TRƯỜNG ĐẠI HỌC MỞ THÀNH PHỐ HỒ CHÍ MINH

KHOA KHOA HỌC CƠ BẢN

ĐỀ TÀI:

ỨNG DỤNG MÁY HỌC TRONG DỰ ĐOÁN NGUY CƠ BỆNH TIM MẠCH

Môn học: Máy học (Machine Learning)

Giảng viên hướng dẫn: Hà Minh Tuấn

Sinh viên thực hiện:
Nguyễn Thảo Quyên
Nguyễn Phạm Triệu Vỹ

Giới thiệu Đề tài: Thách thức và Tiềm năng

Bệnh tim mạch là một trong những nguyên nhân gây tử vong hàng đầu trên thế giới

Nguy cơ mắc bệnh tim mạch không phụ thuộc vào một yếu tố đơn lẻ mà là sự kết hợp phức tạp của nhiều yếu tố như tuổi tác, huyết áp, cholesterol, cân nặng và thói quen sinh hoạt. Do đó, việc ứng dụng máy học để phân tích dữ liệu y sinh và dự đoán nguy cơ mắc bệnh tim mạch là một hướng tiếp cận có ý nghĩa thực tiễn cao.

Dữ liệu Nghiên cứu: Cardiovascular Disease Dataset từ Kaggle

- Dữ liệu Nghiên cứu
- Nguồn: Kaggle
- Tác giả: Olga Sushanova
- Quy mô: Khoảng 70000 bản ghi, mỗi bản ghi đại diện cho 1 bệnh nhân

Dữ liệu dạng bảng, phù hợp cho bài toán học máy có giám sát.

12 biến đầu vào và 1 biến mục tiêu

Biến mục tiêu:

Cardio là biến nhị phân, đóng vai trò then chốt và là biến đầu ra trong việc xác định tình trạng sức khỏe tim mạch của bệnh nhân:

- cardio = 0: Bệnh nhân không mắc bệnh tim mạch
- cardio = 1: Bệnh nhân có nguy cơ hoặc đang mắc bệnh tim mạch

Biến đầu vào	Mô tả	Kiểu dữ liệu / Mã hóa	Đơn vị / Giá trị
age	Tuổi của bệnh nhân, tính bằng số	Số nguyên	Ngày
gender	Giới tính của bệnh nhân	Phân loại	Mã hóa nhị phân
height	Chiều cao của bệnh nhân	Số liên tục	cm
weight	Cân nặng của bệnh nhân	Số liên tục	kg
ap_hi	Huyết áp tâm thu, phản ánh áp lực	Số liên tục	mmHg
ap_lo	Huyết áp tâm trương, phản ánh	Số liên tục	mmHg
cholesterol	Mức cholesterol trong máu	Phân loại thứ bậc	1: Bình thường2: Cao hơn bình
gluc	Mức glucose trong máu	Phân loại thứ bậc	1: Bình thường2: Cao hơn bình
smoke	Thói quen hút thuốc	Nhị phân	0: Không, 1: Có
alco	Thói quen sử dụng rượu bia	Nhị phân	0: Không, 1: Có
active	Mức độ hoạt động thể chất	Nhị phân	0: Không, 1: Có

Tiền xử lý Dữ liệu: Đảm bảo Chất lượng

Quá trình tiền xử lý dữ liệu được thực hiện cẩn thận để loại bỏ nhiễu và đảm bảo tính chính xác, phù hợp cho mô hình học máy.

01

Làm sạch & Chuyển đổi

Loại bỏ cột ID không ảnh hưởng đến quá trình xử lý, kiểm tra và xử lý missing values (không có) và trùng lặp (loại bỏ). Tuổi được chuyển đổi từ ngày sang năm.

02

Hiệu chỉnh Huyết áp

Lấy giá trị tuyệt đối, chặn trong ngưỡng sinh lý (**ap_hi: 60-245 mmHg**, **ap_lo: 40-160 mmHg**) và loại bỏ các trường hợp **ap_lo ≥ ap_hi**.
Làm sạch dữ liệu theo kiến thức y học, giúp mô hình học đúng bản chất

03

Hiệu chỉnh Chiều cao, Cân nặng, BMI

Giới hạn chiều cao (**140-210 cm**), cân nặng (**40-180 kg**). Tính toán và chặn BMI trong khoảng **15-50** theo khuyến nghị của WHO.

04

Thống kê mô tả sau làm sạch

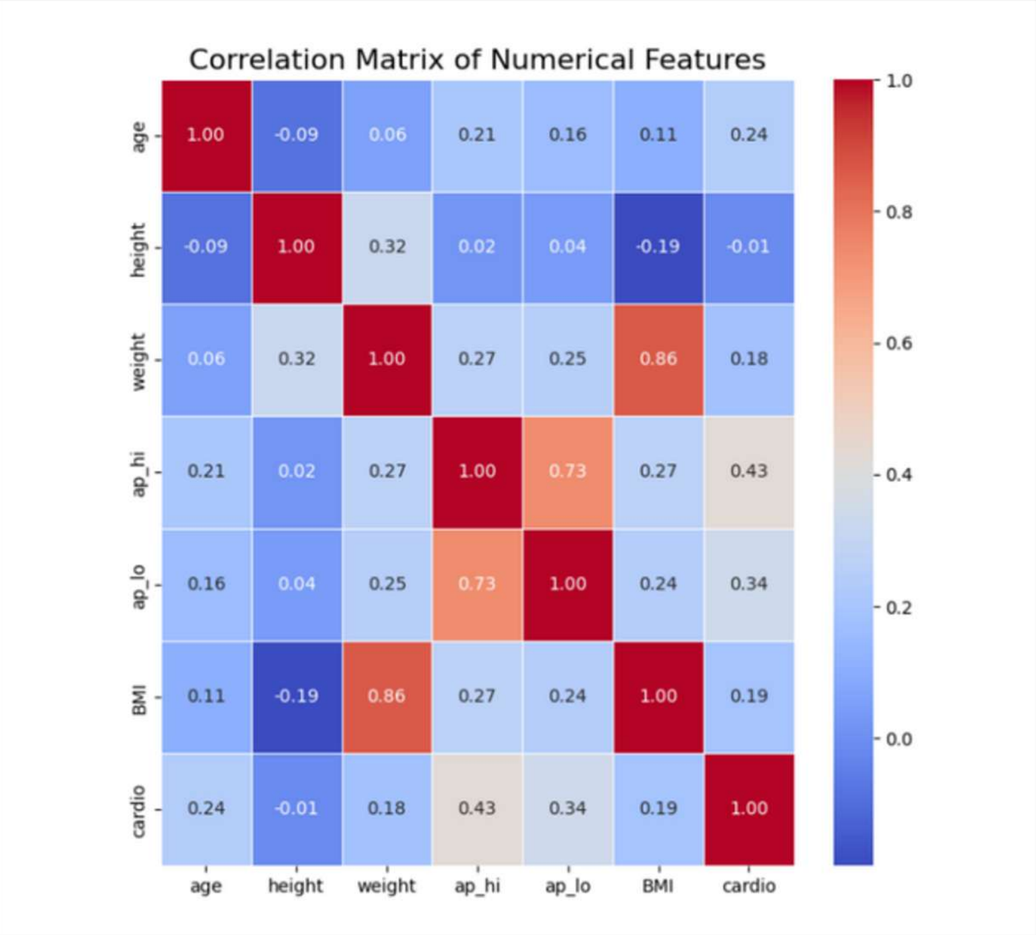
- Tuổi trung vị: **53**
  - **Phạm vi tuổi: 29 – 64**
- Chiều cao trung bình: **165 cm**
- Cân nặng trung bình: **74 kg**
- BMI trung bình: **27.4**
  - **Phân loại: Thừa cân (theo WHO)**
- Huyết áp trung bình: **120/80 mmHg**
  - **Phân loại: Cân bằng**

Các ngưỡng chặn được thiết lập dựa trên kiến thức y sinh thực tế nhằm loại bỏ các giá trị phi sinh học hoặc do lỗi nhập liệu, giúp dữ liệu phản ánh đúng đặc điểm sinh lý con người và tăng độ tin cậy cho mô hình.

# age in days	# gender 1 - women, 2 - men	# height cm	# weight kg	# ap_hi Systolic blood pressure	# ap_lo Diastolic blood pressure	# cholesterol 1: normal, 2: above normal, 3: well above normal	# gluc 1: normal, 2: above normal, 3: well above normal	# smoke whether patient smokes or not
70000 total values	**70000** total values	**70000** total values	**70000** total values	**70000** total values	**70000** total values	**70000** total values	**70000** total values	**70000** total values
17438	1	169	70.0	16020	80	1	1	0
15835	2	169	75.0	14020	80	2	1	0
21361	1	169	71.0	14020	80	3	3	0
16910	2	180	78.0	14020	90	1	1	0
19731	1	160	65.0	14020	90	1	1	0

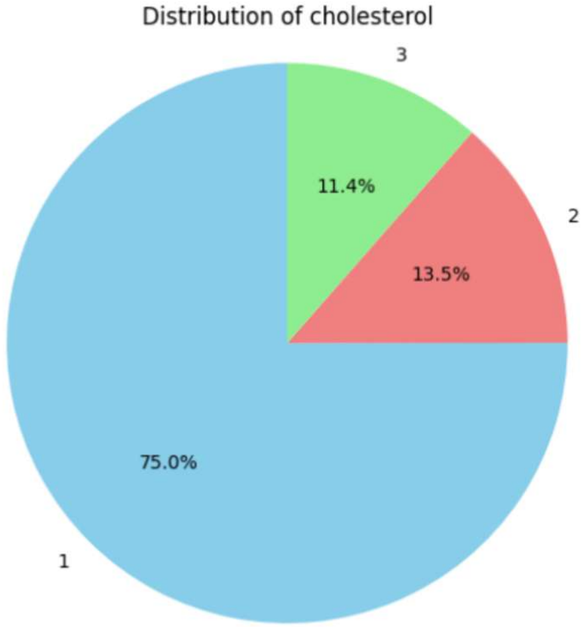
	age	gender	height	weight	ap_hi	ap_lo	cholesterol	gluc	smoke	alco	active	cardio	BMI
0	50	2	168	62.0	110	80	1	1	0	0	1	0	21.967120
1	55	1	156	85.0	140	90	3	1	0	0	1	1	34.927679
2	51	1	165	64.0	130	70	3	1	0	0	0	1	23.507805
3	48	2	169	82.0	150	100	1	1	0	0	1	1	28.710479
4	47	1	156	56.0	100	60	1	1	0	0	0	0	23.011177

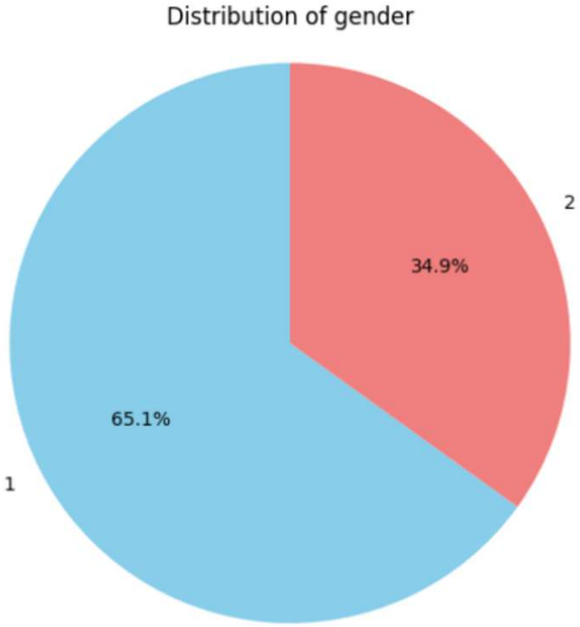
Phân tích Tương quan: Mối quan hệ giữa các Biến

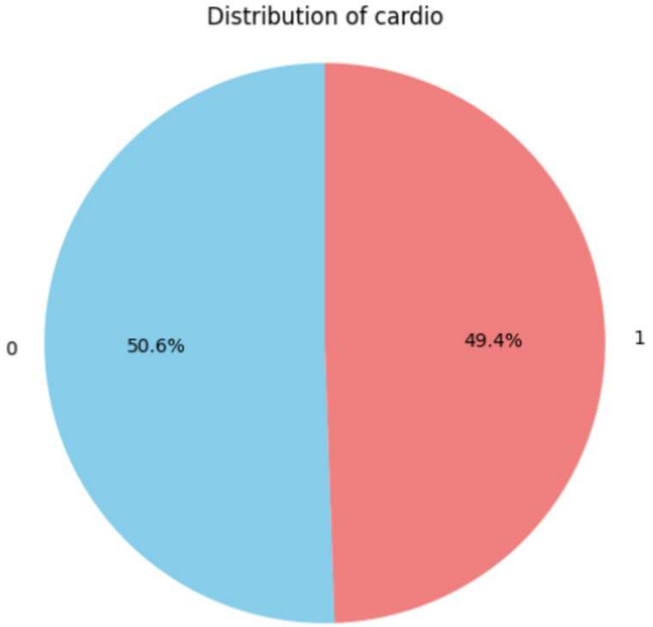


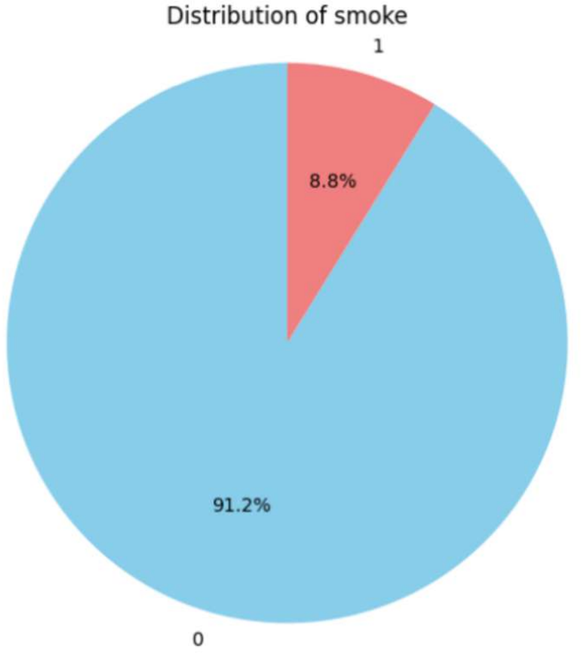
Việc phân tích tương quan giữa các biến đầu vào và biến mục tiêu là rất quan trọng để hiểu rõ hơn về dữ liệu và tránh các vấn đề như đa cộng tuyến.

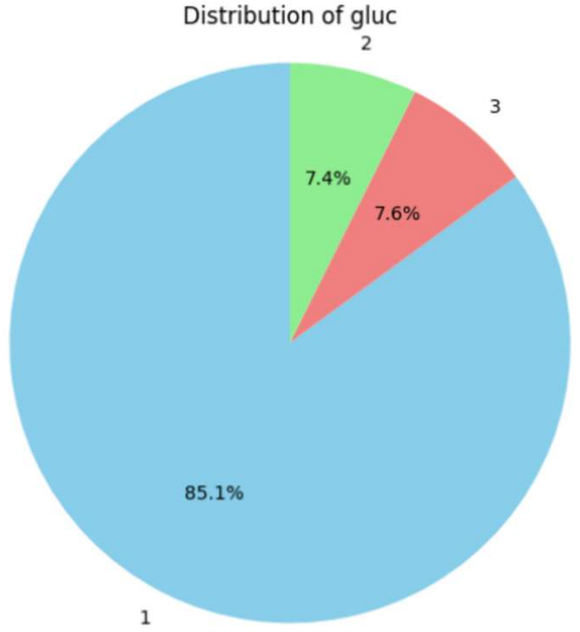
- **Đa cộng tuyến: Weight, height và BMI chứa thông tin trùng lặp (BMI phụ thuộc vào cân nặng và chiều cao), nên giữ đồng thời cả ba gây đa cộng tuyến và khó diễn giải mức độ ảnh hưởng của từng biến.**
- **Huyết áp: ap_hi và ap_lo tương quan mạnh với nhau và đều tương quan dương rõ rệt với cardio, cho thấy huyết áp là yếu tố dự báo quan trọng nhất.**
- **Tuổi & BMI: Age có tương quan dương nhẹ với cardio. BMI tương quan dương với huyết áp, phản ánh béo phì làm tăng huyết áp và gián tiếp làm tăng nguy cơ bệnh tim mạch.**


Distribution of cholesterol
3
11.4%
2
13.5%
75.0%
1

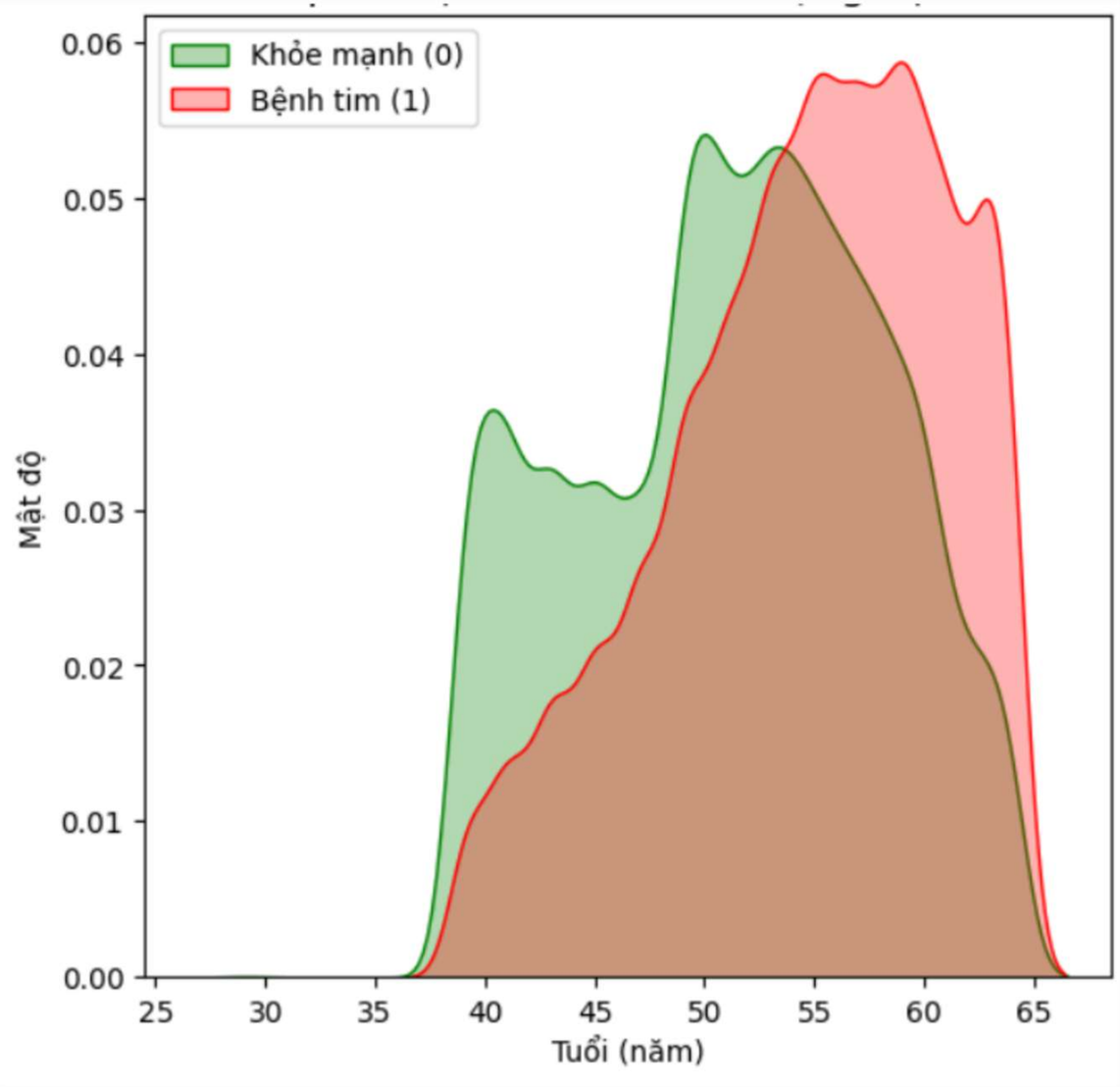

Distribution of gender
2
34.9%
65.1%
1


Distribution of cardio
0
50.6%
49.4%
1

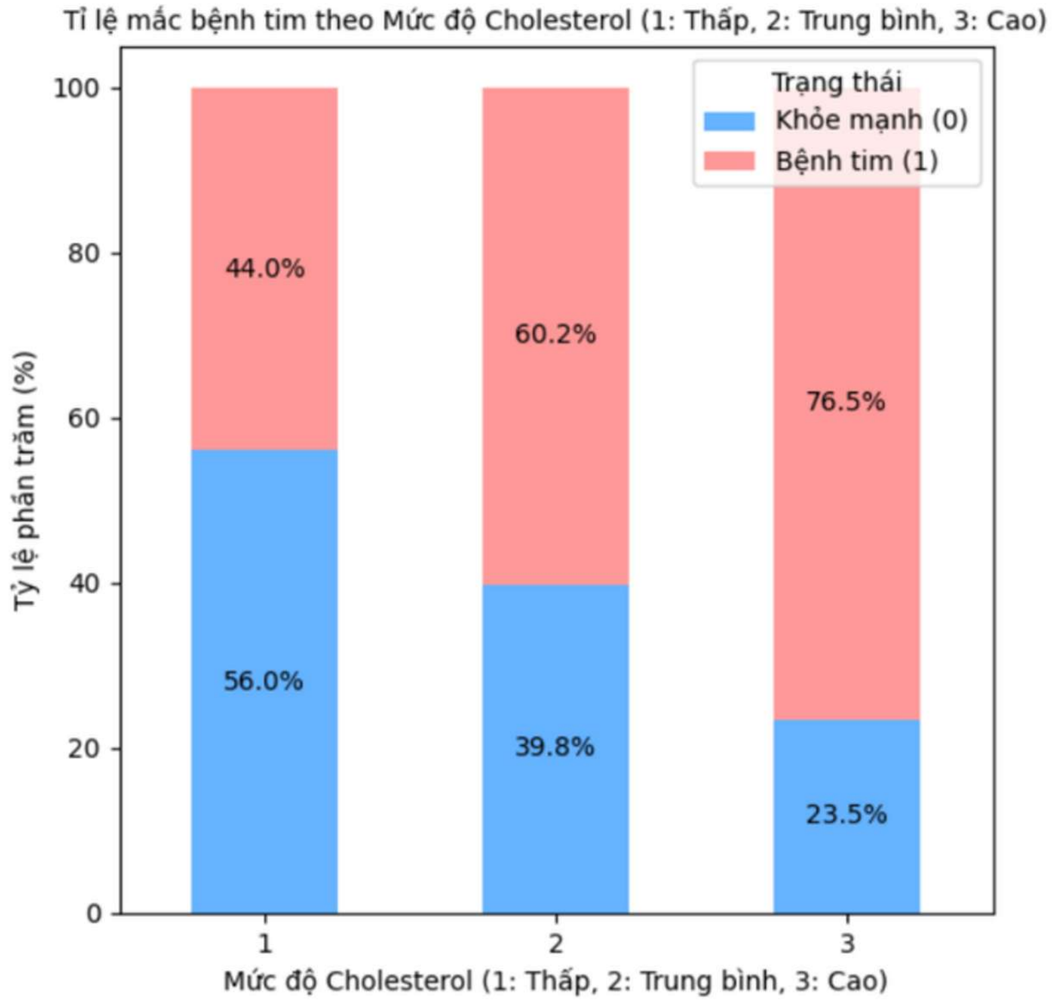

Distribution of smoke
1
8.8%
91.2%
0


Distribution of gluc
2
3
7.4%
7.6%
85.1%
1

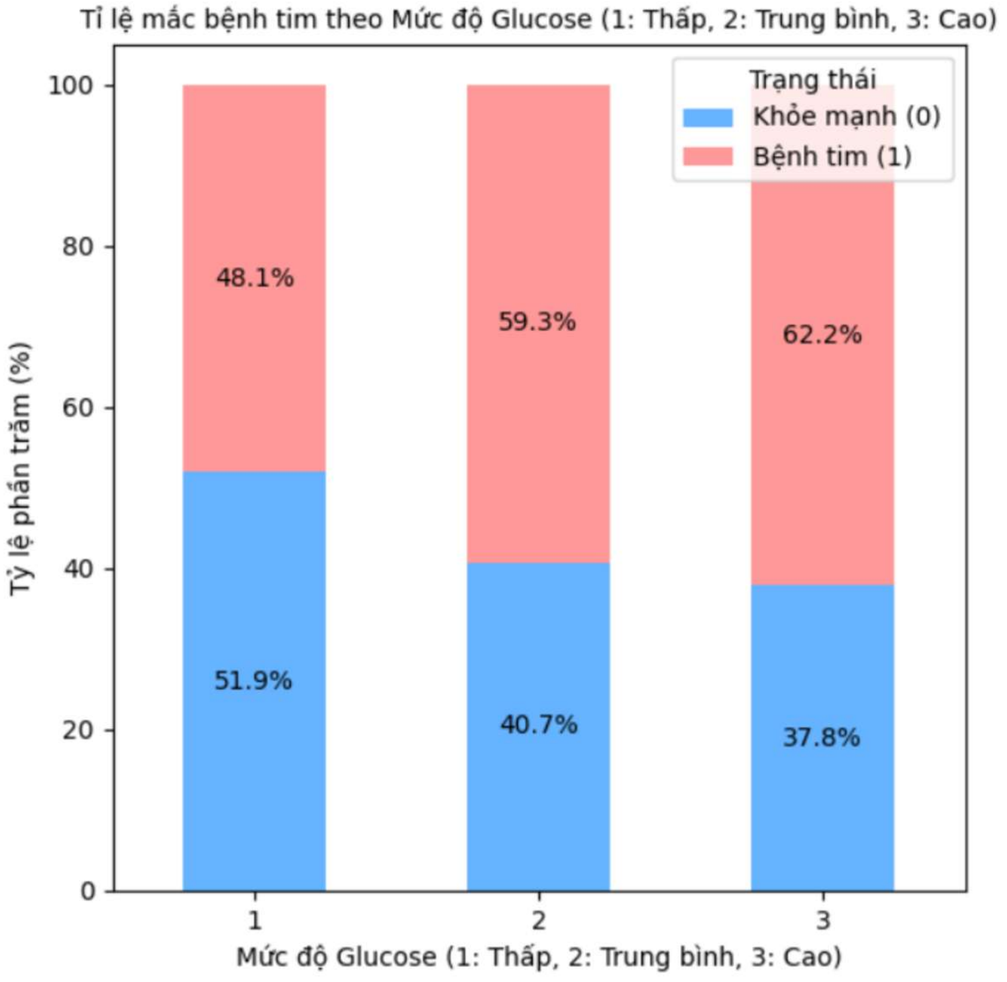
Xu hướng theo Tuổi: Mối liên hệ với Bệnh Tim mạch



- Tỷ lệ mắc bệnh **tăng dần theo độ tuổi** và đạt mức cao ở nhóm người cao tuổi.
- Kết quả khẳng định **age là một predictor quan trọng**, cần được giữ lại trong quá trình huấn luyện mô hình học máy để nâng cao khả năng dự đoán nguy cơ bệnh tim mạch.



Nhóm cholesterol thấp có tỷ lệ mắc bệnh khoảng 44%, tăng lên 60.2% ở mức trung bình và đạt 76.5% ở mức cao. Kết quả khẳng định cholesterol là yếu tố nguy cơ quan trọng, với nguy cơ mắc bệnh tim gia tăng theo mức cholesterol.



Biểu đồ cho thấy tỉ lệ mắc bệnh tim tăng dần theo mức độ glucose. Xu hướng này cho thấy đường huyết cao có liên quan đến nguy cơ mắc bệnh tim, tuy nhiên mức tăng không quá đột biến so với cholesterol.

Thiết lập mô hình & chiến lược đánh giá

Phân chia dữ liệu

Logistic Regression 80% train, 20% test; Cross-validation 5-fold.

Random Forest 80% train, 20% test, Cross-validation 5-fold.

Xgboost 64% train, 20% test, 16% valid

Metrics ưu tiên

Ưu tiên Recall để giảm thiểu False Negative, quan trọng trong y tế.

Hiệu chỉnh threshold

Ngưỡng quyết định được hiệu chỉnh, không cố định ở 0.5.

Feature engineering: tạo ra feature mới nhằm cải tiến performance

pulse_pressure: Giá trị cao → động mạch cứng → nguy cơ tim mạch tăng,

Giúp model học được mối quan hệ phi tuyến giữa huyết áp và nguy cơ bệnh tim

Các chỉ số đánh giá

Accuracy

Độ chính xác tổng thể của mô hình.

Precision

Tỷ lệ dự đoán dương tính đúng.

Recall

Tỷ lệ dự đoán đúng các trường hợp dương tính thực sự.

F1-Score

Cân bằng giữa Precision và Recall.

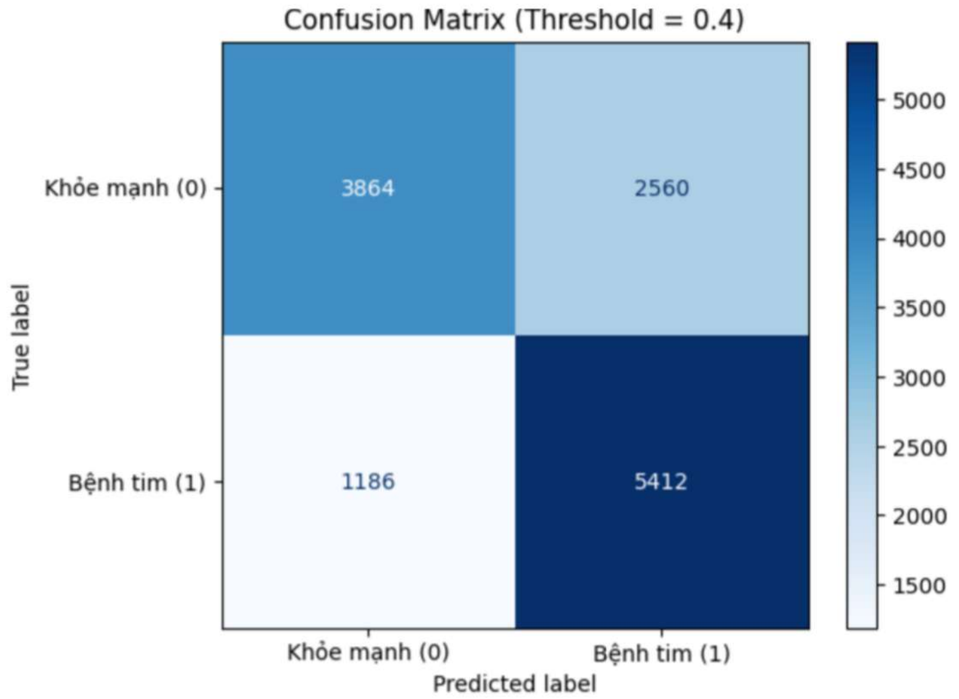
ROC–AUC

Đo lường khả năng phân loại của mô hình.

Lưu ý: Trong bối cảnh y sinh, bỏ sót bệnh nhân nguy hiểm hơn báo động giả, vì vậy Recall được ưu tiên hàng đầu.

Mô hình Logistic Regression

Điều chỉnh tham số	Thành phần	Giá trị
Tối ưu hóa	Phương pháp	GridSearchCV
Tham số tinh chỉnh	Penalty	l1, l2
	C	0.001, 0.01, 0.1, 1, 10, 100
	Class weight	None, balanced
Cấu hình cố định	Solver	liblinear
	Max iterations	1000
Đánh giá	Scoring	F1-score (threshold = 0.5)
	Cross-validation	5-fold

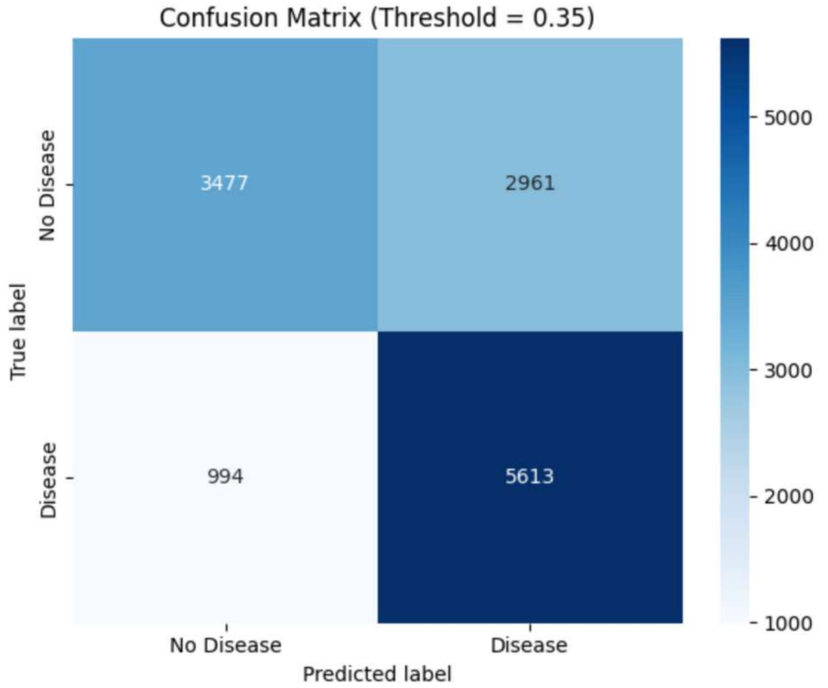


Classification Report:

	precision	recall	f1-score	support
0	0.77	0.60	0.67	6424
1	0.68	0.82	0.74	6598
accuracy			0.71	13022
macro avg	0.72	0.71	0.71	13022
weighted avg	0.72	0.71	0.71	13022

Mô hình RandomForestClassifier

Điều chỉnh tham số	Thành phần	Giá trị
Tối ưu hóa	Phương pháp	RandomizedSearchCV
Tham số tinh chỉnh	n_estimators	100, 300, 500, 800, 1000
	max_depth	5, 10, 20, 30, None
	min_samples_split	2, 5, 10, 20
	min_samples_leaf	1, 5, 10, 20
	max_features	sqrt, log2
	bootstrap	True, False
	class_weight	None, balanced
Đánh giá	Cross-validation	5-fold

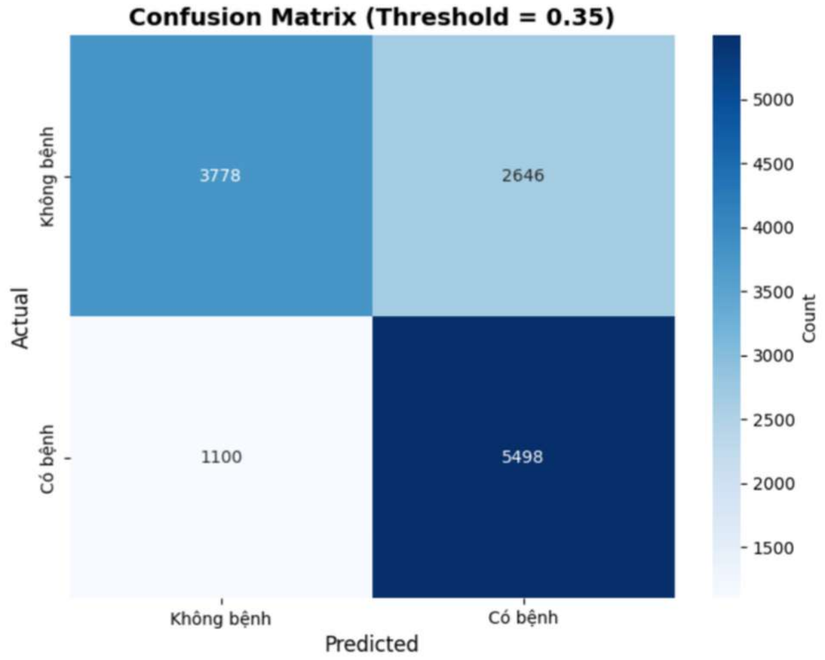


📋 Classification Report:

	precision	recall	f1-score	support
No Disease	0.78	0.54	0.64	6438
Disease	0.65	0.85	0.74	6607
accuracy			0.70	13045
macro avg	0.72	0.69	0.69	13045
weighted avg	0.72	0.70	0.69	13045

Mô hình XGBClassifier

Điều chỉnh tham số	Thành phần	Giá trị
Mô hình	Tên mô hình	XGBClassifier
Tối ưu hóa	Phương pháp	Optuna
Tham số tinh chỉnh	n_estimators	300 – 1000
	max_depth	3 – 20
	learning_rate	0.01 – 0.15
	subsample	0.6 – 1.0
	colsample_bytree	0.6 – 1.0
Tham số tiếp theo	gamma	0 – 5
	min_child_weight	1 – 10
	reg_alpha	0 – 5
	reg_lambda	1 – 10
	scale_pos_weight	0.5 – 2.0
Cấu hình cố định	objective	binary:logistic
	eval_metric	logloss
	early_stopping_rounds	50



	precision	recall	f1-score	support
Không bệnh (0)	0.77	0.59	0.67	6424
Có bệnh (1)	0.68	0.83	0.75	6598
accuracy			0.71	13022
macro avg	0.72	0.71	0.71	13022
weighted avg	0.72	0.71	0.71	13022

So sánh hiệu suất các mô hình (Ngưỡng = 0.35)

Mô hình	Recall	Precision	F1	ROC–AUC
Logistic Regression	0.82	0.68	0.73	0.79
Random Forest	**0.85**	0.65	0.74	0.79
XGBoost	0.83	0.68	**0.75**	**0.80**

Điểm nổi bật

- Random Forest có Recall cao nhất.
- XGBoost đạt F1 và ROC–AUC cao nhất.

Đánh giá tổng hợp & Lựa chọn mô hình

Logistic Regression

Diễn giải cao, ổn định. Phù hợp cho hệ thống hỗ trợ quyết định lâm sàng nhờ tính minh bạch.

Random Forest

Recall cao nhất, nhưng có thể tạo ra nhiều báo động giả, tăng chi phí kiểm tra không cần thiết.

XGBoost

F1-score & ROC–AUC cao nhất, cân bằng tốt giữa Recall và Precision, khai thác hiệu quả feature engineering.

XGBoost là mô hình cân bằng nhất cung cấp hiệu suất tổng thể vượt trội. Trong khi đó, Random Forest phù hợp khi cần tính sàng lọc.