DỰ ĐOÁN BỆNH TIM SỬ DỤNG MÁY HỌC

Hoàng Vũ
Huỳnh Thanh Bình
Nguyễn Minh Tú
Phạm Tấn Khương

Nội dung

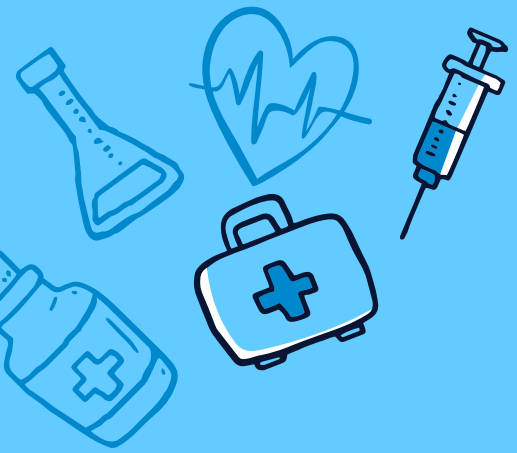
01 Tổng quan vấn đề

02 Lược khảo tài liệu

03 Phương pháp nghiên cứu

04 Thực nghiệm và thảo luận

05 Kết luận và phát triển



01

Tổng quan vấn đề

1.1 Lý do chọn đề tài

- Phát hiện sớm bệnh tim giúp can thiệp kịp thời, giảm nguy cơ tử vong.
- Máy học xử lý dữ liệu lớn, phân tích nhiều yếu tố nguy cơ để tăng độ chính xác chẩn đoán.
- Giảm tải cho bác sĩ, hỗ trợ quản lý sức khỏe cộng đồng.
- Hiệu quả cao của kỹ thuật Ensemble Learning giúp tăng hiệu quả dự đoán.

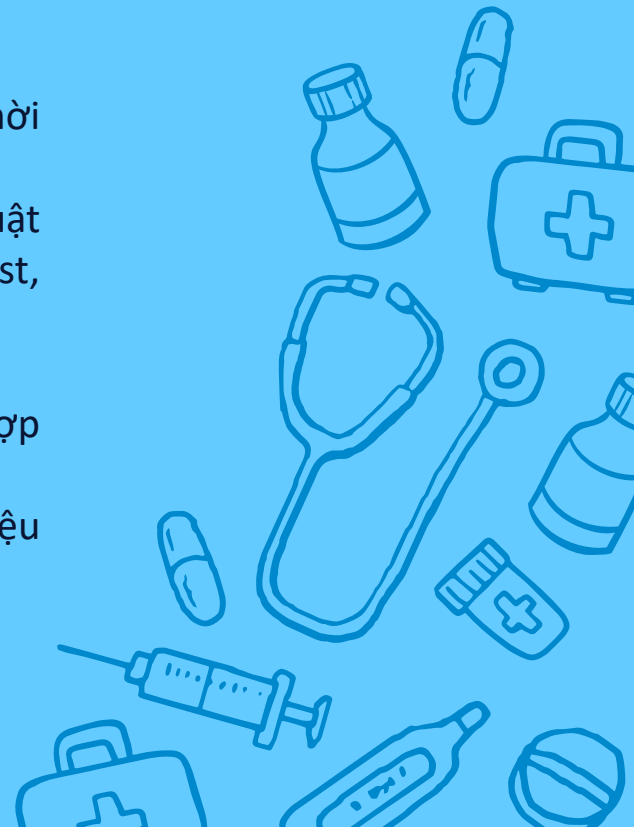
1.2 Vấn đề nghiên cứu & mục tiêu

- Ensemble learning kết hợp nhiều mô hình con (như Decision Tree, Logistic Regression, KNN…)
- **Input**: Các thông tin cá nhân/sức khỏe của người cần dự đoán (tuổi, giới tính, huyết áp, cholesterol, nhịp tim, tiền sử bệnh…).
- **Output**: Nhãn **có nguy cơ mắc bệnh tim hay không** (nhị phân: **1** là có, **0** là không).

1.2 Vấn đề nghiên cứu & mục tiêu

Mục tiêu:

- Đào tạo mô hình máy học với độ chính xác cao và thời gian xử lý nhanh để dự đoán bệnh tim.
- Xây dựng và tối ưu hóa mô hình sử dụng kỹ thuật Ensemble Learning (như Random Forest, XGBoost, Voting...).
- Đánh giá hiệu suất của các mô hình đơn lẻ.
- Lựa chọn các mô hình có độ chính xác cao để kết hợp thành mô hình Ensemble.
- Khai thác ưu điểm của từng mô hình nhằm nâng cao hiệu quả dự đoán tổng thể.



1.3 Câu hỏi & phạm vi nghiên cứu

- Thuật toán máy học nào phù hợp nhất để áp dụng trong việc dự đoán bệnh tim Làm thế nào để lựa chọn và kết hợp các mô hình học máy đơn lẻ nhằm tạo ra một mô hình Ensemble Learning có hiệu suất dự đoán bệnh tim cao??
- Yếu tố dữ liệu nào ảnh hưởng nhiều đến nguy cơ mắc bệnh tim?
- Đánh giá hiệu suất mô hình dựa trên tiêu chí nào là toàn diện nhất?
- Làm sao tối ưu mô hình để cân bằng giữa độ chính xác và tốc độ xử lý?
- **Phạm vi:** Dữ liệu UCI Heart Disease thuộc UCI data Repository, mô hình bao gồm các các thông tin lâm sàng của bệnh nhân

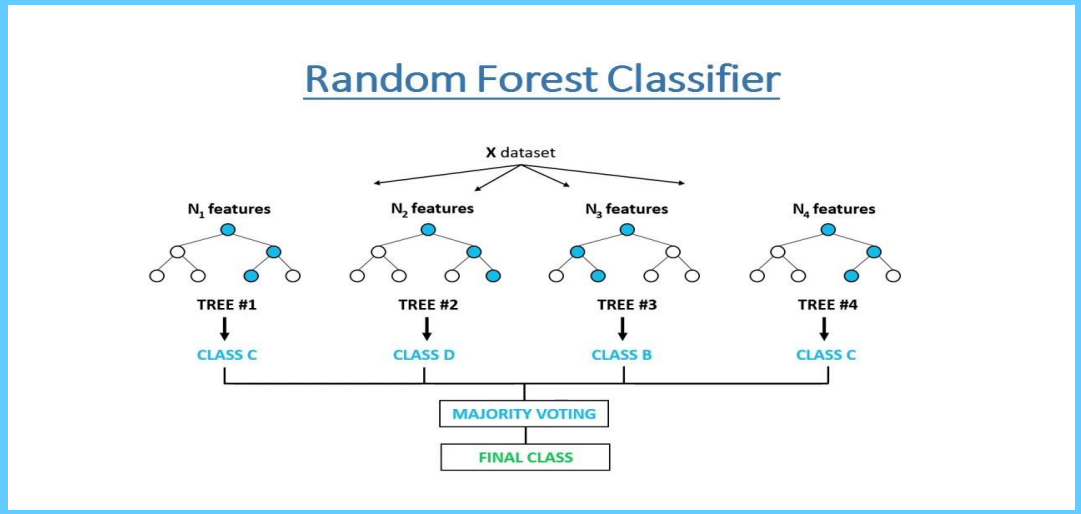
02

Lược khảo tài liệu

2.1 Tổng quan thuật toán

- **Logistic Regression, KNN, SVM, Decision Tree, Random Forest, XGBoost, ANN** đều được áp dụng cho dự đoán bệnh tim.
- Các mô hình **ensemble (Voting, Stacking, Bagging, Boosting)** thường cho kết quả vượt trội so với mô hình đơn lẻ.
- Nghiên cứu liên quan:
  - **An Improved Heart Disease Prediction Using Stacked Ensemble Method:** 9 thuật toán RF, MLP, KNN, XGBoost, SVM, Decision Tree, AdaBoost, Gradient Boosting và MLP, độ chính xác gần 100% (arXiv).

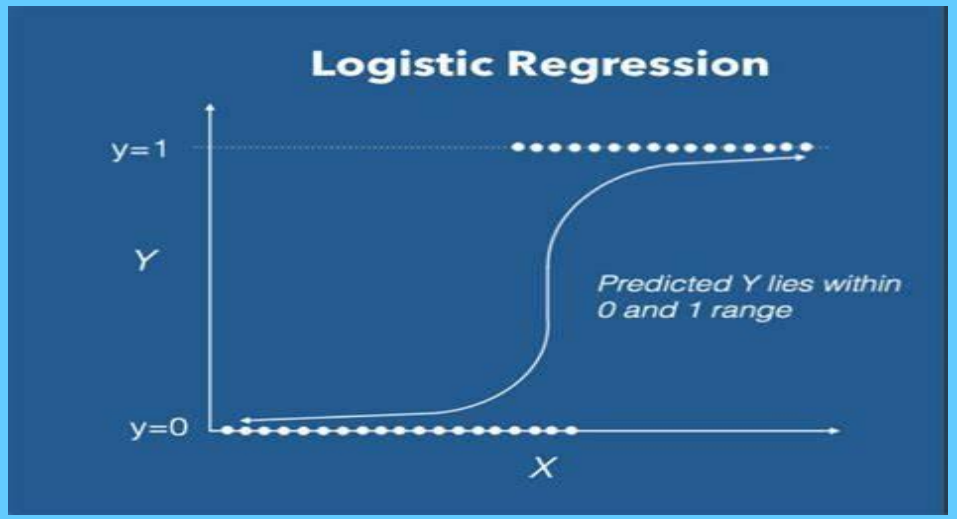
2.1 Tổng quan thuật toán



- Kết hợp nhiều cây quyết định (decision trees) để tăng độ chính xác.
- Mỗi cây học từ tập dữ liệu con + đặc trưng khác nhau.
- Mỗi cây dự đoán một lớp → Bỏ phiếu đa số → ra kết quả cuối cùng.

=> Giảm overfitting, tăng hiệu quả phân loại.

2.1 Tổng quan thuật toán

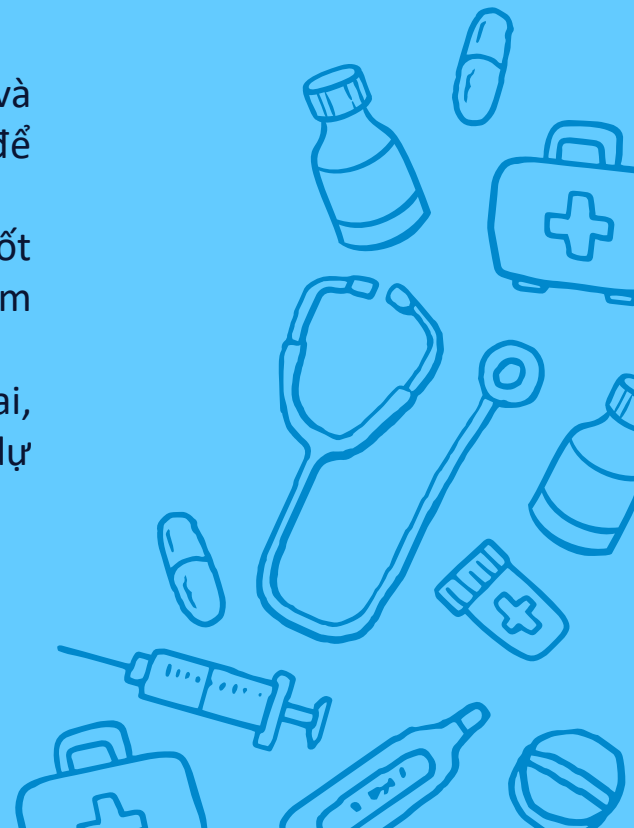


- Dùng để phân loại nhị phân
- Dự đoán đầu ra là xác suất nằm trong khoảng 0 → 1
- Sử dụng hàm sigmoid để chuyển đổi đầu ra tuyến tính thành xác suất.

2.2 Mô hình kết hợp

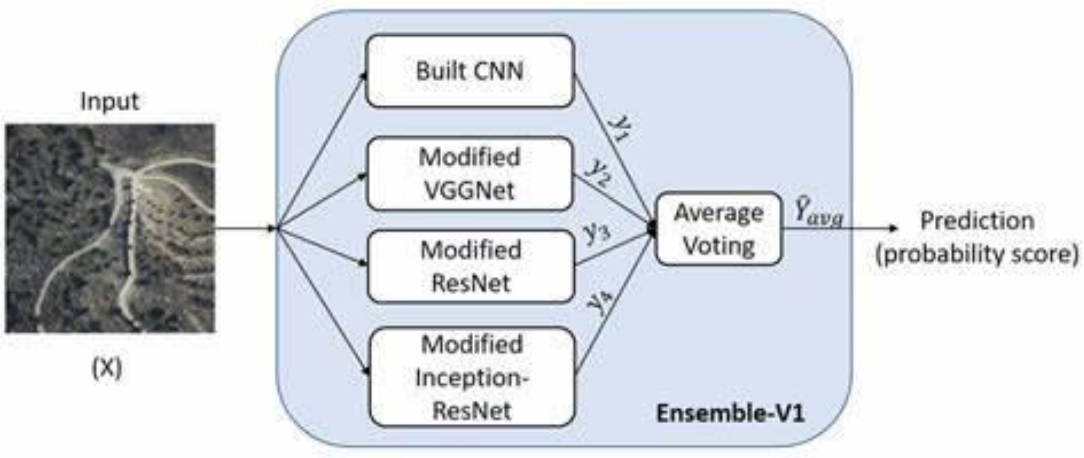
Nghiên cứu này kết hợp hai mô hình Random Forest và Logistic Regression nhằm tận dụng ưu điểm của cả hai để nâng cao hiệu quả dự đoán bệnh tim.

- **Random Forest**: Mô hình ensemble mạnh mẽ, xử lý tốt dữ liệu phức tạp, giảm overfitting, và đánh giá được tầm quan trọng của các đặc trưng.
- **Logistic Regression**: Mô hình đơn giản, dễ triển khai, hiệu quả với dữ liệu tuyến tính và cung cấp xác suất dự đoán.



2.2 Mô hình kết hợp

Phương pháp kết hợp:
Áp dụng **Average Voting**, giúp tăng độ chính xác, giảm overfitting và cải thiện khả năng tổng quát hóa trên dữ liệu mới.



03

Phương pháp nghiên cứu

3.1 Thiết kế nghiên cứu

- Phương pháp **định lượng**: sử dụng dữ liệu số để huấn luyện và đánh giá mô hình.
- Tập trung vào **phân loại** bệnh tim bằng các thuật toán học máy.
- Kết hợp mô hình (SVM, Logistic Regression...) theo **Ensemble Learning** để tăng độ chính xác.

3.2 Đối tượng và mẫu nghiên cứu

- Đối tượng: Bệnh nhân có nguy cơ tim mạch (tuổi, giới tính, huyết áp, cholesterol…).
- Mẫu nghiên cứu: Dữ liệu từ tập Cleveland (UCI).
- Chọn lọc bệnh nhân có yếu tố nguy cơ rõ ràng, đảm bảo tính đại diện và thực tế.

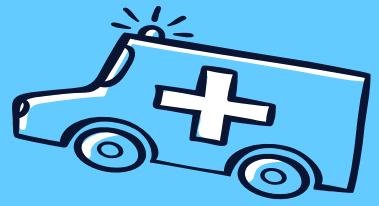
3.3 Thu thập và xử lý dữ liệu

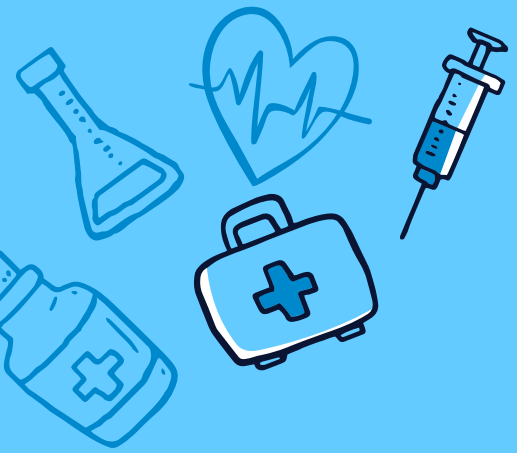
- **Nguồn**: Cleveland Clinic Foundation, công bố năm 1989.
- **Sử dụng 14/76 thuộc tính** có ảnh hưởng lớn đến bệnh tim.
- **Xử lý dữ liệu**:
  - Loại bỏ hoặc điền giá trị thiếu.
  - Chuẩn hóa dữ liệu số (vd: tuổi, huyết áp…) để tăng hiệu suất mô hình.
- **Bảo mật & hợp pháp**: Dữ liệu đã được ẩn danh và công khai cho nghiên cứu.

- Tuổi (age)
- Giới tính (sex)
- Loại đau ngực (cp)
- Huyết áp nghỉ (trestbps)
- Cholesterol huyết thanh (chol)
- Đường huyết lúc đói (fbs)
- Kết quả điện tâm đồ nghỉ (restecg)
- Nhịp tim tối đa đạt được (thalach)
- Đau thắt ngực do gắng sức (exang)
- ST chênh xuống do gắng sức (oldpeak)
- Độ dốc của đoạn ST (slope)
- Số lượng mạch máu chính được nhuộm màu (ca)
- Thalassemia (thal)
- Biến mục tiêu: Sự hiện diện của bệnh tim (num), với giá trị từ 0 (không có bệnh) đến 4 (mức độ nghiêm trọng tăng dần).

3.4 Phân tích dữ liệu

- **Ngôn ngữ & công cụ**: Python, scikit-learn, Jupyter Notebook.
- **Mô hình áp dụng**: SVM, Logistic Regression, kết hợp bằng phương pháp **Voting Ensemble**.
- **Đánh giá mô hình bằng**:
  - Accuracy (độ chính xác)
  - Precision (độ chính xác từng lớp)
  - Recall (khả năng phát hiện đúng bệnh)
  - F1-score (cân bằng giữa precision và recall)





04

Thực nghiệm và thảo luận

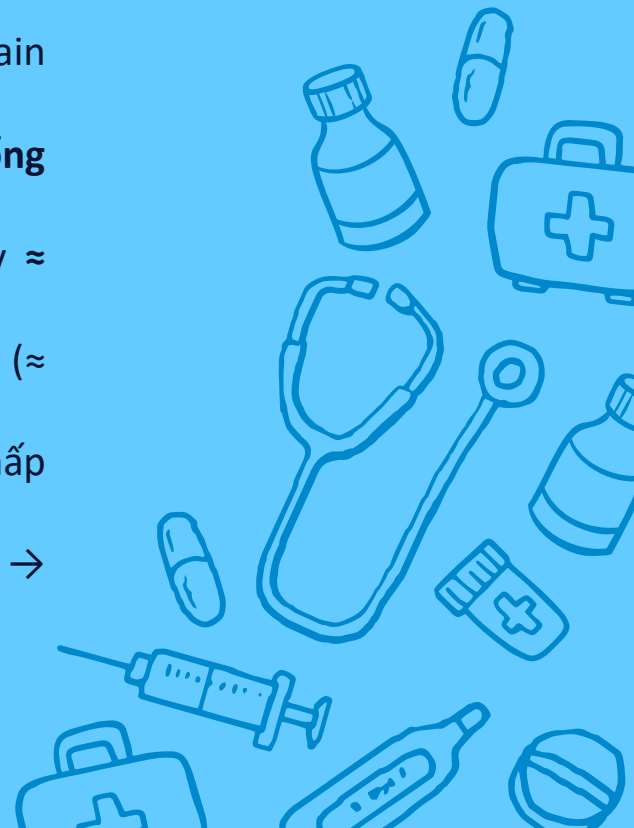
4.1 Kết quả huấn luyện và kiểm thử mô hình

- Đã huấn luyện và đánh giá hiệu suất 6 mô hình:
  KNN, Random Forest, Logistic Regression, Naive Bayes, SVM, Decision Tree
  → Trên tập dữ liệu UCI Heart Disease.

Mô hình	Train Accuracy	Test Accuracy	Precision	Recall
KNN	100.00%	95.08%	94.12%	96.97%
Random Forest	90.87%	91.80%	91.18%	93.94%
Logistic Regression	86.72%	90.16%	90.91%	90.91%
Naive Bayes	83.82%	88.52%	90.63%	87.88%
SVM	94.61%	85.25%	87.50%	84.85%
Decision Tree	93.78%	81.97%	84.38%	81.82%

Đánh giá và phân tích kết quả

- **KNN**: Accuracy kiểm thử **cao nhất (95.08%)**, nhưng train **100%** → **Overfitting**.
- **Random Forest**: Train/Test cân bằng (≈ 91.8%) → **Tổng quát hóa tốt**.
- **Logistic Regression**: Ổn định, dễ giải thích, accuracy ≈ **90.16%**.
- **Naive Bayes**: Hiệu quả bất ngờ với mô hình đơn giản (≈ 88.52%), phù hợp khi dữ liệu tương đối độc lập.
- **SVM**: Huấn luyện tốt (94.61%) nhưng kiểm thử thấp (85.25%) → **Cần tối ưu tham số**.
- **Decision Tree**: Chênh lệch lớn giữa train/test → **Overfitting nặng** nếu không cắt tỉa.

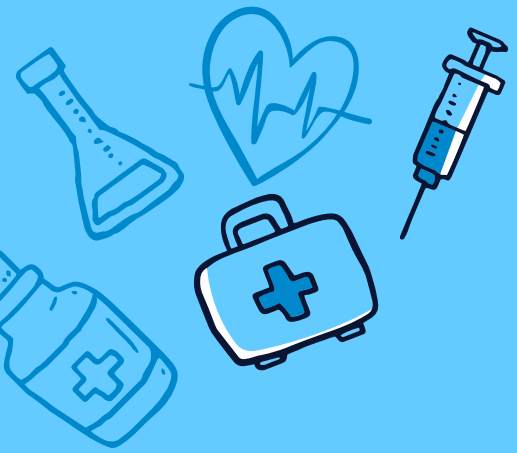


4.2 Đề xuất lựa chọn mô hình

Random Forest và Logistic Regression được chọn vì:

- Độ chính xác cao và ổn định.
- Khả năng tổng quát hóa tốt.

Kết hợp hai mô hình bằng Voting Ensemble để tăng hiệu quả dự đoán bệnh tim.



05

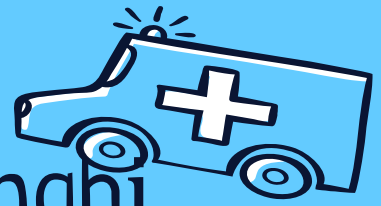
Kết luận và hướng phát triển

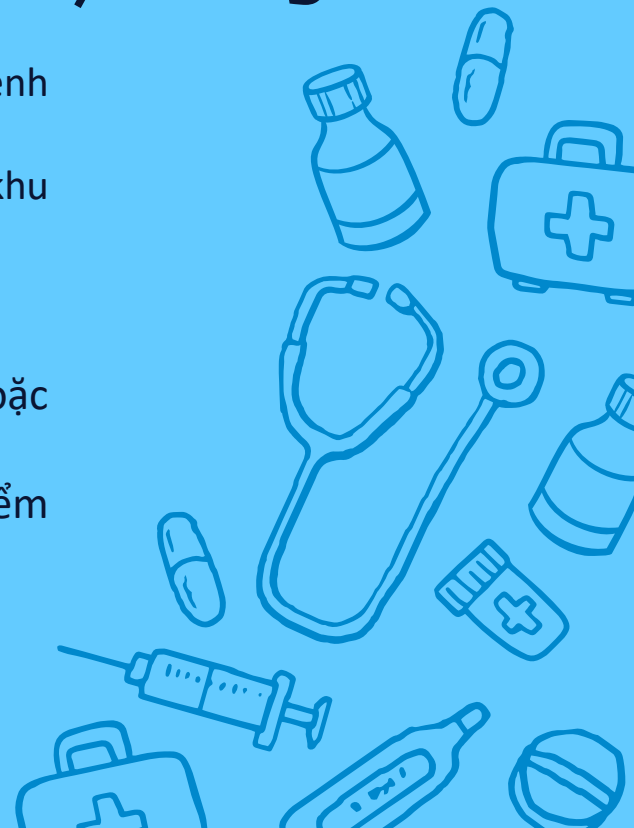
5.1 Kết luận

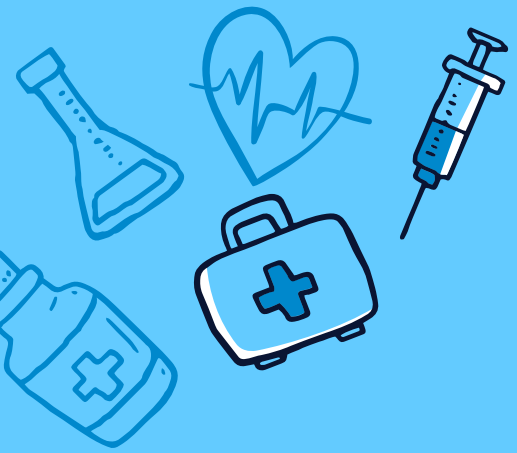
- **Random Forest**: Hiệu suất tốt, **accuracy kiểm thử 91.80%**, tổng quát hóa cao.
- **Logistic Regression**: Ổn định, dễ hiểu, **accuracy 90.16%**, phù hợp ứng dụng thực tế.
- **KNN**: Accuracy kiểm thử **cao nhất (95.08%)**, nhưng **overfitting** (train 100%).
- Kết luận: Mô hình học máy **có thể dự đoán hiệu quả** nguy cơ mắc bệnh tim.
  →Random Forest và Logistic Regression là **hai mô hình đề xuất ứng dụng thực tế**.

5.2 Hướng phát triển & khuyến nghị

- Ứng dụng mô hình vào hệ thống hỗ trợ chẩn đoán bệnh tim trong bệnh viện.
- Tích hợp vào ứng dụng y tế điện tử, đặc biệt tại các khu vực thiếu bác sĩ chuyên khoa.
- Nghiên cứu mở rộng:
  - Tối ưu siêu tham số.
  - Áp dụng mô hình Ensemble nâng cao hoặc Deep Learning.
  - Thử nghiệm trên nhiều bộ dữ liệu hơn để kiểm chứng tính khả dụng.







Thank you