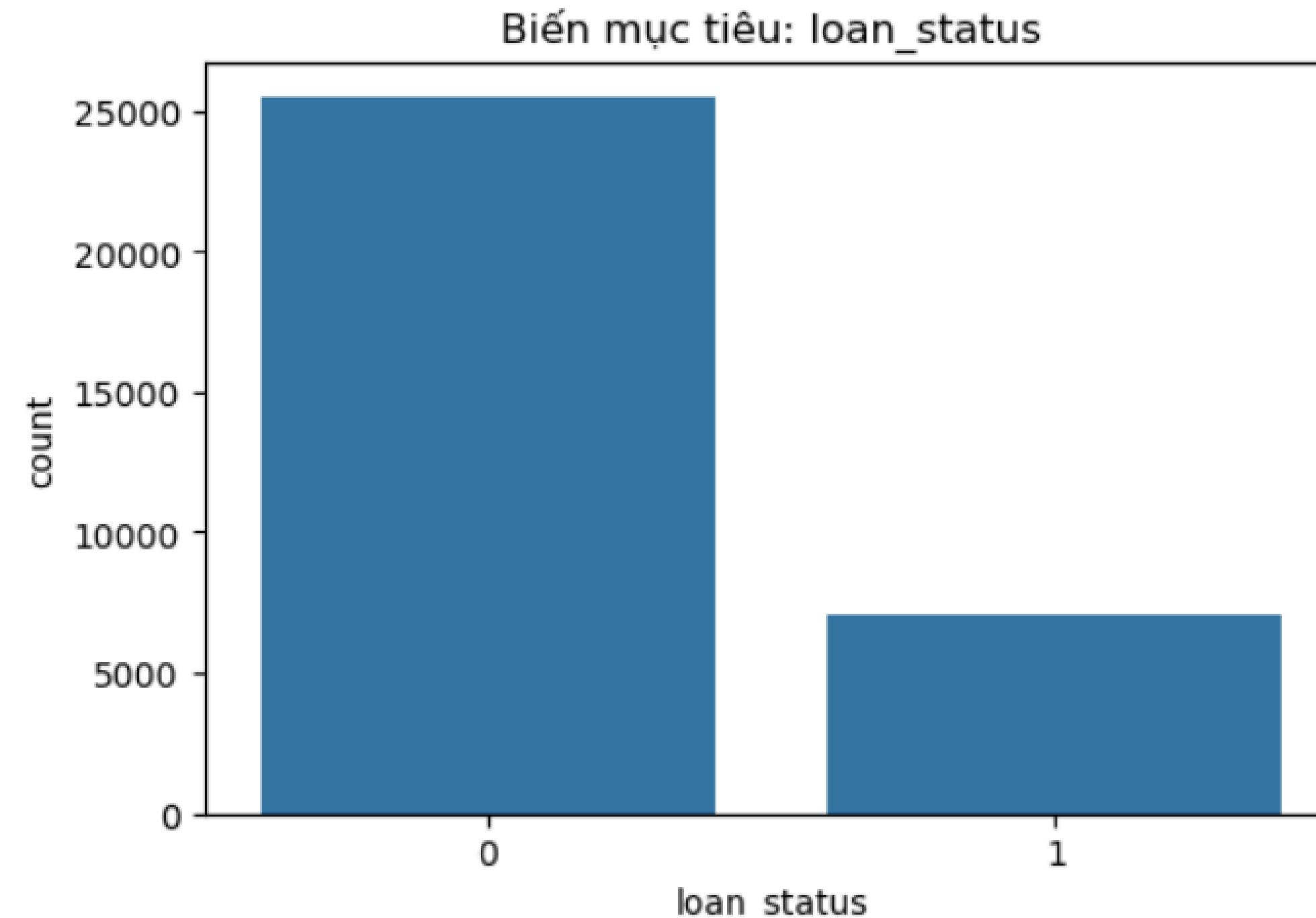
Bộ dữ liệu sử dụng các đặc trưng:

- person_age: Tuổi của khách hàng
- person_income: Thu nhập hàng năm (USD)
- person_home_ownership: Hình thức sở hữu nhà (RENT/OWN/MORTGAGE...)
- person_emp_length: Số năm làm việc
- loan_intent: Mục đích vay (PERSONAL, MEDICAL, EDUCATION...)
- loan_grade: Xếp hạng tín dụng của khoản vay (A–G)
- loan_amnt: Số tiền vay
- loan_int_rate: Lãi suất vay (%)
- loan_percent_income: Tỷ lệ tiền vay / thu nhập
- cb_person_default_on_file: Từng vỡ nợ (Y/N)
- cb_person_cred_hist_length: Thời gian lịch sử tín dụng (năm)

Biến mục tiêu (Binary Target): loan_status

- 0: Khách hàng tốt
- 1: Khách hàng rủi ro (vỡ nợ)

II. GIỚI THIỆU DATASET



- Dữ liệu có sự mất cân bằng lớp.
- Tỷ lệ nhóm 1 (rủi ro cao) thấp hơn đáng kể → cần xử lý imbalance khi train model.

NỘI DUNG

- ▶ 1. Lý do chọn đề tài
- ▶ 2. Giới thiệu dataset
- ▶ **3. Tiền xử lý dữ liệu**
- ▶ 4. Các mô hình học máy áp dụng
- ▶ 5. Kết quả
- ▶ 6. Kết luận

III. TIỀN XỬ LÝ DỮ LIỆU

Xử lý dữ liệu thiếu

- Điền giá trị trung vị (median) cho các biến số liên tục

Mã hóa biến phân loại

- Mapping thủ công với biến nhị phân và có thứ tự
- One-Hot Encoding cho các biến danh mục không có thứ tự

Chuẩn hóa dữ liệu số

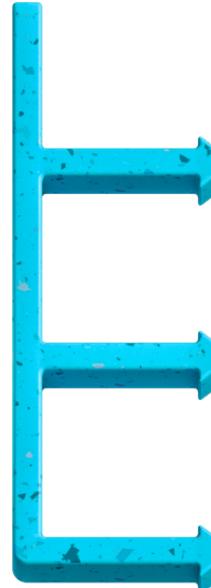
- Sử dụng Standard Scaling

Xử lý mất cân bằng lớp

- Áp dụng trọng số lớp (class weight / scale_pos_weight)

III. TIỀN XỬ LÝ DỮ LIỆU

Chia tập dữ liệu



Train: 70% (22806)

Validation: 15% (4887)

Test: 15% (4888)

III. TIỀN XỬ LÝ DỮ LIỆU

Class Weight

Ý tưởng: Tăng mức “phạt” cho lỗi dự đoán ở các lớp ít mẫu.

Công thức trọng số:

$$w_c = \frac{N}{k \cdot N_c}$$

N = tổng số mẫu

k = số lớp

N_c = số mẫu của lớp c

Áp dụng: Nhân trực tiếp vào hàm loss của lớp tương ứng:

$$\text{Loss}_{\text{total}} = \sum_{i=1}^N w_{y_i} \cdot \text{Loss}(y_i, \hat{y}_i)$$

Mục đích: Giúp mô hình học không thiên về lớp chiếm đa số.

III. TIỀN XỬ LÝ DỮ LIỆU

Scale Pos Weight

Ý tưởng: Chỉ nhân gradient của lớp thiểu số (lớp dương trong nhị phân) để tăng mức ảnh hưởng khi học boosting.

Công thức:

$$\text{scale_pos_weight} = \frac{N_{\text{negative}}}{N_{\text{positive}}}$$

Áp dụng: Gradient của mỗi điểm thuộc lớp dương được nhân với hệ số trên trong quá trình cập nhật cây.

Mục đích: Mô hình ít bỏ sót các trường hợp rủi ro (False Negative) mà không làm biến dạng dữ liệu.

III. TIỀN XỬ LÝ DỮ LIỆU

Chuẩn hóa dữ liệu: StandardScaler



Ý tưởng: Xuất phát từ thống kê.

- Trừ giá trị trung bình ($\text{tâm} = 0$).
- Chia cho độ lệch chuẩn ($\text{phương sai} = 1$).

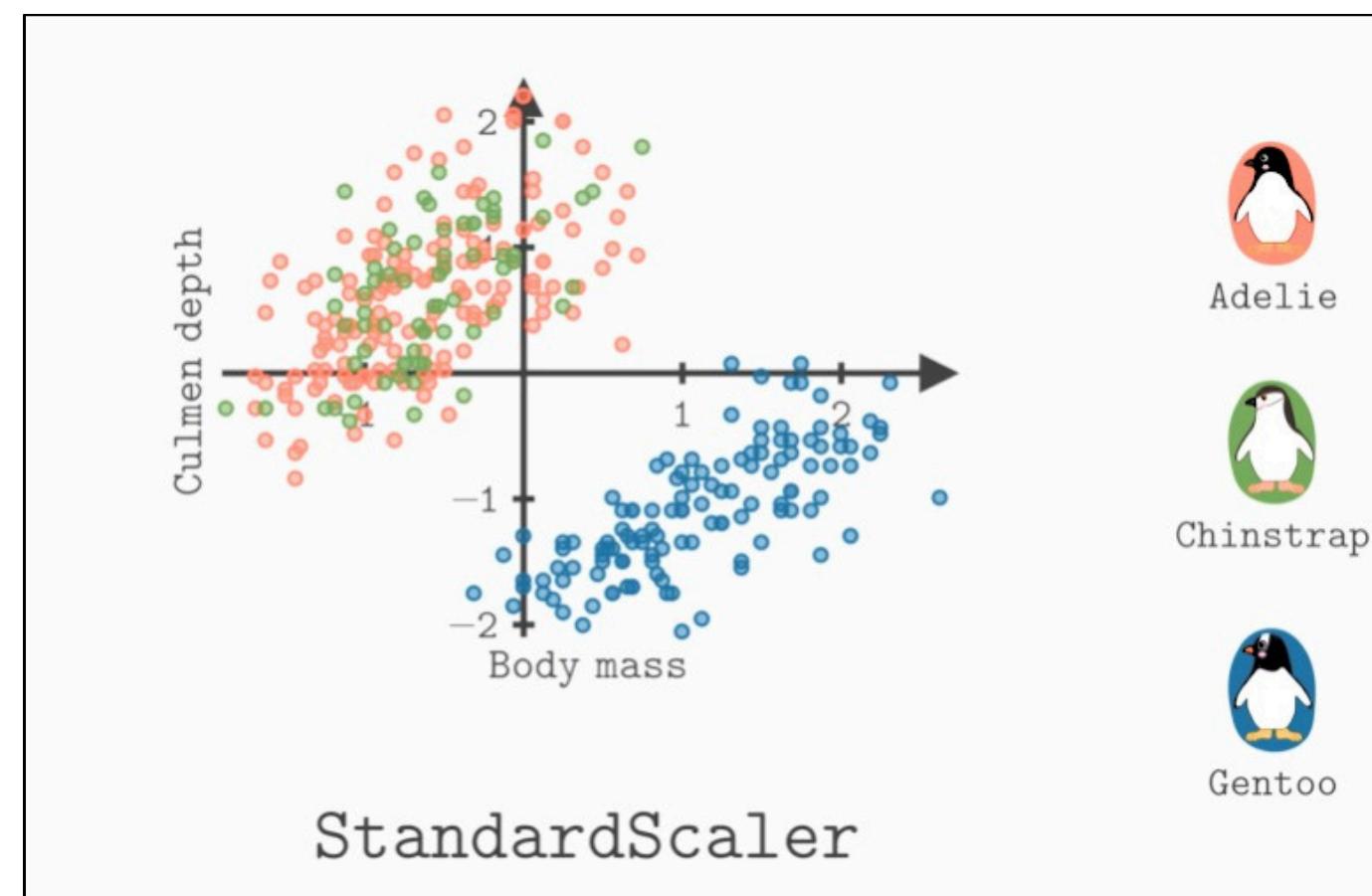
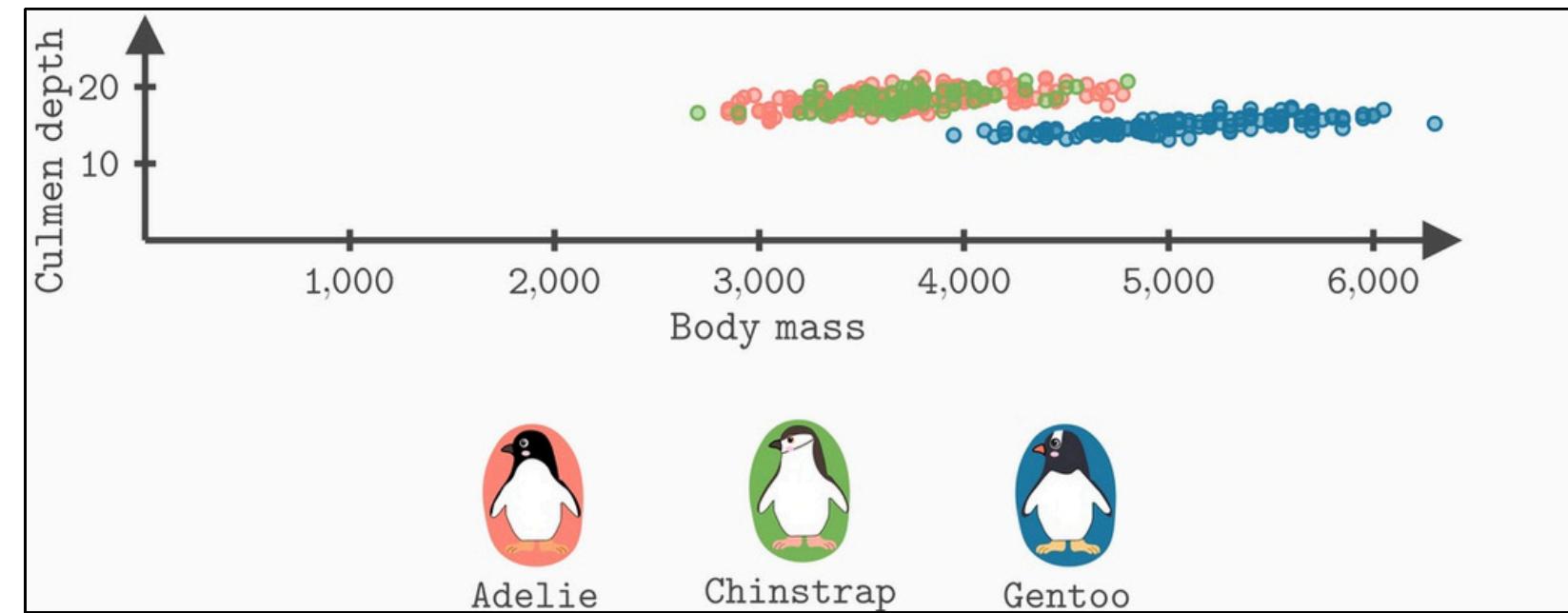
$$\text{StandardScaler : } x_{\text{scaled}} = \frac{x - \bar{x}}{\sigma}$$

Đặc điểm:

- Dữ liệu tập trung quanh 0.
- Loại bỏ ảnh hưởng của thang đo.
- Giúp so sánh các đặc trưng khác đơn vị dễ dàng hơn.

III. TIỀN XỬ LÝ DỮ LIỆU

Ví dụ: khi phân loại loài chim cánh cụt



NỘI DUNG

- ▶ 1. Lý do chọn đề tài
- ▶ 2. Giới thiệu dataset
- ▶ 3. Tiền xử lý dữ liệu
- ▶ **4. Các mô hình học máy áp dụng**
- ▶ 5. Kết quả
- ▶ 6. Kết luận

IV. CÁC MÔ HÌNH HỌC MÁY ÁP DỤNG

Các mô hình học máy

1. Logistic Regression
2. Decision Tree
3. Random Forest
4. LightGBM

IV. CÁC MÔ HÌNH HỌC MÁY ÁP DỤNG

Logistic Regression

Nguyên lý:

- Mô hình tuyến tính dự đoán xác suất nhị phân:

$$P(y = 1|X) = \sigma(\beta_0 + \beta_1 X_1 + \dots + \beta_n X_r)$$

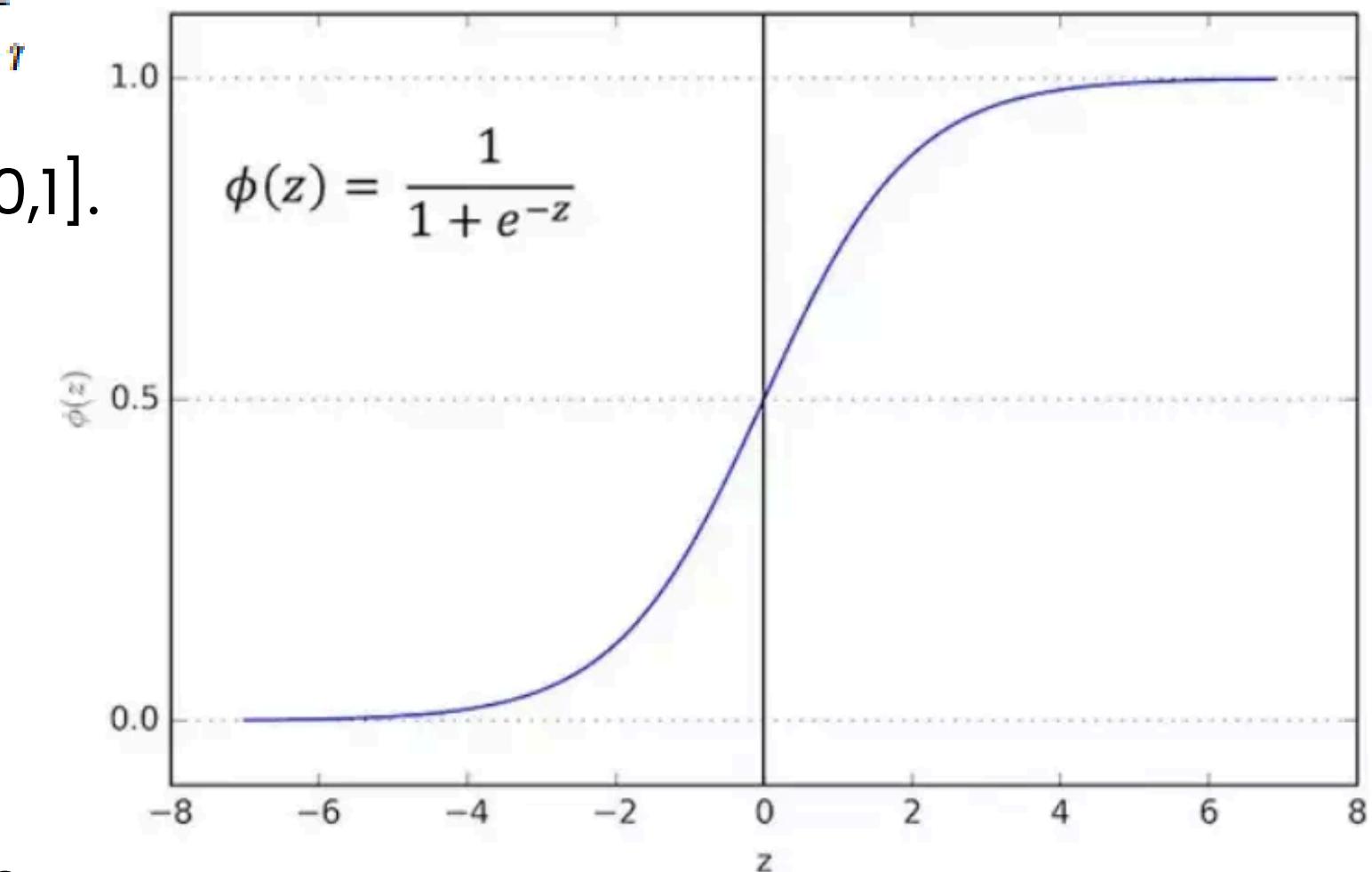
- Sigmoid đưa giá trị tuyến tính vào khoảng [0,1].

Ưu điểm:

- Dễ diễn giải, trực quan
- Tốt cho dữ liệu tuyến tính
- Tích hợp trọng số lớp dễ dàng

Nhược điểm:

- Không học tốt mối quan hệ phi tuyến
- Nhạy với mất cân bằng và thang đo biến

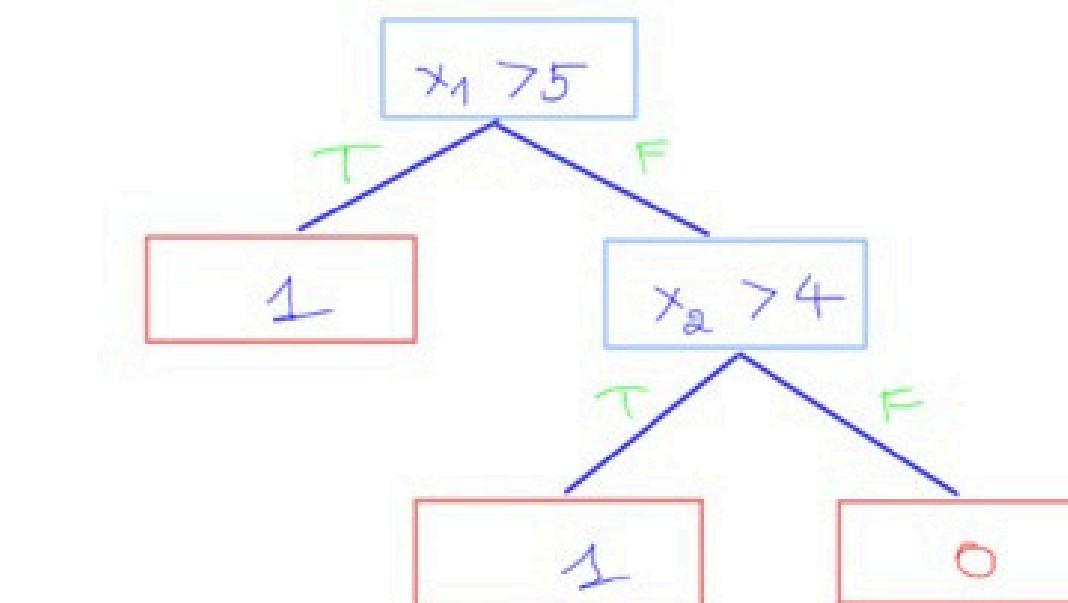
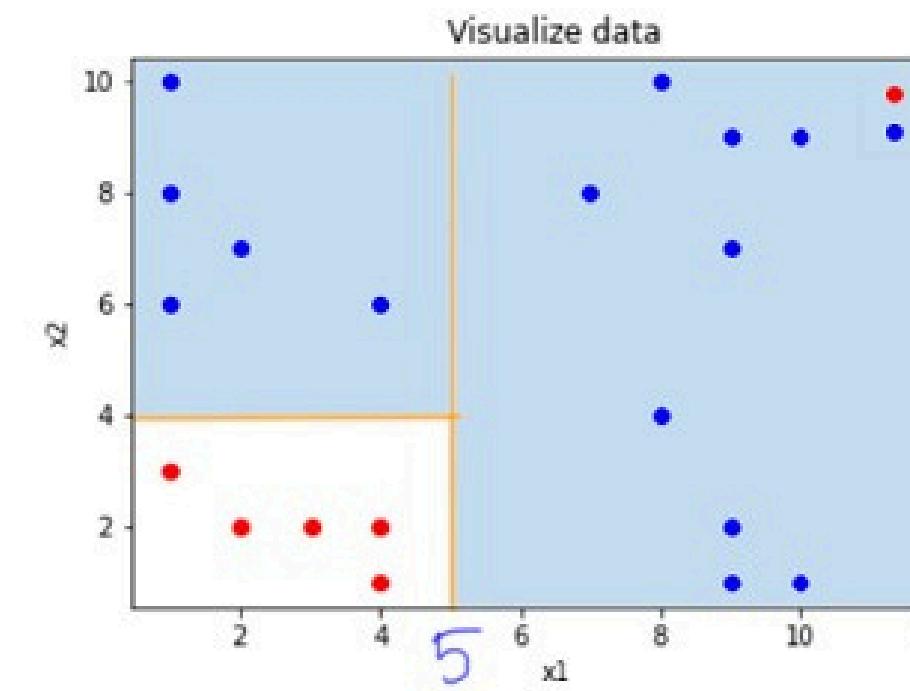


IV. CÁC MÔ HÌNH HỌC MÁY ÁP DỤNG

Decision Tree

Nguyên lý:

- Mô hình chia dữ liệu thành các nhánh dựa trên đặc trưng quan trọng, gọi là “nút”
- Mỗi nút kiểm tra một điều kiện, phân loại hoặc dự đoán kết quả cho dữ liệu.
- Quá trình này lặp lại cho đến khi đạt điều kiện dừng (lá cây).



Dễ overfitting nếu cây quá sâu.

IV. CÁC MÔ HÌNH HỌC MÁY ÁP DỤNG

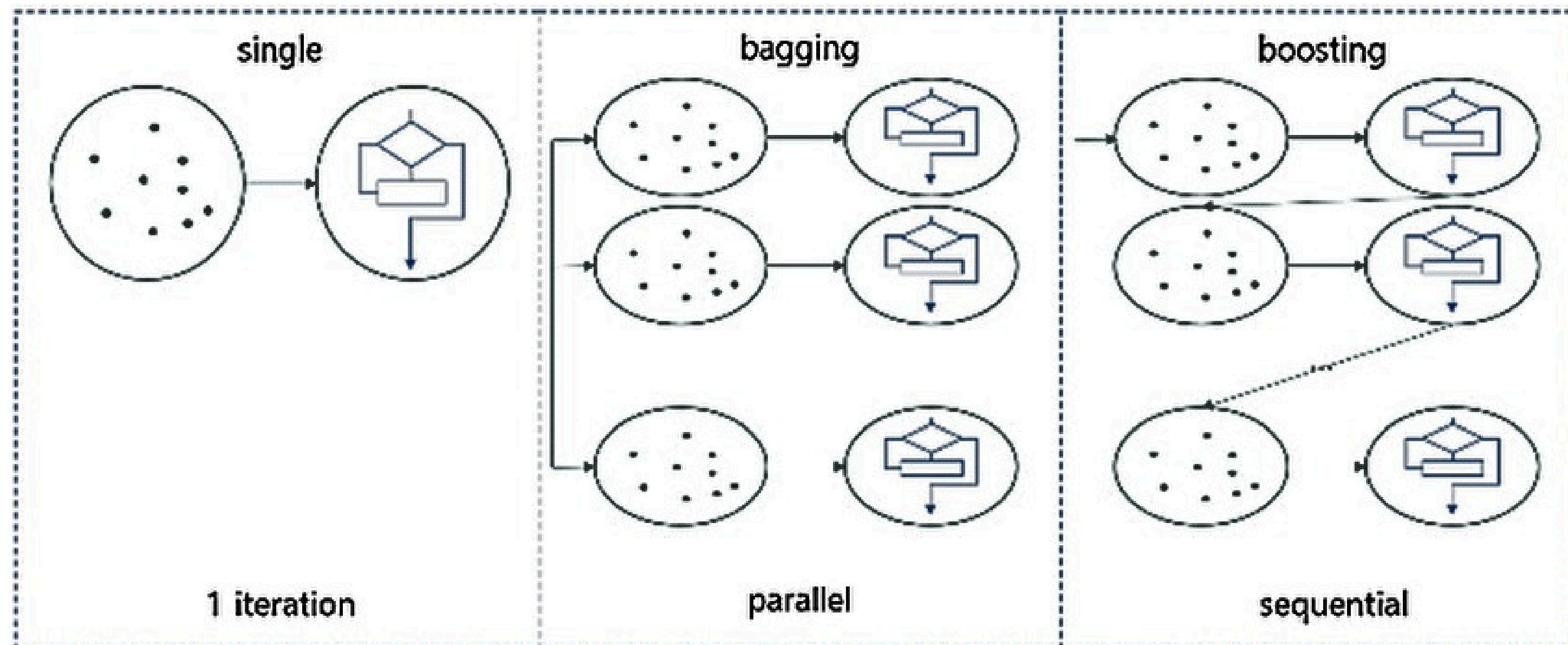
Ensemble Learning

- Khái niệm: Kết hợp nhiều mô hình cơ bản (base models) để tạo ra mô hình mạnh hơn, giảm bias hoặc variance.

Các nhóm chính:

- Bagging – Giảm variance
 - Dùng nhiều model cùng loại trên các subsample khác nhau.
 - Kết quả cuối = trung bình/đa số phiếu của các model.
- Boosting – Giảm bias
 - Mỗi model học sửa lỗi của model trước.
 - Trọng số dữ liệu cập nhật qua từng bước.
- Stacking – Giảm bias, kết hợp nhiều loại model
 - Dùng meta-model để học cách kết hợp kết quả các base models..

IV. CÁC MÔ HÌNH HỌC MÁY ÁP DỤNG



IV. CÁC MÔ HÌNH HỌC MÁY ÁP DỤNG

Random Forest

Nguyên lý:

- Tạo nhiều cây quyết định từ các tập con dữ liệu và tập con đặc trưng
- Kết quả dự đoán là trung bình hoặc đa số của các cây.
- Giảm nguy cơ overfitting so với cây đơn lẻ.

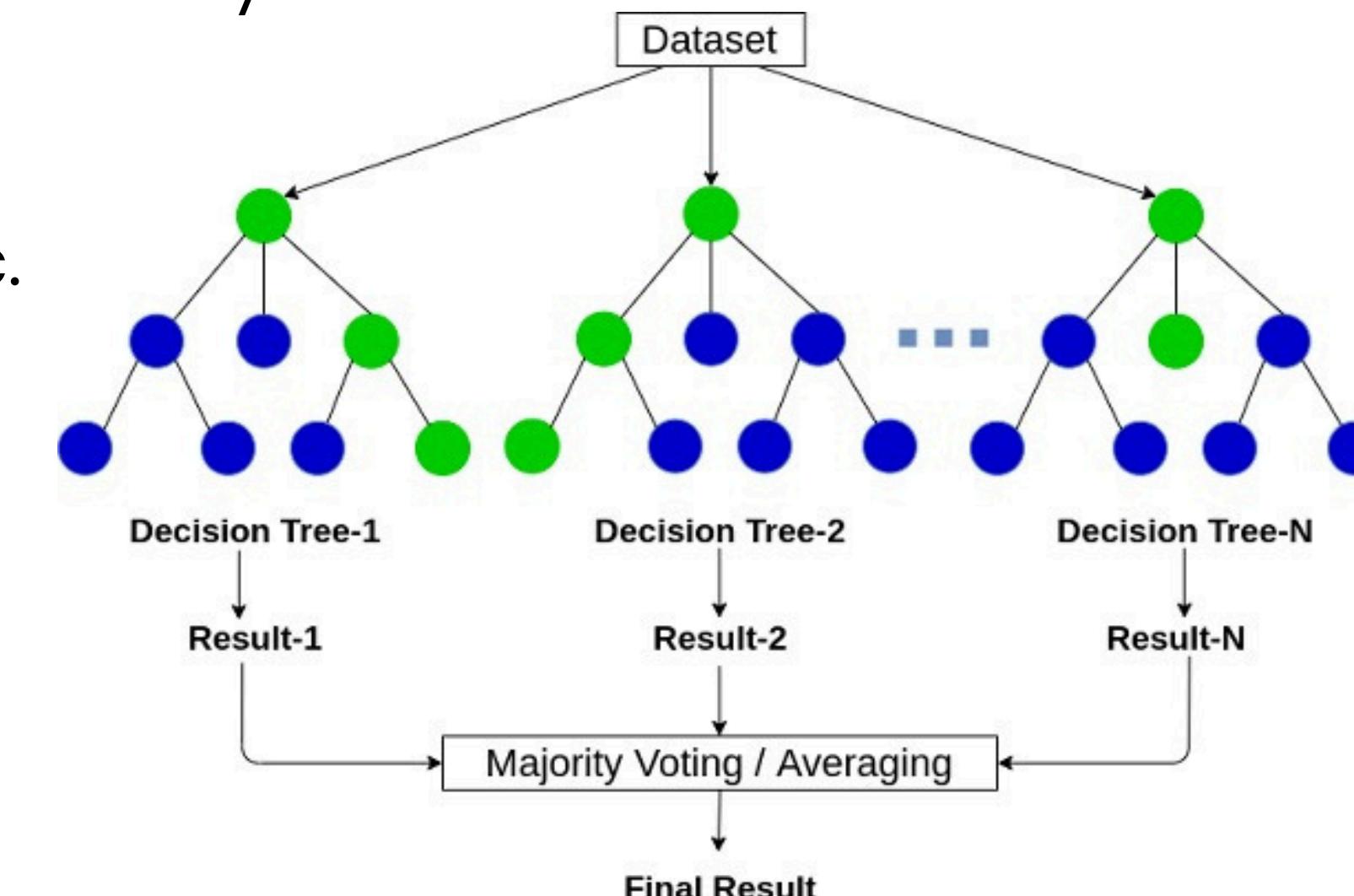
.

Ưu điểm:

- Dự đoán ổn định, chính xác.
- Giảm overfitting so với Decision Tree đơn lẻ.

Nhược điểm:

- Tốn tài nguyên tính toán.



IV. CÁC MÔ HÌNH HỌC MÁY ÁP DỤNG

Gradient Boosting

Nguyên lý:

- Gradient Boosting xây dựng mô hình bằng cách kết hợp nhiều cây quyết định yếu
- Mỗi cây mới được huấn luyện để sửa lỗi (gradient của hàm mất mát) của mô hình trước

Ưu điểm:

- Chính xác, hiệu quả với dữ liệu phức tạp.
- Thích hợp với các bài toán hồi quy và phân loại.

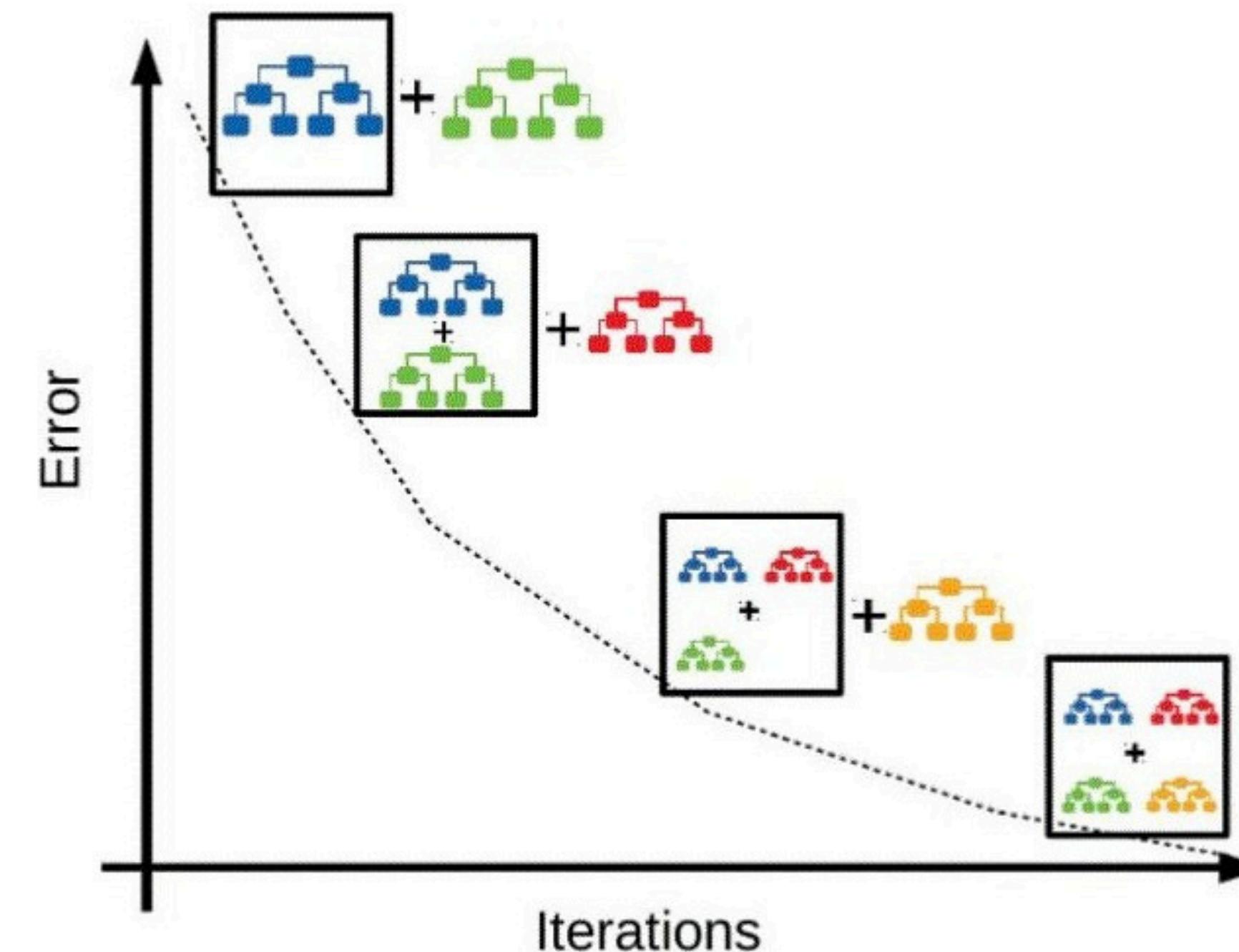
Nhược điểm:

Huấn luyện chậm, tốn thời gian.

Dễ overfitting nếu số lượng cây quá lớn hoặc cây quá sâu.

V. CÁC MÔ HÌNH HỌC MÁY ÁP DỤNG

Gradient Boosting



IV. CÁC MÔ HÌNH HỌC MÁY ÁP DỤNG

LightGBM

- Là phiên bản cải tiến của Gradient Boosting Decision Tree (GBDT).
- Huấn luyện dựa trên gradient và hessian của hàm mất mát

Các kỹ thuật bổ sung

- Leaf-wise tree growth: mỗi lần chia lá làm giảm loss nhiều nhất (nhanh hơn level-wise)
- Histogram-based learning: rời rạc hóa feature thành các bin để tăng tốc và giảm bộ nhớ
- GOSS: giữ toàn bộ mẫu có gradient lớn, giảm số mẫu ít quan trọng để tăng tốc
- EFB: gộp các feature thừa, hiếm khi cùng khác 0 để giảm số chiều dữ liệu

IV. CÁC MÔ HÌNH HỌC MÁY ÁP DỤNG

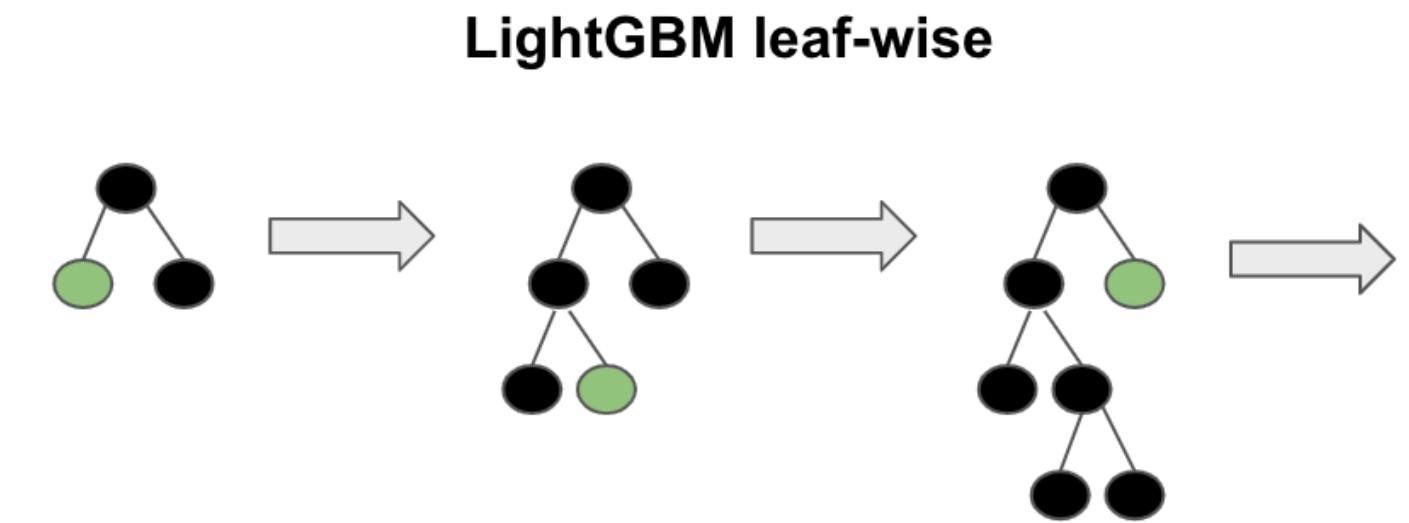
LightGBM

Ưu điểm

- Huấn luyện rất nhanh, phù hợp với dữ liệu lớn
- Tiết kiệm bộ nhớ
- Hiệu suất cao, thường tốt hơn GBDT truyền thống

Nhược điểm

- Dễ overfitting nếu không giới hạn số lá và độ sâu cây
- Cấu trúc mô hình phức tạp, khó giải thích
- Nhạy với tham số, cần tinh chỉnh cẩn thận



CÁC CHỈ SỐ ĐÁNH GIÁ MÔ HÌNH

Accuracy

Tỷ lệ dự đoán đúng trên toàn bộ dữ liệu

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Dễ hiểu nhưng **không phù hợp** khi dữ liệu mất cân bằng

Precision

Trong các khách hàng bị dự đoán **rủi ro**, có bao nhiêu là **rủi ro thật**

$$Precision = \frac{TP}{TP + FP}$$

Quan trọng khi muốn **giảm cảnh báo nhầm**

CÁC CHỈ SỐ ĐÁNH GIÁ MÔ HÌNH

Recall

Trong các khách hàng rủi ro thật, mô hình phát hiện được bao nhiêu

$$Recall = \frac{TP}{TP + FN}$$

Chỉ số quan trọng nhất trong bài toán tín dụng

ROC-AUC

Đo khả năng phân biệt hai lớp trên mọi ngưỡng dự đoán

Không phụ thuộc threshold

Giá trị càng gần 1 → mô hình càng tốt

CÁC CHỈ SỐ ĐÁNH GIÁ MÔ HÌNH

TPR (True Positive Rate / Recall)

$$TPR = \frac{TP}{TP + FN}$$

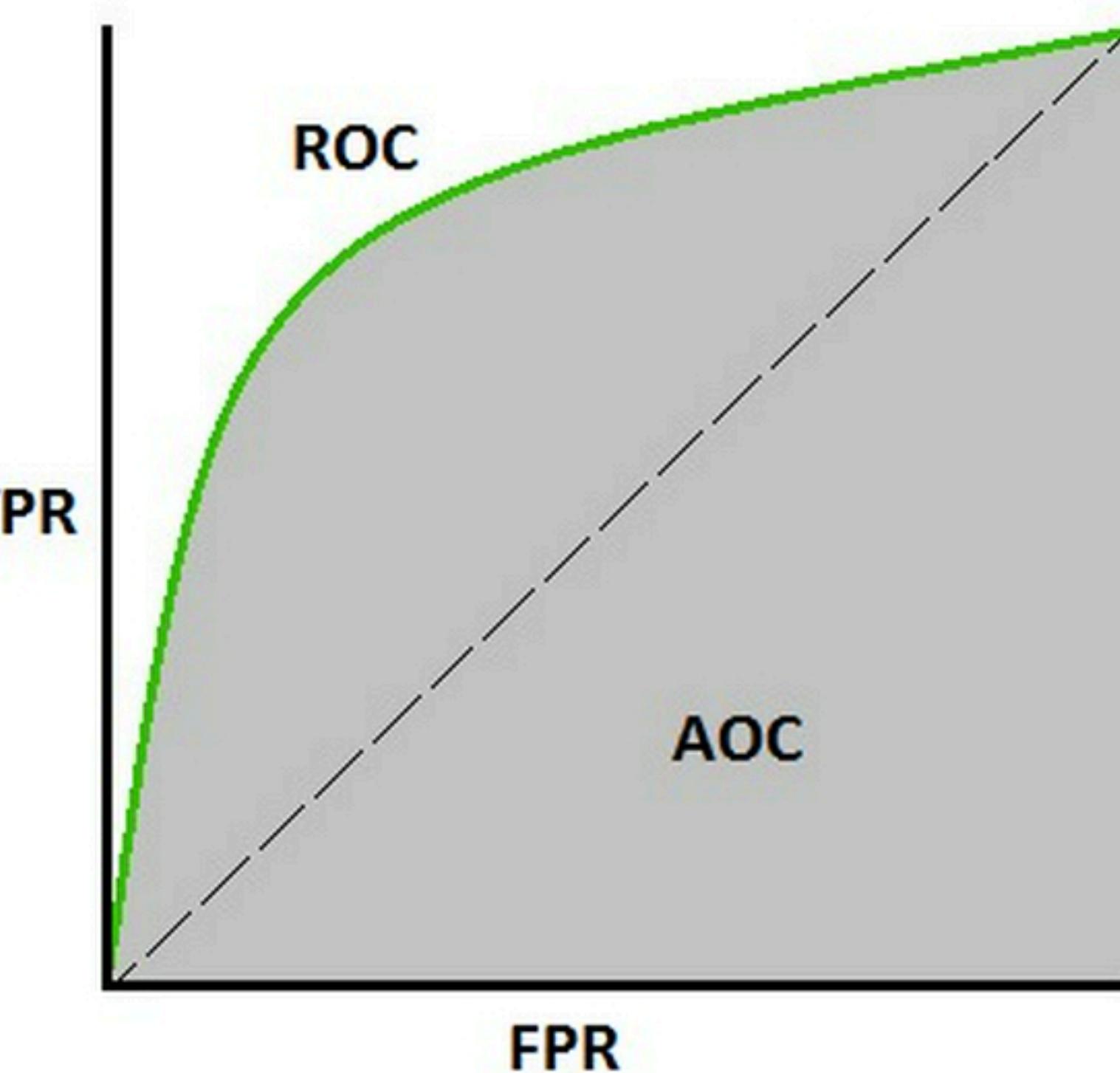
→ Tỷ lệ khách hàng rủi ro được phát hiện đúng

FPR (False Positive Rate)

$$FPR = \frac{FP}{FP + TN}$$

→ Tỷ lệ khách hàng an toàn bị dự đoán nhầm là rủi ro

→ Phản ánh mức độ cảnh báo sai / từ chối nhầm khách hàng tốt



NỘI DUNG

- ▶ 1. Lý do chọn đề tài
- ▶ 2. Giới thiệu dataset
- ▶ 3. Tiền xử lý dữ liệu
- ▶ 4. Các mô hình học máy áp dụng
- ▶ **5. Kết quả**
- ▶ 6. Kết luận

V. KẾT QUẢ

	Model	Accuracy	Precision	Recall	ROC-AUC
0	Logistic Regression	0.791735	0.515040	0.786317	0.868276
1	Decision Tree	0.913871	0.844350	0.742268	0.905213
2	Random Forest	0.921031	0.868906	0.751640	0.930776
3	LightGBM	0.894640	0.716641	0.855670	0.950867

NỘI DUNG

- ▶ 1. Lý do chọn đề tài
- ▶ 2. Giới thiệu dataset
- ▶ 3. Tiền xử lý dữ liệu
- ▶ 4. Các mô hình học máy áp dụng
- ▶ 5. Kết quả
- ▶ 6. Kết luận**

VI. KẾT LUẬN

- Bài toán đánh giá rủi ro tín dụng được mô hình hóa hiệu quả dưới dạng phân loại nhị phân.
- Các mô hình học máy dựa trên cây (Decision Tree, Random Forest, LightGBM) cho kết quả vượt trội hơn Logistic Regression nhờ khả năng học quan hệ phi tuyến trong dữ liệu.
- LightGBM đạt hiệu suất tốt nhất trên tập kiểm tra với ROC-AUC cao và Recall tốt, thể hiện khả năng phân biệt rõ giữa khách hàng rủi ro và an toàn.
- Việc điều chỉnh ngưỡng dự đoán (threshold) giúp cân bằng hợp lý giữa Recall và Precision, phù hợp với mục tiêu hạn chế bỏ sót khách hàng rủi ro trong thực tế tín dụng.



DEMO



THANK YOU