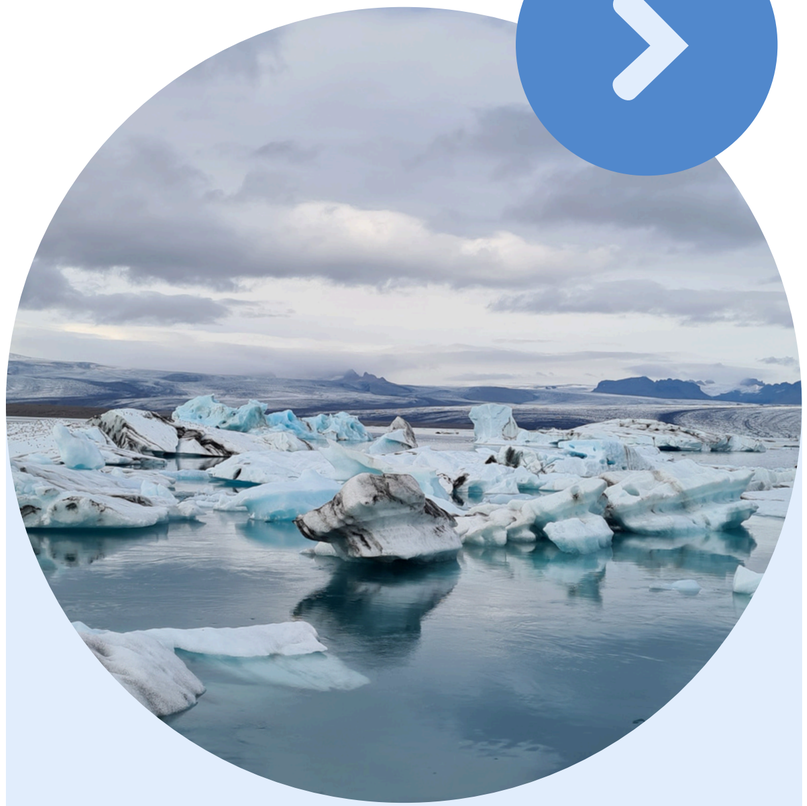
HỆ THỐNG PHÁT HIỆN BỆNH TIM

Mai Phương Nam - 22IT178

Trần Phước Anh - 22IT009

Mục lục



1. GIỚI THIỆU

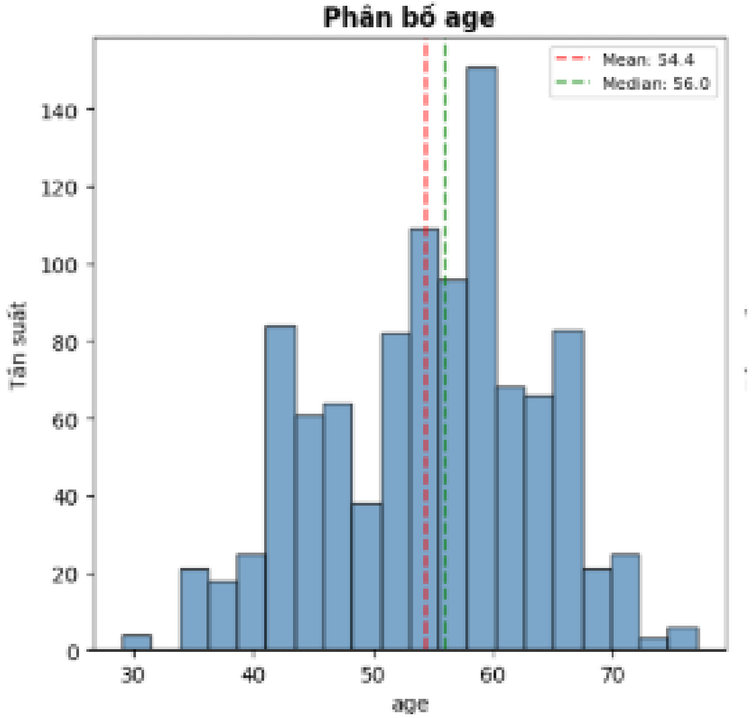
Đề tài “Hệ thống dự đoán nguy cơ mắc bệnh tim và triển khai trên nền tảng web bằng Django” nhằm mục tiêu kết hợp giữa AI và lập trình web để tạo ra một công cụ trực quan, dễ sử dụng, giúp người dùng tự đánh giá nguy cơ sức khỏe tim mạch của mình một cách nhanh chóng và tiện lợi.

2. THU THẬP & KHÁM PHÁ DỮ LIỆU

Nguồn: Heart Disease Dataset (~1025mẫu, 14 đặc trưng)

Mô tả thống kê dữ liệu

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
count	1025.000000	1025.000000	1025.000000	1025.000000	1025.00000	1025.000000	1025.000000	1025.000000	1025.000000	1025.000000	1025.000000	1025.000000	1025.000000	1025.000000
mean	54.434146	0.695610	0.942439	131.611707	246.00000	0.149268	0.529756	149.114146	0.336585	1.071512	1.385366	0.754146	2.323902	0.513171
std	9.072290	0.460373	1.029641	17.516718	51.59251	0.356527	0.527878	23.005724	0.472772	1.175053	0.617755	1.030798	0.620660	0.500070
min	29.000000	0.000000	0.000000	94.000000	126.00000	0.000000	0.000000	71.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	48.000000	0.000000	0.000000	120.000000	211.00000	0.000000	0.000000	132.000000	0.000000	0.000000	1.000000	0.000000	2.000000	0.000000
50%	56.000000	1.000000	1.000000	130.000000	240.00000	0.000000	1.000000	152.000000	0.000000	0.800000	1.000000	0.000000	2.000000	1.000000
75%	61.000000	1.000000	2.000000	140.000000	275.00000	0.000000	1.000000	166.000000	1.000000	1.800000	2.000000	1.000000	3.000000	1.000000
max	77.000000	1.000000	3.000000	200.000000	564.00000	1.000000	2.000000	202.000000	1.000000	6.200000	2.000000	4.000000	3.000000	1.000000

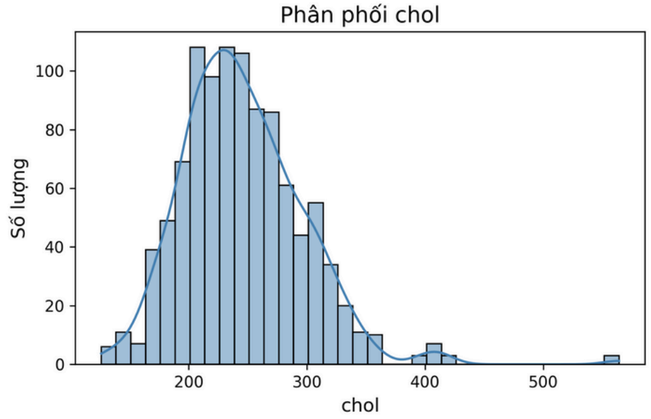


Kiểm tra dữ liệu thiếu

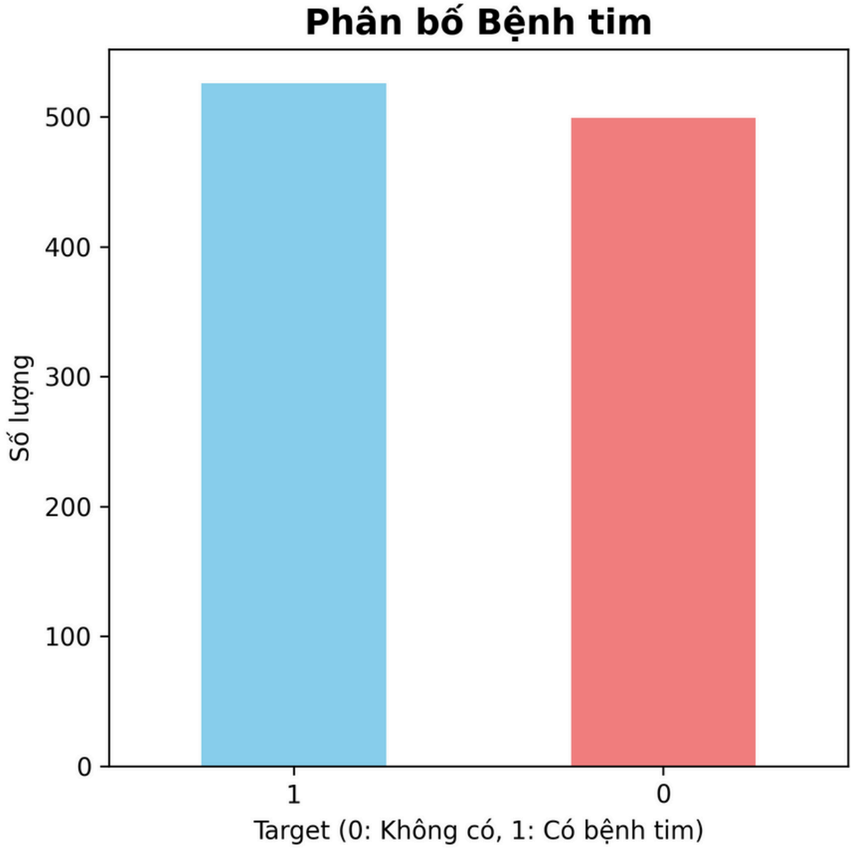
```
Số lượng giá trị thiếu theo từng cột:
age: 0
sex: 0
cp: 0
trestbps: 0
chol: 0
fbs: 0
restecg: 0
thalach: 0
exang: 0
oldpeak: 0
slope: 0
ca: 0
thal: 0
target: 0
```

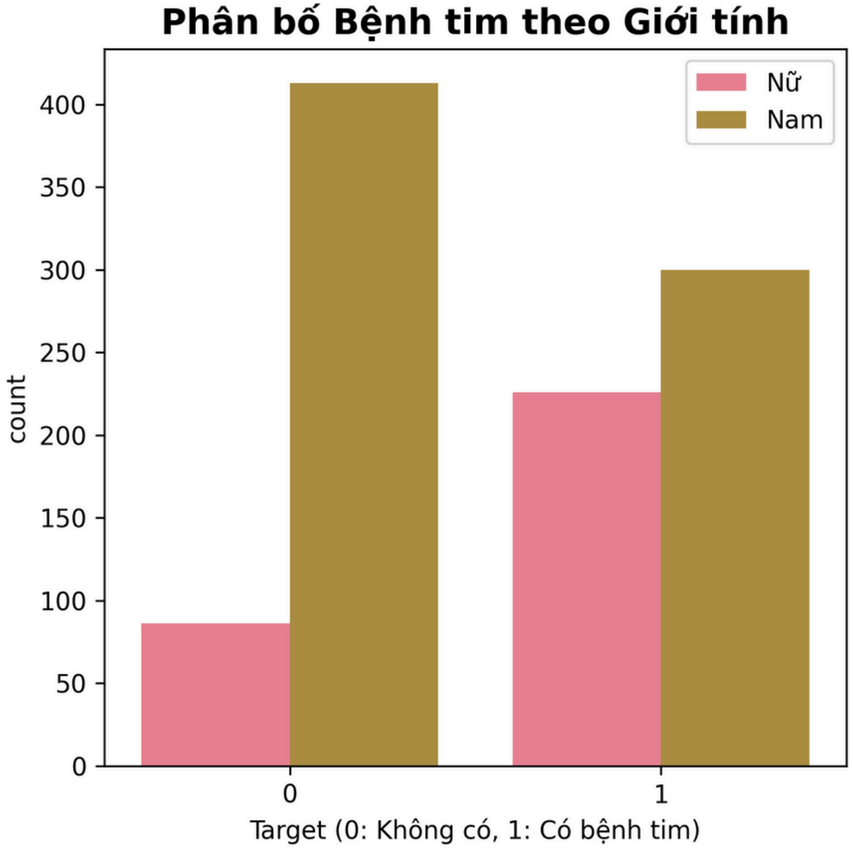
Phân tích biến mục tiêu

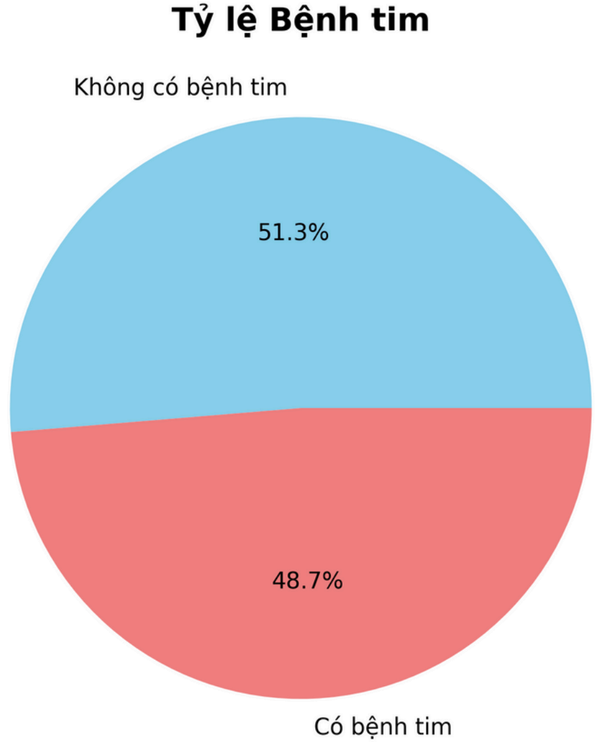
```
🎯 PHÂN TÍCH BIẾN MỤC TIÊU
==================================================
Phân bố lớp:
Không có bệnh tim (0): 499 (48.7%)
Có bệnh tim (1): 526 (51.3%)
```



2. THU THẬP & KHÁM PHÁ DỮ LIỆU







2. THU THẬP & KHÁM PHÁ DỮ LIỆU

Phân tích các biến số (Numerical Cols)

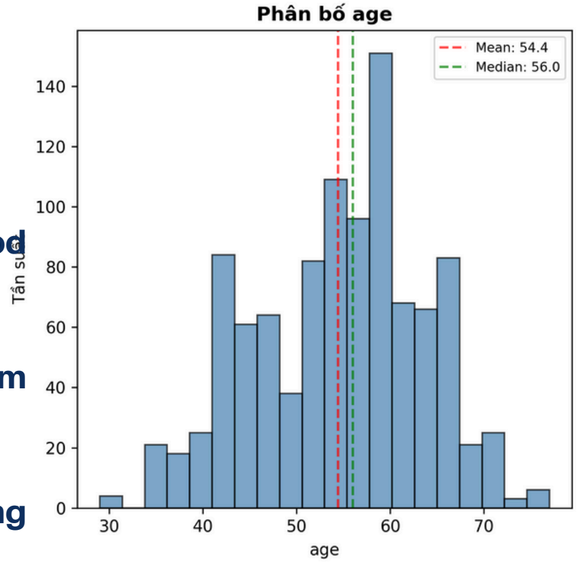
1. age : Tuổi của bệnh nhân (tính bằng năm)

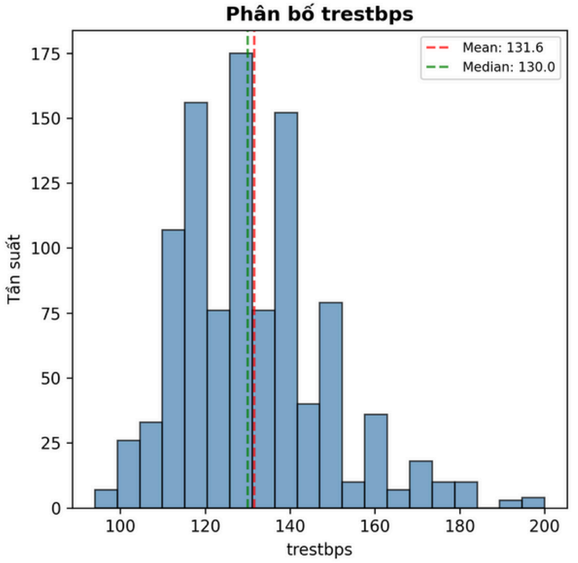
2.trestbps: Huyết áp lúc nghỉ (resting blood pressure), đơn vị mm Hg.

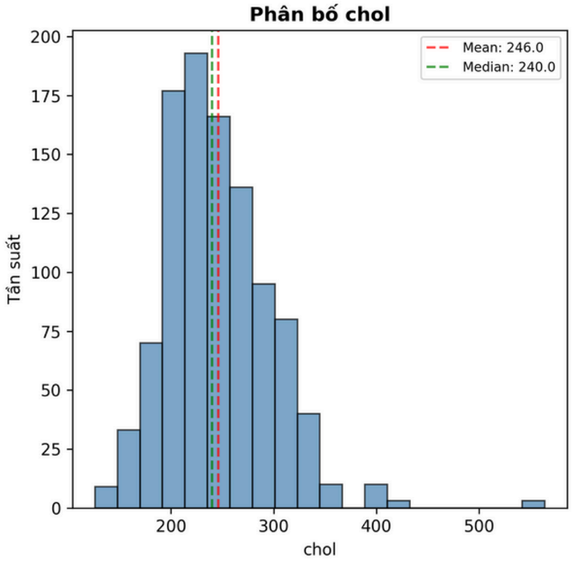
3.chol: Mức cholesterol huyết thanh (serum cholesterol) (mg/dl).

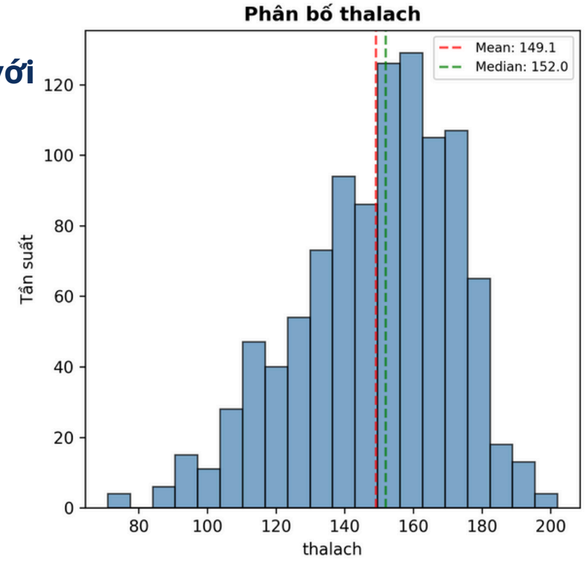
4.thalach: Nhịp tim tối đa đạt được khi gắng sức (maximum heart rate achieved).

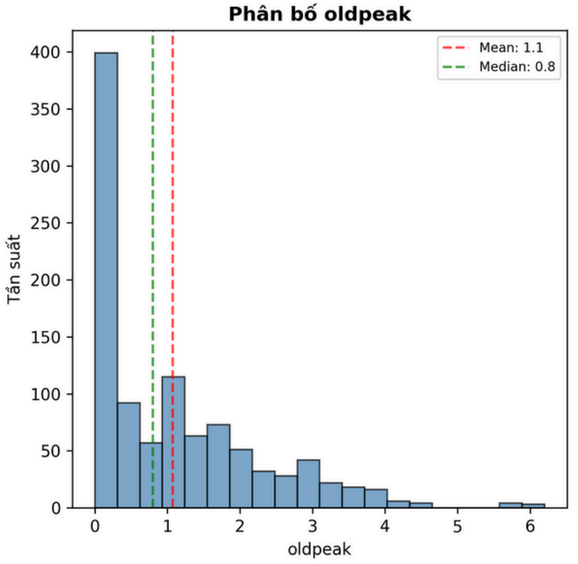
5.oldpeak: Độ chênh ST do gắng sức so với nghỉ ngơi.



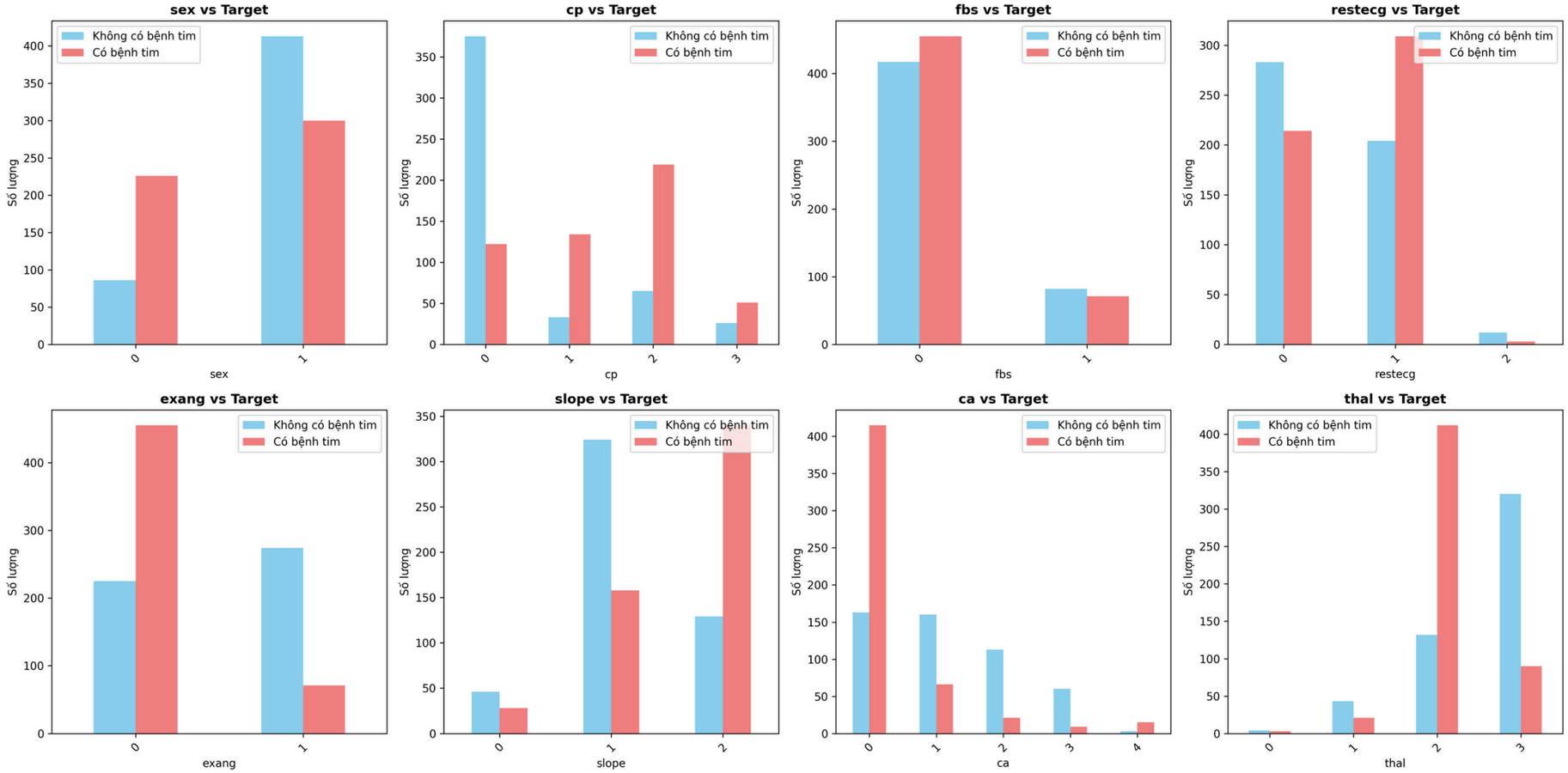








2. THU THẬP & KHÁM PHÁ DỮ LIỆU



Phân tích các biến phân loại (Categorical Cols)

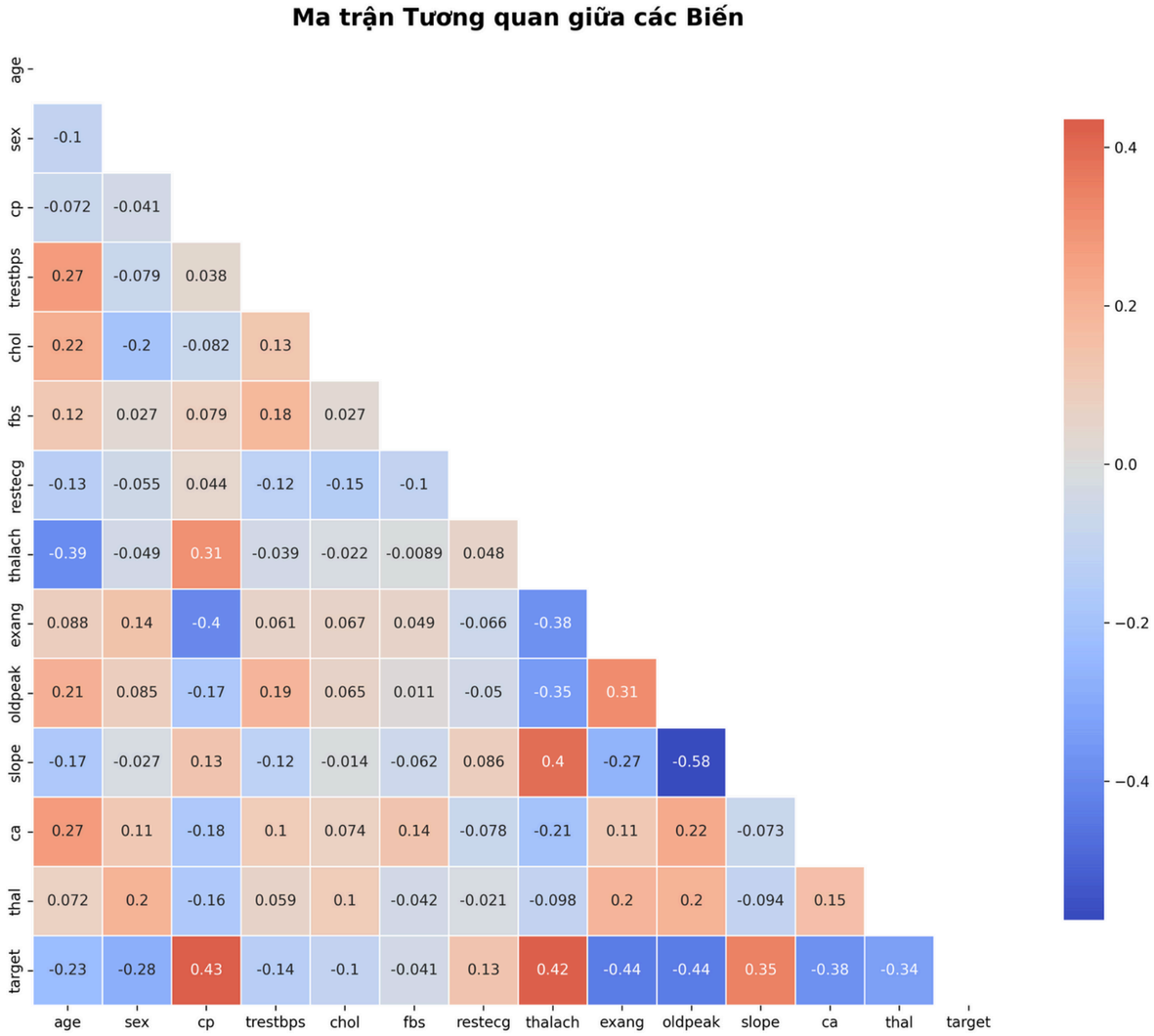
1. cp : Đau thắc ngực
2. fbs: Lượng đường huyết lúc đói > 120 mg/dl? (1 = Có, 0 = Không).
3.restecg:Kết quả điện tâm đồ lúc nghỉ (resting electrocardiographic results):
4. exang: Có bị đau thắt ngực do gắng sức không? (1 = Có, 0 = Không).
5.slope: Hình dạng dốc của đoạn ST trong bài test gắng sức:
6. ca: Số lượng mạch vành chính nhìn thấy khi chụp X-quang huỳnh quang (0-3).
7. thal: Kết quả test Thalassemia:

2. THU THẬP & KHÁM PHÁ DỮ LIỆU

Ma trận tương quan

Tương quan với biến Target (theo độ lớn):

oldpeak: 0.438
exang: 0.438
cp: 0.435
thalach: 0.423
ca: 0.382
slope: 0.346
thal: 0.338
sex: 0.280
age: 0.229
trestbps: 0.139
restecg: 0.134
chol: 0.100
fbs: 0.041



3. XỬ LÝ DỮ LIỆU

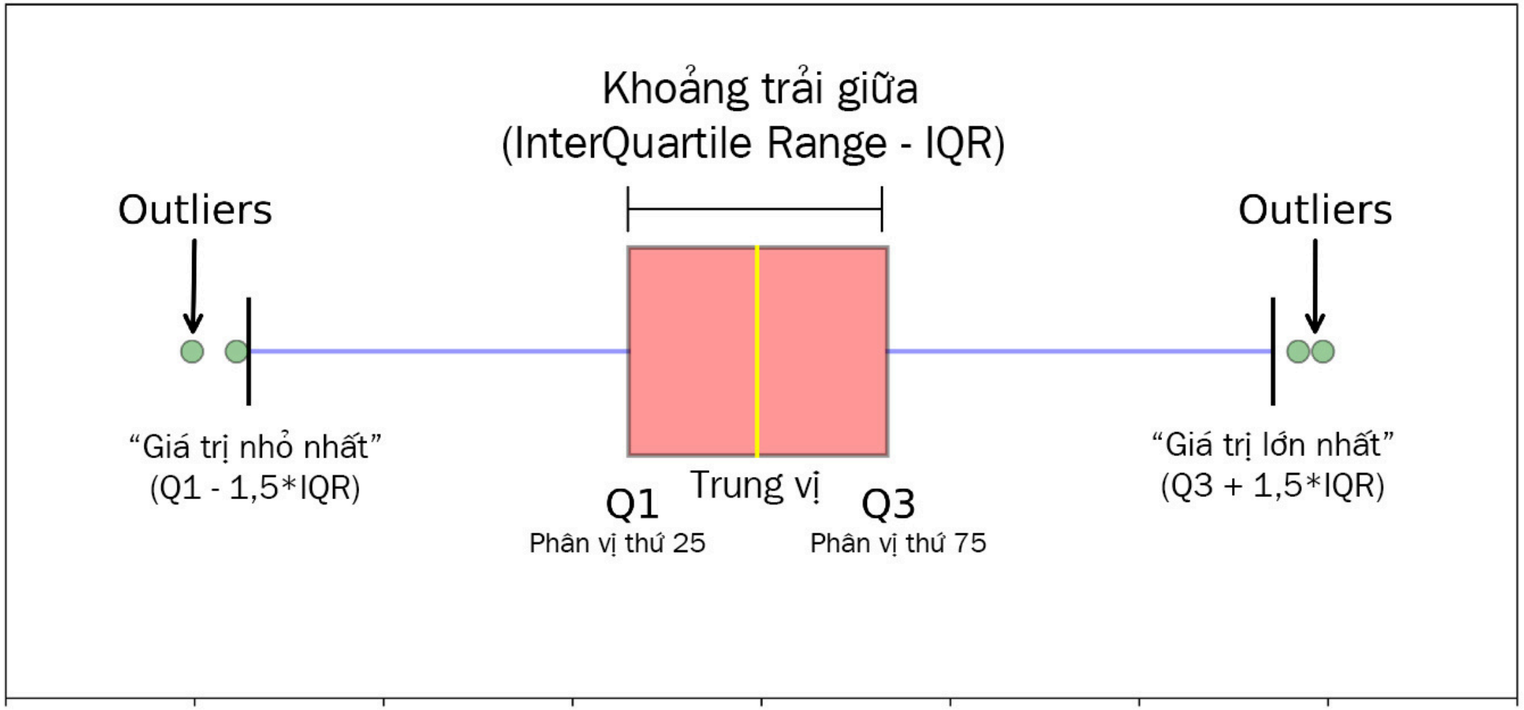
Kiểm tra và xử lý outliers

Chia tách & phân loại & chuẩn hoá

3. XỬ LÝ DỮ LIỆU

Kiểm tra và xử lý outliers

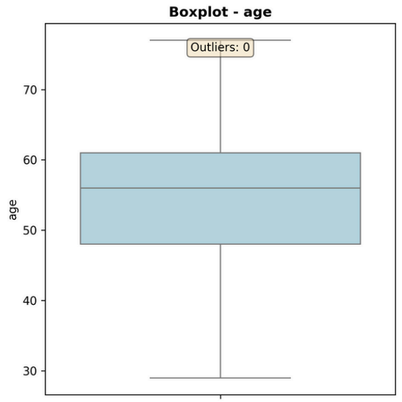
Sử dụng phương pháp IQR để kiểm tra Outliers

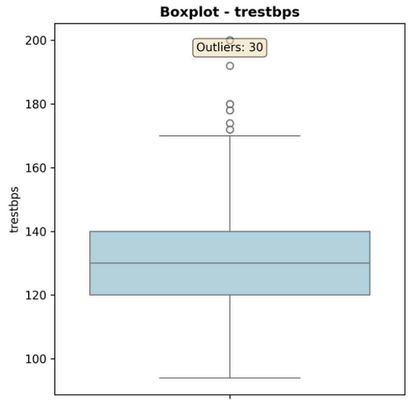


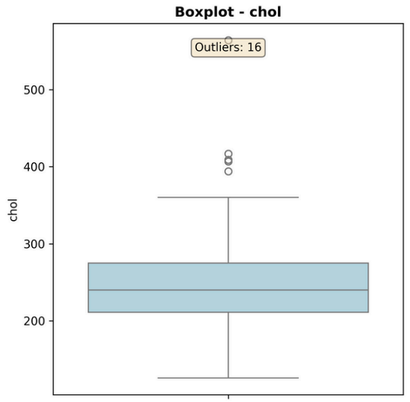
3. XỬ LÝ DỮ LIỆU

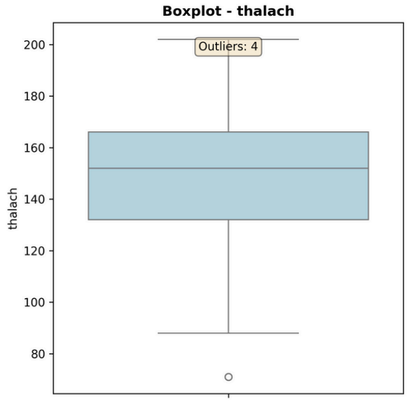
Kiểm tra và xử lý outliers

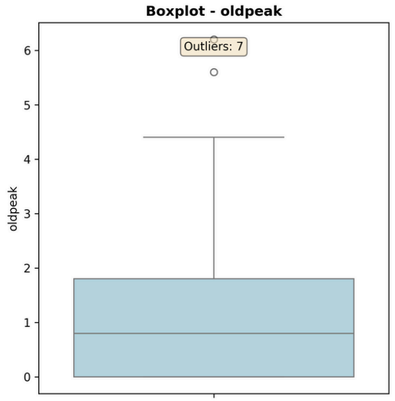
Sử dụng phương pháp IQR để kiểm tra Outliers











Số lượng outliers đã được xử lý:

age: 0 outliers

trestbps: 30 outliers

chol: 16 outliers

thalach: 4 outliers

oldpeak: 7 outliers

3. XỬ LÝ DỮ LIỆU

Phân loại các Features

Dựa vào tính chất, phân chia các feature thành 3 nhóm:

1. Numerical Features (age,trestbps, chol, thalach, oldpeak)
2. Ordinal Features (cp , restecg, slope, ca, thal)
3. Nominal Features (sex, fbs, exang)

Chuẩn hoá dữ liệu

Sử dụng các thư viện:

1. OrdinalEncoder
2. StandardScaler
3. ColumnTransformer

Chia tách dữ liệu

Chia tách dữ liệu thành các train set và test set

Tỉ lệ chia tách train set : test set là 8:2

Kết quả:
Kích thước tập train: (820, 13)
Kích thước tập test: (205, 13)
Tỷ lệ target trong train: {1: 0.513, 0: 0.487}
Tỷ lệ target trong test: {1: 0.512, 0: 0.488}

3. XỬ LÝ DỮ LIỆU

So sánh tương quan với target

So sánh tương quan với Target:

Feature	Before	After	Change
oldpeak	0.438	0.453	+0.014
exang	0.438	0.424	-0.014
cp	0.435	0.402	-0.033
thalach	0.423	0.450	+0.027
ca	0.382	0.373	-0.009
slope	0.346	0.338	-0.008
thal	0.338	0.299	-0.039
sex	0.280	0.289	+0.009
age	0.229	0.244	+0.014
trestbps	0.139	0.117	-0.022
restecg	0.134	0.116	-0.019
chol	0.100	0.139	+0.039
fbs	0.041	0.064	+0.023

4. HUẤN LUYỆN MÔ HÌNH

1. Logistic Regression

- **Là thuật toán học có giám sát dùng để dự đoán biến phân loại.**
- **Thường áp dụng trong phân loại nhị phân (ví dụ: có bệnh/không, spam/không spam).**
- **Đầu ra là một giá trị xác suất từ 0 đến 1.**

- **Sử dụng hàm sigmoid để biến đổi đầu ra tuyến tính thành xác suất :**
  σ(z) = 1 / (1 + e^(-z))
  σ(z) ≥ 0.5 → dự đoán lớp 1 (positive)
  σ(z) < 0.5 → dự đoán lớp 0 (negative)

- **Sử dụng hàm mất mát Binary Cross-Entropy:**
  L = - [y * log(ŷ) + (1 - y) * log(1 - ŷ)]

4. HUẤN LUYỆN MÔ HÌNH

2. Random Forest

- Random Forest là một tập hợp của nhiều cây quyết định (Decision Trees).
- Kết quả đầu ra được xác định bằng bỏ phiếu đa số (với phân loại) hoặc trung bình (với hồi quy).
- Là thuật toán thuộc nhóm ensemble learning - học kết hợp để tăng độ chính xác.

Nhược điểm:

1. Khó giải thích hơn một cây đơn.
2. Tốn tài nguyên hơn về thời gian và bộ nhớ.
3. Dễ bị chậm với tập dữ liệu rất lớn (nhiều cây hoặc nhiều đặc trưng).

Ưu điểm:

1. Độ chính xác cao, nhất là với dữ liệu phi tuyến.
2. Giảm overfitting nhờ voting trung bình.
3. Hoạt động tốt với dữ liệu có nhiều đặc trưng (features).
4. Tự động xử lý missing values tốt (tuỳ thư viện).

4. HUẤN LUYỆN MÔ HÌNH

3. Support Vector Machine

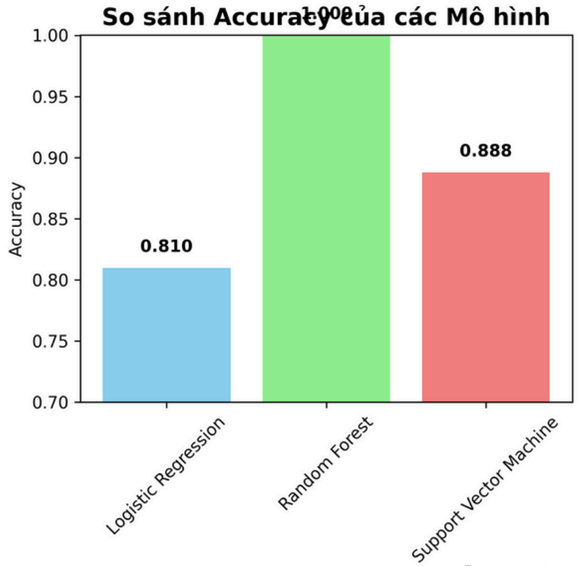
- SVM là một thuật toán học có giám sát được sử dụng cho phân loại và hồi quy.
- Mục tiêu chính: tìm siêu phẳng (hyperplane) tối ưu để phân tách các lớp dữ liệu.
- Tối ưu hóa khoảng cách (margin) giữa siêu phẳng và các điểm dữ liệu gần nhất của mỗi lớp.

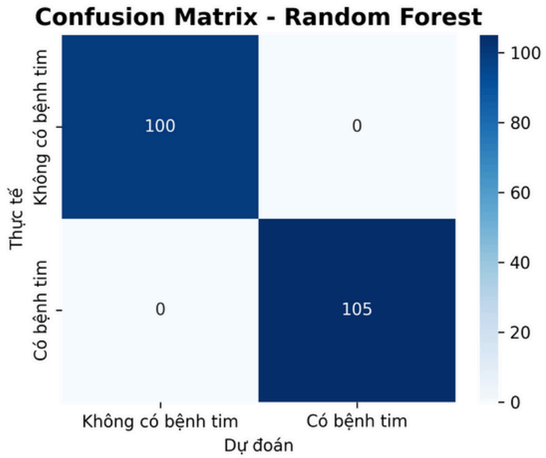
- Support Vectors: Các điểm dữ liệu gần siêu phẳng nhất, đóng vai trò quyết định biên phân chia.
- Margin: Khoảng cách từ siêu phẳng đến các support vectors - càng lớn càng tốt.
- Hyperplane: Đường/tập mặt phân chia dữ liệu trong không gian đặc trưng.

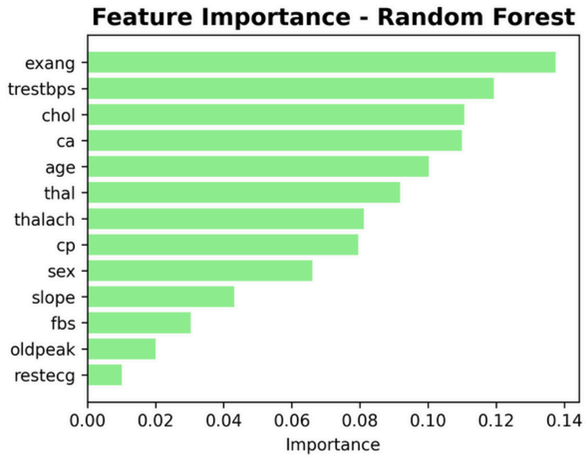
4. HUẤN LUYỆN MÔ HÌNH

Logistic Regression	>>>>>	Accuracy: 0,8098
Random Forest	>>>>>	Accuracy: 1,0000
Support Vector Machine	>>>>>	Accuracy: 0,8878

5. KẾT LUẬN







- Logistic Regression đạt 0.8098 – thể hiện hiệu suất ổn định, dễ triển khai và giải thích, phù hợp với các bài toán phân loại tuyến tính.
- Support Vector Machine (SVM) đạt 0.8878 – cho thấy khả năng phân loại tốt hơn Logistic Regression, đặc biệt hiệu quả trong các không gian có chiều cao.
- Random Forest đạt 1.0000 (100%) – hiệu suất hoàn hảo trên tập kiểm tra, cho thấy mô hình rất mạnh, có thể học tốt cấu trúc dữ liệu, nhưng cần kiểm tra thêm khả năng overfitting nếu dữ liệu huấn luyện và kiểm tra chưa đủ đa dạng.

THANK YOU!