

**TRƯỜNG ĐẠI HỌC SƯ PHẠM KỸ THUẬT TP.HCM KHOA CÔNG NGHỆ THÔNG TIN**

\*\*\*\*\*\*\*\*

**BÁO CÁO GIỮA KỲ**

**MÔN HỌC: NHẬP MÔN KHOA HỌC DỮ LIỆU**

**BÁO CÁO VỀ TẬP DỮ LIỆU NGHIÊN CỨU DỰ ĐOÁN BỆNH TIM**

**GVHD: ThS. Lê Minh Tân**

**Lớp: Sáng thứ 5**

**Sinh viên thực hiện:**

**Hồng Anh Khoa - 22110351**

**Trần Vũ Khanh - 22110349**

**Trịnh Tấn Hào - 22110315**

**Tp. Hồ Chí Minh, ngày 26 tháng 3 năm 2025**

**MỤC LỤC**

[PHẦN 1: GIỚI THIỆU VỀ DỮ LIỆU 1](#_Toc193906906)

[1.1. Phân tích Bộ dữ liệu Heart Failure Prediction Dataset (Fedesoriano): 1](#_Toc193906907)

[1.1.1 Nguồn gốc và Bối cảnh của Bộ dữ liệu: 1](#_Toc193906908)

[1.1.2 Mô tả Chi tiết về các Biến: 2](#_Toc193906909)

[1.2. Khám phá Chi tiết về Bộ dữ liệu Bệnh tim (David Lapp): 3](#_Toc193906910)

[1.2.1 Bối cảnh Bộ dữ liệu: 3](#_Toc193906911)

[1.2.2 Mô tả Chi tiết về các Biến: 4](#_Toc193906912)

[1.3. Nghiên cứu về Bộ dữ liệu Dự đoán Bệnh tim (Rashad Mammadov): 4](#_Toc193906913)

[1.3.1 Tổng quan về Bộ dữ liệu: 4](#_Toc193906914)

[1.3.2 Mô tả Chi tiết về các Biến: 5](#_Toc193906915)

[PHẦN 2: PHÂN TÍCH, THỐNG KÊ DỮ LIỆU SAU KHI XỬ LÝ 6](#_Toc193906916)

[2.1. Thông tin cơ bản: 6](#_Toc193906917)

[2.2. Phân tích các đặc trưng: 8](#_Toc193906918)

[PHẦN 3: NGHIÊN CỨU KIẾN THỨC LIÊN QUAN 15](#_Toc193906919)

[3.1. Hiểu về suy tim 15](#_Toc193906920)

[3.2. Yếu tố nguy cơ gây ra bệnh tim 15](#_Toc193906921)

[3.3. Các Khái niệm Khoa học Dữ liệu Cốt lõi cho Mô hình Dự đoán: 16](#_Toc193906922)

[3.3.1. Phân loại trong Khoa học Dữ liệu: 16](#_Toc193906923)

[3.3.2. Các Thuật toán Phân loại Chính: 16](#_Toc193906924)

[KẾT LUẬN 18](#_Toc193906925)

[TÀI LIỆU THAM KHẢO 18](#_Toc193906926)

# PHẦN 1: GIỚI THIỆU VỀ DỮ LIỆU

Bệnh tim mạch (Cardiovascular disease) là nguyên nhân gây tử vong hàng đầu trên toàn cầu, ước tính có khoảng 17,9 triệu ca tử vong mỗi năm. Con số đáng kinh ngạc này nhấn mạnh gánh nặng sức khỏe to lớn liên quan đến các tình trạng này.

Bốn trên năm ca tử vong do CVD là do nhồi máu cơ tim và đột quỵ, với một tỷ lệ đáng kể xảy ra sớm ở những người dưới 70 tuổi. Tỷ lệ tử vong sớm này cho thấy sự cần thiết của các chiến lược can thiệp sớm hiệu quả.

**Nhận định:** Quy mô tử vong do CVD cho thấy tầm quan trọng của nghiên cứu trong lĩnh vực này. Số lượng ca tử vong cao và thực tế là nhiều ca xảy ra trước tuổi 70 nhấn mạnh nhu cầu cấp thiết về các biện pháp phòng ngừa, phát hiện sớm và quản lý tốt hơn đối với các bệnh tim mạch.

**Tiềm năng của Khoa học Dữ liệu và Học máy trong chuẩn đoán bệnh tim mạch:**

Các mô hình học máy có thể vô giá trong việc phát hiện sớm và quản lý những người có nguy cơ tim mạch cao. Các mô hình này có thể phân tích các mẫu phức tạp trong dữ liệu bệnh nhân để xác định những người có nguy cơ.

Các mô hình này có thể phân tích các yếu tố nguy cơ khác nhau như tăng huyết áp, tiểu đường và tăng lipid máu để dự đoán khả năng mắc bệnh tim. Bằng cách tận dụng các yếu tố này, học máy có thể hỗ trợ đánh giá rủi ro cá nhân hóa.

**Nhận định:** Khoa học dữ liệu cung cấp các công cụ mạnh mẽ để phân tích dữ liệu y tế phức tạp và xây dựng các mô hình dự đoán. Các thuật toán học máy có thể xử lý lượng lớn dữ liệu và xác định các mối quan hệ tinh tế giữa các chỉ số sức khỏe khác nhau và khả năng phát triển bệnh tim, có khả năng dẫn đến chẩn đoán sớm hơn và các biện pháp phòng ngừa hiệu quả hơn.

## 1.1. Phân tích Bộ dữ liệu Heart Failure Prediction Dataset (Fedesoriano):

### 1.1.1 Nguồn gốc và Bối cảnh của Bộ dữ liệu:

* + - **Nguồn trên Kaggle:** Bộ dữ liệu "Heart Failure Prediction Dataset" có sẵn trên Kaggle dưới tên người dùng "Fedesoriano".
    - **Ngày tạo:** Bộ dữ liệu được tạo vào tháng 9 năm 2021.
    - **Liên kết:** <https://www.kaggle.com/datasets/fedesoriano/heart-failure-prediction>
    - **Kết hợp các Bộ dữ liệu:** Bộ dữ liệu này độc đáo vì nó kết hợp năm bộ dữ liệu bệnh tim khác nhau, trước đây độc lập (Cleveland, Hungary, Switzerland, Long Beach VA và Stalog (Heart) Data Set).
    - **Số lượng Quan sát:** Bộ dữ liệu kết hợp ban đầu có 1190 quan sát, đã giảm xuống còn 918 sau khi loại bỏ 272 bản sao.
    - **Biến mục tiêu:** Heart disease: 1 (bệnh tim), 0 (bình thường)
    - **Kho lưu trữ Nguồn:** Các bộ dữ liệu gốc đến từ Kho lưu trữ Học máy UCI. <https://archive.ics.uci.edu/dataset/45/heart+disease>

### 1.1.2 Mô tả Chi tiết về các Biến:

Bộ dữ liệu chứa 11 đặc trưng (biến đầu vào) và 1 lớp đầu ra (biến mục tiêu). Các biến này đại diện cho các đặc điểm lâm sàng và nhân khẩu học khác nhau của bệnh nhân.

Dữ liệu bao gồm các biến liên quan đến bệnh tim của bệnh nhân, cụ thể như sau:

• Tuổi (Age): Biến liên tục, đo bằng năm.

• Giới tính (Sex): Biến phân loại nhị phân (M: Nam, F: Nữ).

• Loại đau ngực (ChestPainType): Biến phân loại danh nghĩa với 4 loại (TA: Đau thắt ngực điển hình(1), ATA: Không điển hình(2), NAP: Không do đau thắt ngực(3), ASY: Không triệu chứng(0)).

• Huyết áp lúc nghỉ (RestingBP): Biến liên tục, đo bằng mm Hg.

• Mức cholesterol (Cholesterol): Biến liên tục, đo bằng mg/dl.

• Mức đường huyết lúc đói (FastingBS): Biến nhị phân (1: >120 mg/dl, 0: ngược lại).

• Kết quả điện tâm đồ lúc nghỉ (RestingECG): Biến phân loại danh nghĩa với 3 loại (Normal: Bình thường, ST: Bất thường sóng ST-T, LVH: Phì đại thất trái).

• Nhịp tim tối đa đạt được (MaxHR): Biến liên tục, giá trị từ 60 đến 202.

• Đau thắt ngực do gắng sức (ExerciseAngina): Biến nhị phân (Y: Có, N: Không).

• ST chênh xuống do gắng sức (Oldpeak): Biến liên tục.

• Độ dốc đoạn ST đỉnh gắng sức (ST\_Slope): Biến phân loại thứ tự với 3 mức độ (Up: Dốc lên, Flat: Nằm ngang, Down: Dốc xuống).

• Biến mục tiêu (HeartDisease): Xác định sự hiện diện của bệnh tim (1: Bệnh, 0: Bình thường).

## 1.2. Khám phá Chi tiết về Bộ dữ liệu Bệnh tim (David Lapp):

### 1.2.1 Bối cảnh Bộ dữ liệu:

* + - **Nguồn trên Kaggle:** Bộ dữ liệu "Heart Disease Dataset" có sẵn trên Kaggle dưới tên người dùng "**David Lapp**".
    - **Liên kết:** <https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset>
    - **Nguồn gốc Lịch sử:** Bộ dữ liệu này được cập nhật trên Kaggle năm 2019 và có ý nghĩa lịch sử vì nó có từ năm 1988 và là một tập hợp dữ liệu từ bốn địa điểm địa lý khác nhau: Cleveland, Hungary, Switzerland và Long Beach V.
    - **Số lượng Thuộc tính:** Bộ dữ liệu gốc chứa một tập hợp toàn diện gồm 76 thuộc tính. Tuy nhiên, do hầu hết các nghiên cứu tập trung vào một tập hợp có thể quản lý được, nên tất cả các thí nghiệm đã công bố thường đề cập đến việc sử dụng một tập con gồm 14 thuộc tính chính, bao gồm cả thuộc tính được dự đoán.
    - **Biến Mục tiêu:** Trường "target" cho biết sự hiện diện của bệnh tim ở bệnh nhân 1 (bệnh tim), 0 (bình thường).
    - **Nhận định:** Bộ dữ liệu này, mặc dù đã cũ, vẫn là một nguồn tài nguyên có giá trị do các nghiên cứu sâu rộng đã được thực hiện trên tập con gồm 14 thuộc tính được xác định rõ ràng của nó.

### 1.2.2 Mô tả Chi tiết về các Biến:

Phiên bản thường được sử dụng của bộ dữ liệu này bao gồm 14 thuộc tính, bao gồm các phép đo lâm sàng và đặc điểm bệnh nhân khác nhau, trong đó một thuộc tính là biến mục tiêu quan trọng.

Dữ liệu bao gồm các biến liên quan đến bệnh tim của bệnh nhân, cụ thể như sau:

* **Tuổi (age):** Biến liên tục, đo bằng năm.
* **Giới tính (sex):** Biến nhị phân (1: Nam, 0: Nữ).
* **Loại đau ngực (cp):** Biến phân loại với 4 mức độ.
* **Huyết áp lúc nghỉ (trestbps):** Biến liên tục, đo bằng mm Hg.
* **Mức cholesterol (chol):** Biến liên tục, đo bằng mg/dl.
* **Mức đường huyết lúc đói (fbs):** Biến nhị phân (1: >120 mg/dl, 0: ngược lại).
* **Kết quả điện tâm đồ lúc nghỉ (restecg):** Biến phân loại với 3 mức độ.
* **Nhịp tim tối đa đạt được (thalach):** Biến liên tục.
* **Đau thắt ngực do gắng sức (exang):** Biến nhị phân (1: Có, 0: Không).
* **ST chênh xuống do gắng sức (oldpeak):** Biến liên tục.
* **Độ dốc đoạn ST đỉnh gắng sức (slope):** Biến phân loại với 3 mức độ.
* **Số mạch máu chính được tô màu (ca):** Biến rời rạc (0-3).
* **Tình trạng thalassemia (thal):** Biến phân loại với 3 mức độ.
* **Biến mục tiêu (target):** Xác định bệnh tim (1: Bệnh, 0: Không bệnh).

## 1.3. Nghiên cứu về Bộ dữ liệu Dự đoán Bệnh tim (Rashad Mammadov):

### 1.3.1 Tổng quan về Bộ dữ liệu:

* + - **Nguồn trên Kaggle:** Bộ dữ liệu "Heart Disease Prediction Dataset" có sẵn để sử dụng công khai trên Kaggle dưới tên người dùng "Rashad Mammadov".
    - **Ngày tạo:** Bộ dữ liệu được tạo vào tháng 6 năm 2024.
    - **Liên kết:** <https://www.kaggle.com/datasets/rashadrmammadov/heart-disease-prediction>
    - **Mô tả Chung:** Bộ dữ liệu này cung cấp một tập hợp thông tin liên quan đến các cá nhân và các yếu tố nguy cơ khác nhau góp phần vào sự phát triển của bệnh tim. Nó bao gồm các chi tiết nhân khẩu học, tiền sử bệnh, lựa chọn lối sống và các triệu chứng được báo cáo liên quan đến bệnh tim.
    - **Biến Mục tiêu:** Biến mục tiêu quan trọng trong bộ dữ liệu này được gắn nhãn "Heart Disease" và nó đóng vai trò là một chỉ số cho biết liệu một cá nhân đã được chẩn đoán mắc bệnh tim (giá trị 1) hay chưa (giá trị 0).
    - **Nhận định:** Bộ dữ liệu này cung cấp một phạm vi các yếu tố nguy cơ tiềm ẩn rộng hơn, đặc biệt bao gồm các lựa chọn lối sống như hút thuốc và uống rượu, điều này có thể cung cấp các quan điểm và hiểu biết khác nhau so với hai bộ dữ liệu đầu tiên tập trung nhiều hơn vào các phép đo lâm sàng.

### 1.3.2 Mô tả Chi tiết về các Biến:

Bộ dữ liệu này bao gồm tổng cộng 16 biến, cung cấp một tập hợp các đặc trưng phong phú cho việc phân tích và mô hình hóa dự đoán .

Dữ liệu bao gồm các biến sức khỏe của cá nhân, cụ thể như sau:

* **Tuổi (Age):** Biến liên tục, đo bằng năm.
* **Giới tính (Gender):** Biến phân loại (Nam, Nữ).
* **Mức cholesterol (Cholesterol):** Biến liên tục, đo bằng mg/dL.
* **Huyết áp (Blood Pressure):** Biến liên tục, đo bằng mmHg.
* **Nhịp tim (Heart Rate):** Biến liên tục, đo bằng nhịp mỗi phút (bpm).
* **Tình trạng hút thuốc (Smoking):** Biến phân loại với 3 mức độ (Chưa bao giờ, Đã từng, Hiện tại).
* **Tần suất uống rượu (Alcohol Intake):** Biến phân loại với 3 mức độ (Không, Vừa phải, Nhiều).
* **Số giờ tập thể dục (Exercise Hours):** Biến liên tục, đo bằng số giờ mỗi tuần.
* **Tiền sử gia đình mắc bệnh tim (Family History):** Biến nhị phân (Có, Không).
* **Chẩn đoán tiểu đường (Diabetes):** Biến nhị phân (Có, Không).
* **Béo phì (Obesity):** Biến nhị phân (Có, Không).
* **Mức độ căng thẳng (Stress Level):** Biến số rời rạc, đo trên thang điểm từ 1 đến 10.
* **Mức đường huyết (Blood Sugar):** Biến liên tục, đo bằng mg/dL.
* **Đau thắt ngực do gắng sức (Exercise Induced Angina):** Biến nhị phân (Có, Không).
* **Loại đau ngực (Chest Pain Type):** Biến phân loại với 4 mức độ (Đau thắt ngực điển hình, Không điển hình, Không do đau thắt ngực, Không triệu chứng).
* **Bệnh tim (Heart Disease - Target):** Biến mục tiêu xác định sự hiện diện của bệnh tim (0: Không, 1: Có).

**PHẦN 2: PHÂN TÍCH, THỐNG KÊ DỮ LIỆU SAU KHI XỬ LÝ**

Bộ dữ liệu này được xử lý bằng cách xử lý các bộ dữ liệu riêng biệt về 1 mẫu chung sau đó hợp lại thành 1 bộ hoàn chỉnh.

Các bước phân tích cơ bản:  
1. Đọc dữ liệu từ các file

2. Loại bỏ 2 cột không cần thiết ‘ca’, ‘thal’ của tập “**David Lapp”.** Kết hợp 2 tập “**Fedesoriano**” và “**David Lapp**” đồng bộ tên cột theo bộ “**David Lapp”.**

3. Xử lý dữ liệu thiếu của tập “**Rashad Mammadov**” bằng phương pháp KNN Imputer

4. Chuẩn hóa và đồng bộ 3 tập dữ liệu.

5. Kết hợp tập dữ liệu “**Rashad Mammadov”** với tập đã kết hợp trước đó.

6. Kiểm tra giá trị trùng lập.

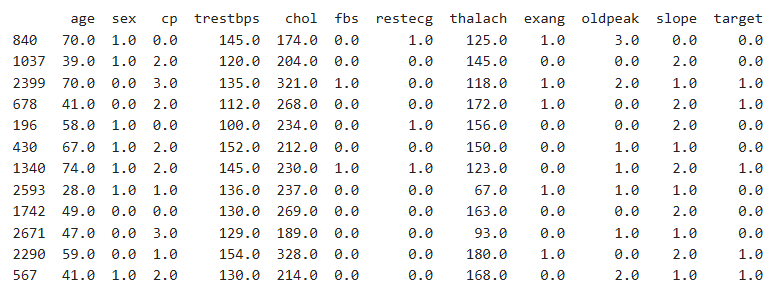
7. lấy 2000 mẫu ngẫu nhiên tạp thành dataset mới.

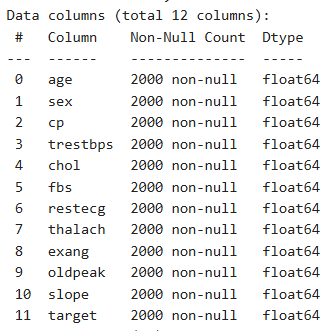
**2.1. Thông tin cơ bản:**

- Số lượng mẫu: 2000

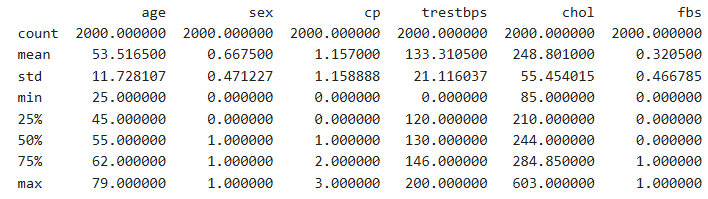
- Số đặc trưng: 12

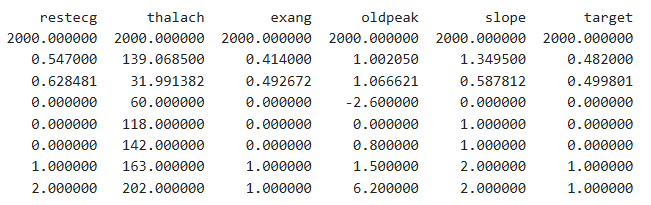
**\* Đọc 12 hàng đầu tiên:**





**\* Thống kê mô tả (đặc trưng số):**

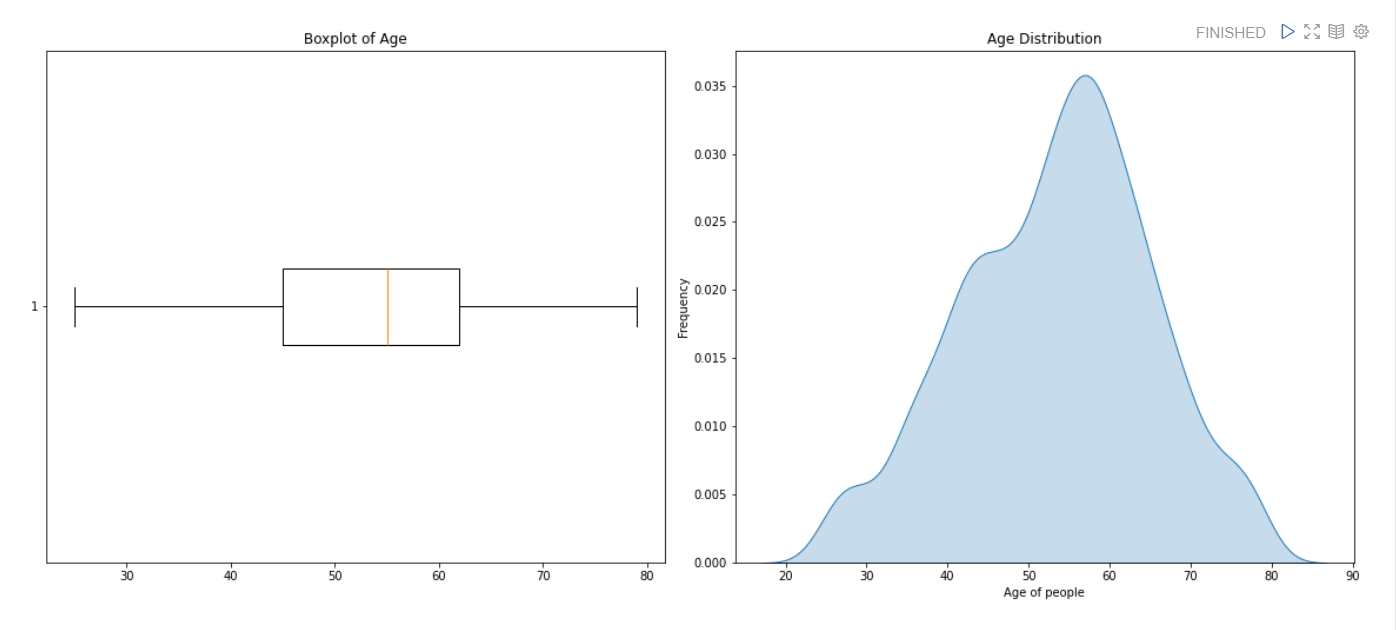




[9 rows x 13 columns]

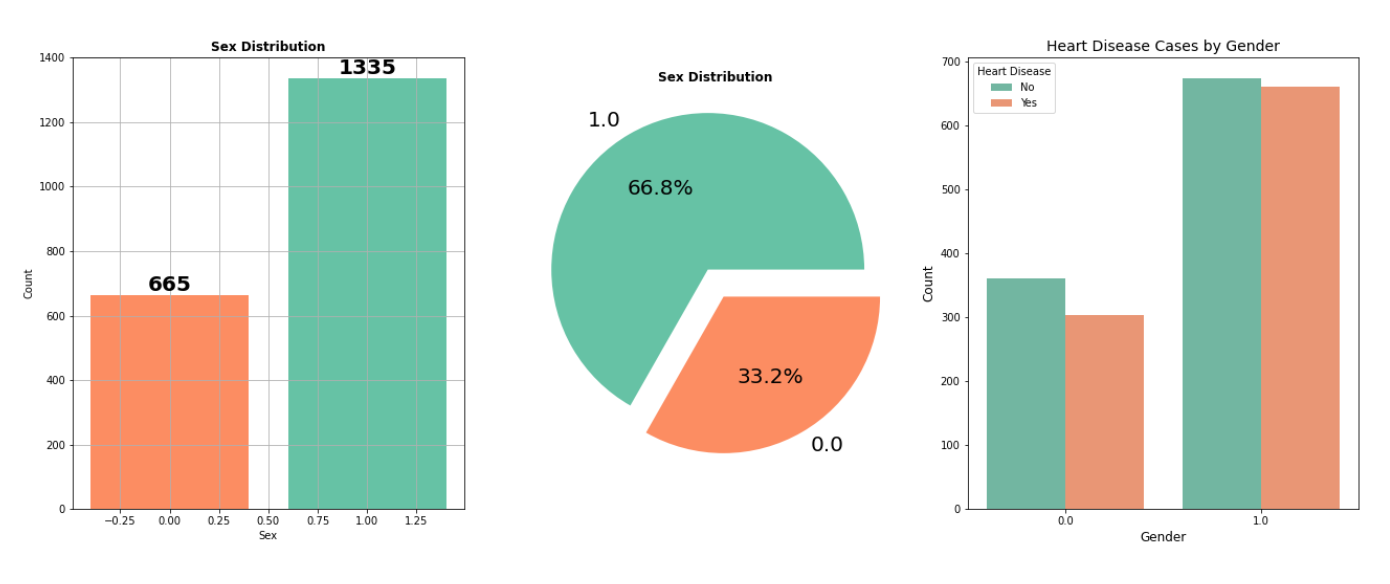
- Số dòng trùng lặp: 406

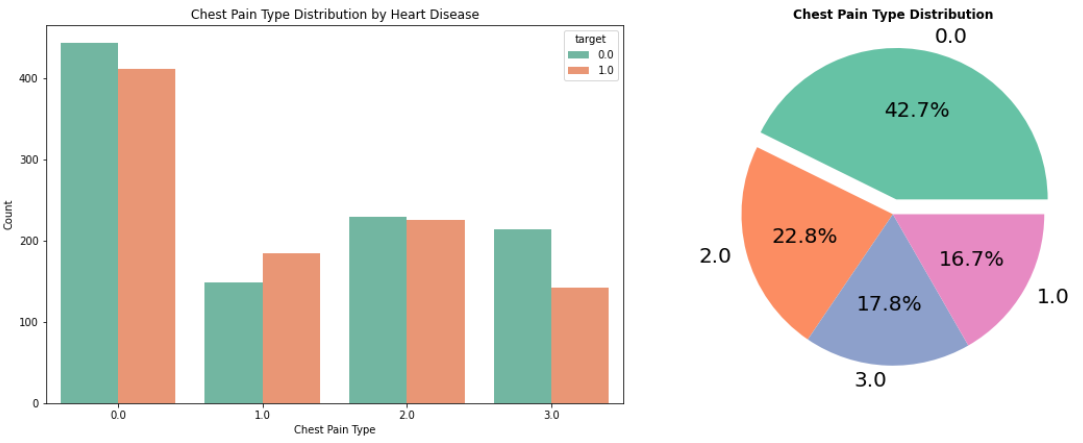
**2.2. Phân tích các đặc trưng:**

**\* Đặc trưng về độ tuổi (age):**

50% dữ liệu tập trung vào khoảng 45-60 tuổi.

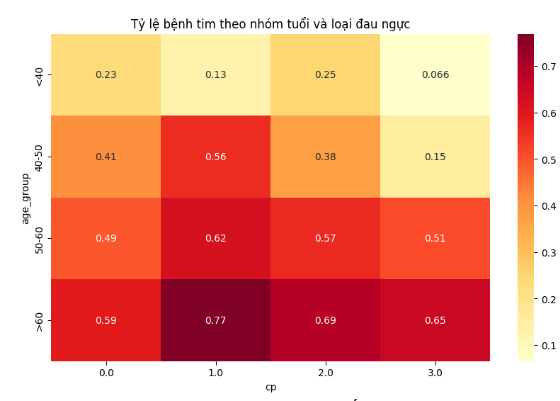
* Không có giá trị ngoại lệ lớn, dữ liệu tương đối cân bằng.
* Phân phối gần với phân phối chuẩn nhưng hơi lệch phải, cho thấy có một số người cao tuổi hơn mức trung bình.

 **\* Đặc trưng về giới tính (sex):**

* Có sự chênh lệch giới tính giữa nam và nữ, với số lượng nam nhiều gấp đôi nữ .
* Số người mắc bệnh tim gần bằng số người không mắc bệnh tim.
* **\* Đặc trưng về Loại đau ngực (cp):**
* ASY: Không triệu chứng(0),
* TA: Đau thắt ngực điển hình(1),
* ATA: Không điển hình(2),
* NAP: Không do đau thắt ngực(3).

ASY là loại phổ biến nhất, cho thấy nhiều bệnh nhân có vấn đề về tim mà không có triệu chứng rõ ràng.

**\* Tỷ lệ bệnh tim theo nhóm tuổi và loại đau ngực:**

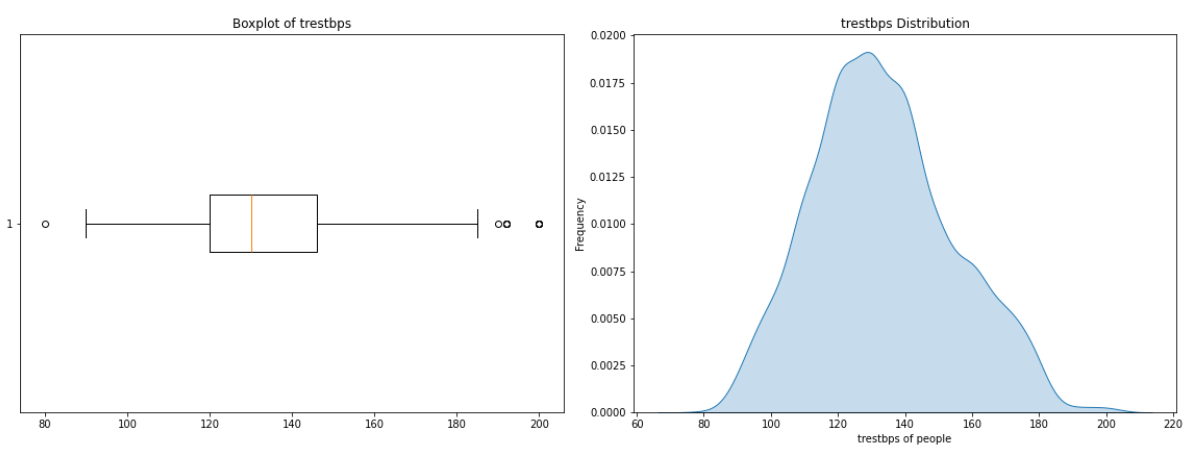
****

Xu hướng theo nhóm tuổi:

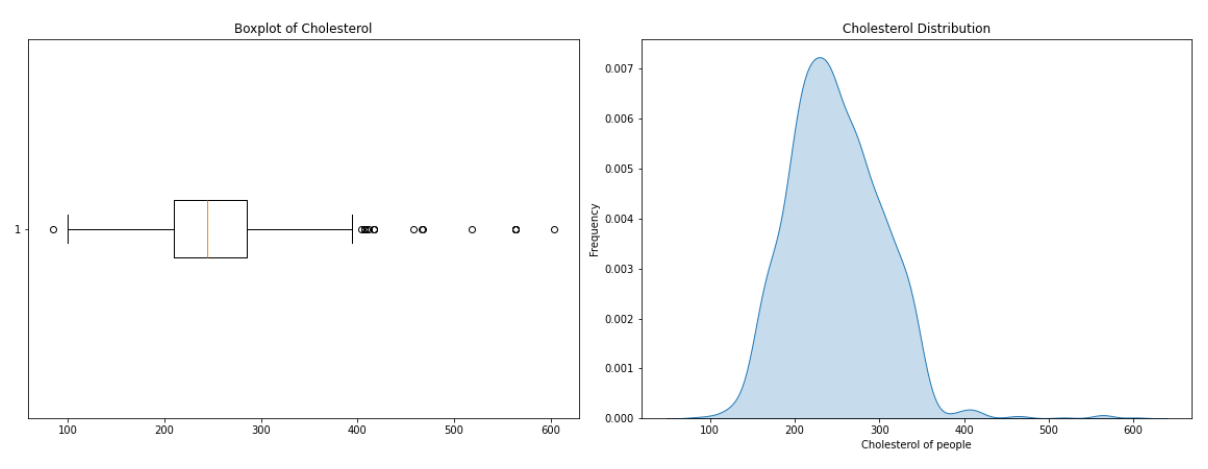
Tỷ lệ bệnh tim có xu hướng tăng dần theo độ tuổi, đặc biệt ở nhóm tuổi từ 40-50 (0.59) và 30-60 (0.77). Điều này phù hợp với nghiên cứu y khoa rằng nguy cơ mắc bệnh tim thường tăng theo tuổi tác.

Biến động tỷ lệ:

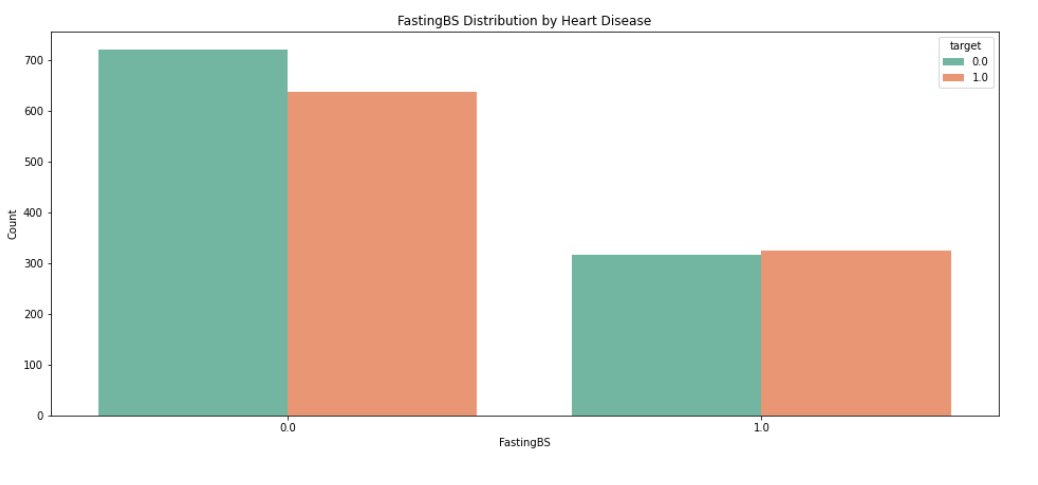
Có sự biến động đáng kể trong tỷ lệ bệnh tim, từ thấp nhất là 0.05 đến cao nhất là 2.0. Giá trị 2.0 có vẻ bất thường và cần kiểm tra lại tính hợp lệ của dữ liệu.

**\* Đặc trưng về Huyết áp lúc nghỉ ngơi (trestbps):**

* Huyết áp tâm thu lúc nghỉ ngơi (trestbps): Dữ liệu cho thấy huyết áp tâm thu lúc nghỉ ngơi của người trong mẫu có phân phối gần như chuẩn, với giá trị trung bình nằm trong khoảng 120-140.
* Điểm ngoại lệ: Có một số người có huyết áp tâm thu lúc nghỉ ngơi rất thấp hoặc rất cao,

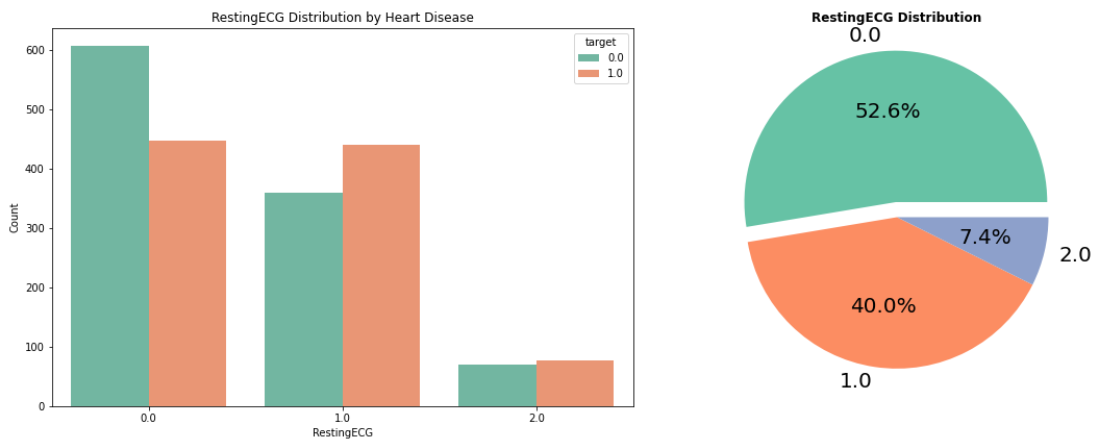
**\* Đặc trưng về Lượng Cholesterol (chol):**

* Dữ liệu cho thấy lượng cholesterol của nhóm người này có phân phối lệch phải, với nhiều người có mức cholesterol cao hơn mức trung bình.
* Có một số người có mức cholesterol rất cao, cần được chú ý.
* Mức cholesterol cao là một yếu tố nguy cơ của bệnh tim mạch.

**\* Đặc trưng về Lượng đường trong máu (fbs):**

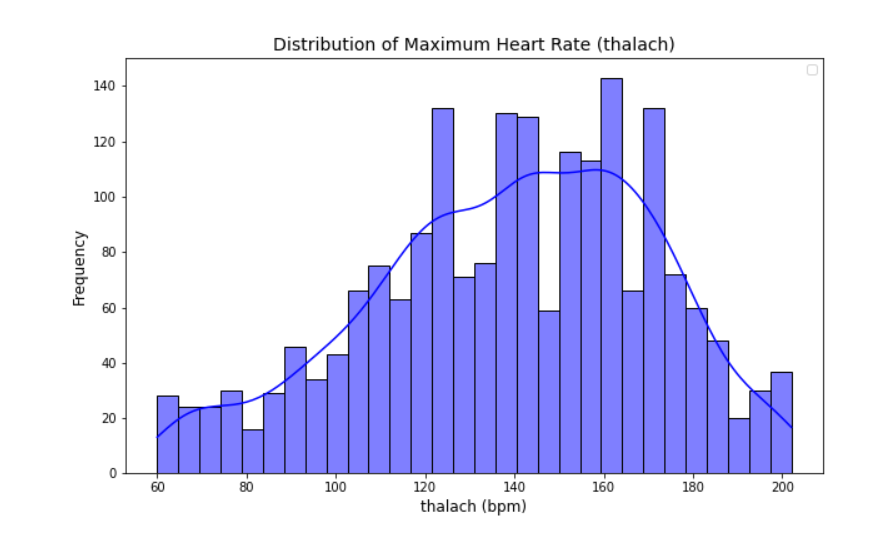
Đối với những người có FastingBS = 1 (đường huyết cao), số lượng người mắc bệnh và không mắc bệnh khá tương đồng, nhưng có xu hướng gần bằng nhau.

Đối với những người có FastingBS = 0 (đường huyết bình thường), số lượng người không mắc bệnh tim cao hơn một chút so với số người mắc bệnh.

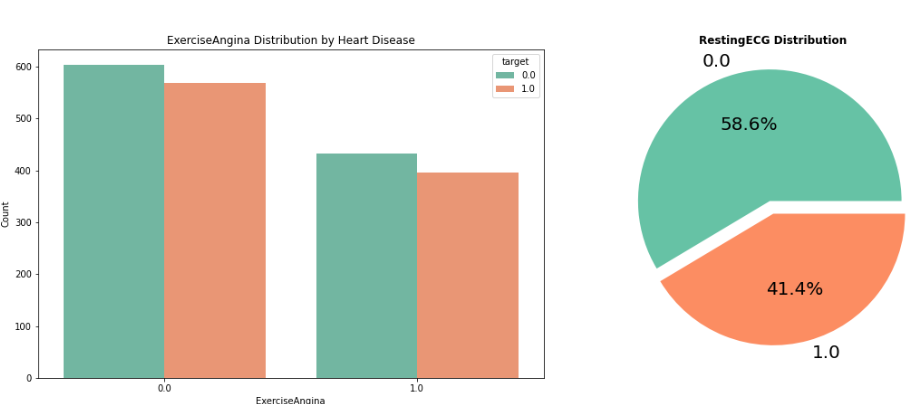
**\* Đặc trưng về Kết quả điện tâm đồ khi nghỉ ngơi (restecg):**

RestingECG = 0 và 1 chiếm phần lớn dữ liệu, trong đó nhóm không mắc bệnh nhiều hơn một chút ở cả hai trường hợp.

RestingECG = 2 ít gặp nhất (7.4%) nhưng có xu hướng liên quan nhiều hơn đến bệnh tim.

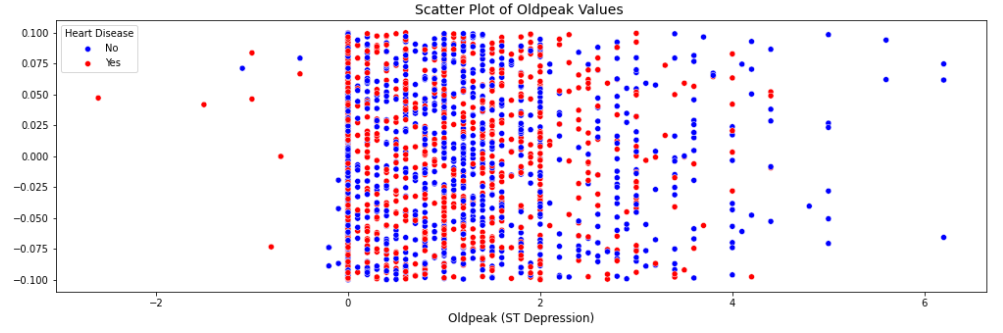
**\* Đặc trưng về Nhịp tim tối đa đạt được (thalach):**

Nhịp tim tối đa có phân bố chuẩn, cần kiểm tra thêm mối quan hệ với bệnh tim.

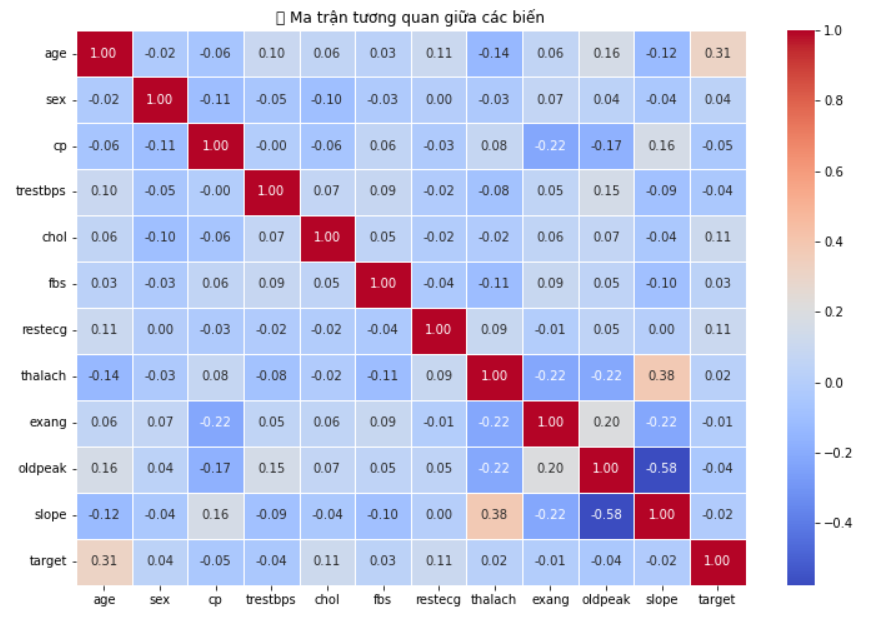
**\* Đặc trưng về Cơn đau thắt ngực do tập thể dục (exang):**

ExerciseAngina có ảnh hưởng đến nguy cơ mắc bệnh, nhóm có đau thắt ngực khi gắng sức có tỷ lệ mắc bệnh cao hơn.

RestingECG chủ yếu thuộc nhóm 0 và 1, có thể ảnh hưởng đến nguy cơ mắc bệnh nhưng cần phân tích sâu hơn.

**\* Đặc trưng về Mức độ chênh ST khi tập thể dục so với lúc nghỉ ngơi (oldpeak):**

Khi Oldpeak tăng (>2), số lượng điểm xanh (No) có vẻ chiếm ưu thế hơn điểm đỏ, cho thấy những người có mức ST Depression cao hơn thường không mắc bệnh tim.

**\*Phân tích tương quan giữa các biến:**

Dựa vào ma trận tương quan giữa các biến trong bộ dữ liệu chẩn đoán bệnh tim, chúng ta có thể đánh giá tính phù hợp của dữ liệu này cho việc huấn luyện mô hình dự đoán bệnh tim.

Quan sát cột "target" (biến mục tiêu), chúng ta thấy rằng biến "age" có mối tương quan dương nhẹ (0.31) với bệnh tim. Tiếp theo là "oldpeak"(độ chênh ST) (-0.05), "cp" (đau ngực) (-0.05), "thalach" (0.02), và một số biến khác có mức độ tương quan rất thấp hoặc gần như không đáng kể.

Mặc dù không có biến nào có mức độ tương quan vượt quá 0.5 với biến mục tiêu, nhưng giá trị tương quan dương và âm nhẹ giữa một số biến "oldpeak" và "slope" với "target" cho thấy vẫn có thể có mối quan hệ phi tuyến tính hoặc có sự kết hợp của nhiều yếu tố quyết định bệnh tim.

Bộ dữ liệu này có thể dùng để huấn luyện mô hình dự đoán bệnh tim nhưng cần tối ưu hóa đặc trưng và chọn mô hình phù hợp để đạt hiệu suất tốt.

# PHẦN 3: NGHIÊN CỨU KIẾN THỨC LIÊN QUAN

## 3.1. Hiểu về suy tim

**Định nghĩa**: Suy tim (CHF) là tình trạng tim không bơm đủ máu để đáp ứng nhu cầu cơ thể, không có nghĩa là tim ngừng đập mà chỉ bị suy giảm chức năng.

**Nguyên nhân**: Thường do bệnh động mạch vành, nhồi máu cơ tim, huyết áp cao, tiểu đường, béo phì, bệnh van tim. Một số nguyên nhân khác gồm dị tật tim bẩm sinh, nhiễm trùng, lối sống không lành mạnh.

**Các loại suy tim**: Ảnh hưởng đến tim trái, tim phải hoặc cả hai. Suy tim trái phổ biến hơn, gồm suy tim với phân suất tống máu giảm (HFrEF) và bảo tồn (HFpEF).

**Triệu chứng**: Khó thở, mệt mỏi, sưng chân, khó thở khi nằm, tăng cân, ho dai dẳng, nhịp tim bất thường.

## 3.2. Nguyên nhân và cách hạn chế bệnh tim

\*Nguyên nhân:

* + - **Tình trạng y tế**: Bệnh động mạch vành, nhồi máu cơ tim, huyết áp cao, tiểu đường, béo phì, bệnh van tim, rối loạn nhịp tim, dị tật tim bẩm sinh, bệnh thận, tuyến giáp, ngưng thở khi ngủ.
    - **Lối sống**: Hút thuốc, chế độ ăn uống kém, ít vận động, uống rượu nhiều, sử dụng ma túy.
    - **Yếu tố khác**: Tuổi cao, tiền sử gia đình, chủng tộc, yếu tố di truyền, Stress,..

\*Cách phòng ngừa:

Để phòng ngừa bệnh tim, cần duy trì chế độ ăn uống lành mạnh, tập thể dục thường xuyên, kiểm soát cân nặng, không hút thuốc, hạn chế rượu bia, kiểm tra sức khỏe định kỳ và giảm căng thẳng. Những biện pháp này giúp giảm nguy cơ mắc bệnh tim và cải thiện sức khỏe lâu dài.

## 3.3. Các Khái niệm Khoa học Dữ liệu Cốt lõi cho Mô hình Dự đoán:

### 3.3.1. Phân loại trong Khoa học Dữ liệu:

* + - **Định nghĩa:** Phân loại là một nhiệm vụ học máy có giám sát liên quan đến việc gán các điểm dữ liệu vào các danh mục hoặc lớp được xác định trước.
    - Mục tiêu là học một ánh xạ từ các đặc trưng đầu vào đến một nhãn đầu ra rời rạc.
    - **Tính phù hợp với Dự đoán Bệnh tim:** Dự đoán sự hiện diện hay vắng mặt của bệnh tim (kết quả nhị phân: 0 hoặc 1) là một bài toán phân loại.
    - **Các loại Phân loại:** Đề cập ngắn gọn đến phân loại nhị phân (hai lớp) và phân loại đa lớp (hơn hai lớp). Dự đoán bệnh tim trong bối cảnh này là nhị phân.
    - **Nhận định:** Hiểu phân loại như một khái niệm cốt lõi định hình nhiệm vụ dự đoán bệnh tim bằng cách sử dụng các bộ dữ liệu được cung cấp.

### 3.3.2. Các Thuật toán Phân loại Chính:

**Các Thuật toán Học máy Phổ biến:**

* + - **Hồi quy Logistic:** Thường được sử dụng như một mô hình cơ sở.
    - **Random forest (Random forest):** Thường đạt được độ chính xác cao và là một lựa chọn phổ biến.
    - **K-Nearest Neighbors (KNN):** Được khám phá như một kỹ thuật phân loại khác.
    - **Perceptron đa lớp (MLP):** Ngày càng được sử dụng vì khả năng mô hình hóa các mối quan hệ phi tuyến tính phức tạp.
    - **XGBoost:** Một thuật toán tăng cường độ dốc được biết đến với hiệu suất cao.
  + **Bảng 4: Tóm tắt các Bài báo Nghiên cứu về Dự đoán Bệnh tim**

|  |  |  |  |
| --- | --- | --- | --- |
| **Nghiên cứu (Tác giả, Năm)** | **Bộ dữ liệu đã sử dụng** | **Các Thuật toán Chính đã Áp dụng** | **Độ chính xác/Chỉ số Chính được Báo cáo** |
| Zeng (2023) 4 | Fedesoriano | Cây quyết định, KNN, SVM, XGBoost | SVM: Độ chính xác 88,8% |
| Alotaibi (2019) 5 | Tập con của Fedesoriano | Random Forest | Độ chính xác 89,14% |
| Scholarly Review Online (2023/2024) 5 | Fedesoriano | Random forest, XGBoost | RF: 92,90%, XGB: 91,56% Độ chính xác |
| Mohan et al15. | Cleveland UCI | RF lai và Phương pháp Tuyến tính | Độ chính xác 88,7% |
| Liu et al. (2025) 6 | Hồ sơ Sức khỏe Điện tử (Meta-analysis) | Random forest, Học sâu | RF: AUC 0,865, DL: AUC 0,847 |
| Dangare & Apte 7 | Cơ sở dữ liệu Bệnh tim | Mạng nơ-ron | Độ chính xác 100% |
| Malavika G. et al8. | Kho lưu trữ UCI | LR, KNN, SVM, NB, DT, RF | RF: Độ chính xác 91,80% |
| нагрузка 8 | Bệnh tim Nam Phi | LR, SVM, KNN, ANN | SVM: Độ chính xác 78,1% |
| обнаружение 8 | Không xác định | RF | Độ chính xác 100% |
| обнаружение 10 | Không xác định | Perceptron đa lớp | Độ chính xác 87,28% |

# KẾT LUẬN

Trong báo cáo này, chúng tôi đã tiến hành phân tích ba bộ dữ liệu liên quan đến dự đoán bệnh tim, bao gồm Heart Failure Prediction Dataset (Fedesoriano), Bộ dữ liệu Bệnh tim (David Lapp) và Dự đoán Bệnh tim (Rashad Mammadov). Việc khám phá chi tiết về nguồn gốc, bối cảnh và các biến của từng bộ dữ liệu giúp hiểu rõ đặc điểm của dữ liệu đầu vào, từ đó làm cơ sở cho các bước xử lý và phân tích sau này.

Thông qua quá trình phân tích và thống kê dữ liệu, chúng tôi đã xác định được những đặc trưng quan trọng ảnh hưởng đến bệnh tim và thực hiện các bước xử lý dữ liệu nhằm cải thiện chất lượng thông tin đầu vào. Các yếu tố nguy cơ, như tuổi tác, huyết áp, cholesterol và tiền sử bệnh lý, được xem xét để đánh giá mức độ ảnh hưởng đến nguy cơ mắc bệnh tim.

Bên cạnh đó, báo cáo cũng tập trung vào các kiến thức khoa học liên quan, bao gồm cơ chế suy tim, yếu tố nguy cơ và các thuật toán phân loại trong khoa học dữ liệu. Việc áp dụng các phương pháp học máy vào bài toán dự đoán bệnh tim giúp cung cấp những gợi ý quan trọng cho việc xây dựng mô hình dự đoán hiệu quả.

Tóm lại, nghiên cứu này góp phần làm rõ cách dữ liệu y tế có thể được phân tích và sử dụng để hỗ trợ dự đoán bệnh tim. Kết quả thu được có thể là nền tảng cho các nghiên cứu sâu hơn nhằm phát triển các mô hình chính xác và tối ưu hơn trong lĩnh vực chăm sóc sức khỏe.

# TÀI LIỆU THAM KHẢO

1. Heart Failure Prediction Dataset - Kaggle, truy cập vào tháng 3 24, 2025, [https://www.kaggle.com/datasets/Fedesoriano/heart-failure-prediction](https://www.kaggle.com/datasets/fedesoriano/heart-failure-prediction)
2. Heart Disease Dataset - Kaggle, truy cập vào tháng 3 24, 2025, <https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset>
3. Heart Disease prediction - Kaggle, truy cập vào tháng 3 24, 2025, <https://www.kaggle.com/datasets/rashadrmammadov/heart-disease-prediction>
4. (PDF) The Prediction of Heart Failure based on Four Machine Learning Algorithms, truy cập vào tháng 3 24, 2025, <https://www.researchgate.net/publication/369864504_The_Prediction_of_Heart_Failure_based_on_Four_Machine_Learning_Algorithms>
5. Prediction of Heart Failure Using Random Forest and XG Boost - Scholarly Review Journal, truy cập vào tháng 3 24, 2025, <https://www.scholarlyreview.org/api/v1/articles/121782-prediction-of-heart-failure-using-random-forest-and-xgboost.pdf>
6. Machine learning based prediction models for cardiovascular disease risk using electronic health records data: systematic review and meta-analysis - Oxford Academic, truy cập vào tháng 3 24, 2025, <https://academic.oup.com/ehjdh/article/6/1/7/7845948>
7. Improved Study of Heart Disease Prediction System using Data Mining Classification Techniques - ResearchGate, truy cập vào tháng 3 24, 2025, <https://www.researchgate.net/publication/258651311_Improved_Study_of_Heart_Disease_Prediction_System_using_Data_Mining_Classification_Techniques>
8. Prediction of Heart Disease Based on Machine Learning Using Jellyfish Optimization Algorithm - PMC, truy cập vào tháng 3 24, 2025, <https://pmc.ncbi.nlm.nih.gov/articles/PMC10378171/>
9. Machine Learning-Based Predictive Models for Detection of Cardiovascular Diseases - PMC, truy cập vào tháng 3 24, 2025, <https://pmc.ncbi.nlm.nih.gov/articles/PMC10813849/>
10. Effective Heart Disease Prediction Using Machine Learning Techniques - MDPI, truy cập vào tháng 3 24, 2025, <https://www.mdpi.com/1999-4893/16/2/88>