BÁO CÁO ĐỒ ÁN CUỐI KỲ  
DỰ ĐOÁN BỆNH TIM MẠCH SỬ DỤNG MACHINE LEARNING

**Môn học: AI  
Ngày báo cáo: 07/07/2025**

# MỤC LỤC

1. 1. GIỚI THIỆU
2. 2. TỔNG QUAN VỀ DỮ LIỆU
3. 3. PHÂN TÍCH KHÁM PHÁ DỮ LIỆU (EDA)
4. 4. TIỀN XỬ LÝ DỮ LIỆU
5. 5. KỸ THUẬT FEATURE ENGINEERING
6. 6. XÂY DỰNG VÀ ĐÁNH GIÁ MÔ HÌNH
7. 7. TRIỂN KHAI ỨNG DỤNG WEB
8. 8. KẾT LUẬN VÀ KIẾN NGHỊ
9. 9. TÀI LIỆU THAM KHẢO

# 1. GIỚI THIỆU

## 1.1. Bối cảnh nghiên cứu

Bệnh tim mạch là một trong những nguyên nhân hàng đầu gây tử vong trên toàn thế giới.   
 Theo Tổ chức Y tế Thế giới (WHO), mỗi năm có khoảng 17.9 triệu người tử vong do các bệnh   
 tim mạch, chiếm 31% tổng số ca tử vong toàn cầu. Việc phát hiện sớm và dự đoán chính xác   
 nguy cơ mắc bệnh tim mạch đóng vai trò quan trọng trong việc điều trị và phòng ngừa.  
   
 Với sự phát triển của trí tuệ nhân tạo và machine learning, việc ứng dụng các thuật toán   
 học máy để dự đoán bệnh tim mạch đã trở thành một hướng nghiên cứu được quan tâm.   
 Các mô hình machine learning có thể phân tích các yếu tố nguy cơ phức tạp và đưa ra   
 dự đoán chính xác, hỗ trợ các bác sĩ trong quá trình chẩn đoán.

## 1.2. Mục tiêu nghiên cứu

Mục tiêu chính của đồ án này là xây dựng một hệ thống dự đoán bệnh tim mạch sử dụng   
 các thuật toán machine learning, bao gồm:  
   
 - Phân tích và làm sạch dữ liệu về các yếu tố nguy cơ tim mạch  
 - Áp dụng các kỹ thuật feature engineering để cải thiện chất lượng dữ liệu  
 - So sánh hiệu suất của các thuật toán machine learning khác nhau  
 - Xây dựng mô hình ensemble để đạt được độ chính xác cao nhất  
 - Triển khai ứng dụng web với giao diện thân thiện cho người dùng

## 1.3. Phạm vi nghiên cứu

Đồ án sử dụng bộ dữ liệu Heart Failure Prediction Dataset từ Kaggle, bao gồm 918 mẫu dữ liệu   
 với 11 đặc trưng đầu vào và 1 biến mục tiêu. Dự án tập trung vào việc xây dựng mô hình   
 phân loại nhị phân để dự đoán có/không có bệnh tim mạch.

# 2. TỔNG QUAN VỀ DỮ LIỆU

## 2.1. Nguồn dữ liệu

Dữ liệu được thu thập từ Kaggle - Heart Failure Prediction Dataset, được tạo bởi việc   
 kết hợp 5 bộ dữ liệu tim mạch phổ biến khác nhau. Bộ dữ liệu này chứa 918 quan sát với   
 11 đặc trưng được sử dụng phổ biến để dự đoán bệnh tim mạch.

## 2.2. Mô tả các đặc trưng

Bộ dữ liệu bao gồm các đặc trưng sau:  
   
 1. Age: Tuổi của bệnh nhân (năm)  
 2. Sex: Giới tính (Male/Female)  
 3. ChestPainType: Loại đau ngực (TA, ATA, NAP, ASY)  
 4. RestingBP: Huyết áp tâm trương (mmHg)  
 5. Cholesterol: Mức cholesterol trong máu (mg/dl)  
 6. FastingBS: Đường huyết lúc đói > 120 mg/dl (1: Có, 0: Không)  
 7. RestingECG: Kết quả điện tâm đồ lúc nghỉ (Normal, ST, LVH)  
 8. MaxHR: Nhịp tim tối đa đạt được  
 9. ExerciseAngina: Đau thắt ngực khi tập thể dục (Yes/No)  
 10. Oldpeak: ST depression  
 11. ST\_Slope: Độ dốc của đoạn ST (Up, Flat, Down)  
   
 Biến mục tiêu:  
 - HeartDisease: Có bệnh tim mạch (1) hoặc không (0)

# 3. PHÂN TÍCH KHÁM PHÁ DỮ LIỆU (EDA)

## 3.1. Phân tích chất lượng dữ liệu

Quá trình phân tích chất lượng dữ liệu cho thấy:  
   
 - Kích thước dữ liệu: 918 mẫu với 12 cột  
 - Không có giá trị missing trong dataset gốc  
 - Tuy nhiên, phát hiện các giá trị bất thường:  
 + RestingBP có 1 giá trị bằng 0 (không hợp lý về y học)  
 + Cholesterol có 172 giá trị bằng 0 (chiếm 18.7% dữ liệu)  
 + Oldpeak có một số giá trị âm (không phù hợp với định nghĩa y học)  
   
 Phân bố biến mục tiêu:  
 - Có bệnh tim mạch: 508 mẫu (55.3%)  
 - Không có bệnh tim mạch: 410 mẫu (44.7%)  
 - Dữ liệu tương đối cân bằng

## 3.2. Phân tích outliers

Phân tích outliers sử dụng box plot cho thấy:  
   
 - Age: Phân phối khá đều, không có outliers đáng kể  
 - RestingBP: Có nhiều outliers > 170 mmHg và giá trị 0 bất thường  
 - Cholesterol: Có outliers > 300 mg/dl và nhiều giá trị 0 không hợp lý  
 - MaxHR: Có một số outliers nhưng trong phạm vi y học chấp nhận được  
 - Oldpeak: Có giá trị âm không phù hợp với định nghĩa y học

## 3.3. Phân tích tương quan

Ma trận tương quan giữa các biến số cho thấy:  
   
 - Oldpeak có tương quan mạnh với HeartDisease (r = 0.43)  
 - MaxHR có tương quan nghịch với HeartDisease (r = -0.38)  
 - Age có tương quan dương với HeartDisease (r = 0.28)  
 - Cholesterol có tương quan yếu với HeartDisease  
 - Các biến phân loại như ChestPainType, ST\_Slope cũng có ảnh hưởng đáng kể

# 4. TIỀN XỬ LÝ DỮ LIỆU

## 4.1. Xử lý dữ liệu bất thường

Quá trình tiền xử lý dữ liệu bao gồm:  
   
 1. Loại bỏ các mẫu có RestingBP = 0 (1 mẫu bị loại)  
 2. Xử lý giá trị Oldpeak âm bằng cách gán về 0  
 3. Sử dụng model-based imputation cho Cholesterol:  
 - Áp dụng XGBoost Regressor để dự đoán giá trị Cholesterol bị thiếu  
 - Sử dụng log transformation để xử lý skewness  
 - Tạo missing flag để đánh dấu các giá trị được impute  
 4. Winsorization: Giới hạn các giá trị cực trị  
 - RestingBP: 80-180 mmHg  
 - Cholesterol: 100-500 mg/dl  
 - MaxHR: 60-200 bpm  
 - Oldpeak: 0-5

## 4.2. Chuẩn hóa dữ liệu phân loại

Xử lý dữ liệu phân loại:  
   
 1. Chuẩn hóa format: Chuyển về uppercase, loại bỏ khoảng trắng  
 2. Mapping giá trị:  
 - Sex: M -> Male, F -> Female  
 - ExerciseAngina: Y -> Yes, N -> No  
 3. Chuyển đổi kiểu dữ liệu thành category để tối ưu bộ nhớ

# 5. KỸ THUẬT FEATURE ENGINEERING

## 5.1. Tạo đặc trưng mới

Các đặc trưng mới được tạo:  
   
 1. Age\_MaxHR\_Ratio: Tỷ lệ tuổi/nhịp tim tối đa  
 - Phản ánh khả năng tim mạch theo tuổi  
   
 2. BP\_Stress\_Diff: Chênh lệch huyết áp và stress  
 - Công thức: RestingBP - (Oldpeak \* 10)  
   
 3. AgeGroup: Nhóm tuổi phân loại  
 - <40, 40-55, 55-65, 65+  
   
 4. Chol\_Level: Mức cholesterol phân loại  
 - Normal (<200), Borderline (200-240), High (>240)  
   
 5. Cardiac\_Risk\_Score: Điểm nguy cơ tim mạch tổng hợp  
 - Kết hợp Age, Cholesterol, RestingBP, MaxHR với trọng số  
   
 6. Polynomial features: Age\_Squared, Cholesterol\_Squared  
 - Nắm bắt mối quan hệ phi tuyến

## 5.2. Chuẩn hóa dữ liệu

Áp dụng StandardScaler cho các đặc trưng số:  
   
 - Chuẩn hóa về mean = 0, std = 1  
 - Chỉ áp dụng cho các biến số, không áp dụng cho biến phân loại  
 - Đảm bảo tất cả các thuật toán ML có thể hoạt động hiệu quả

# 6. XÂY DỰNG VÀ ĐÁNH GIÁ MÔ HÌNH

## 6.1. Lựa chọn thuật toán

Dự án so sánh 3 thuật toán chính:  
   
 1. Logistic Regression:  
 - Thuật toán tuyến tính, dễ hiểu và giải thích  
 - Thích hợp cho dữ liệu tuyến tính  
   
 2. Random Forest:  
 - Ensemble method, xử lý tốt overfitting  
 - Có thể nắm bắt mối quan hệ phi tuyến  
   
 3. XGBoost:  
 - Gradient boosting, hiệu suất cao  
 - Tối ưu cho dữ liệu có cấu trúc

## 6.2. Tối ưu siêu tham số

Sử dụng RandomizedSearchCV với 5-fold cross-validation:  
   
 XGBoost parameters:  
 - n\_estimators: 100-1000  
 - max\_depth: 3-10  
 - learning\_rate: 0.01-0.3  
 - subsample: 0.6-1.0  
 - colsample\_bytree: 0.6-1.0  
 - reg\_alpha, reg\_lambda: 0-1  
   
 Random Forest parameters:  
 - n\_estimators: 100-1000  
 - max\_depth: 5-30  
 - min\_samples\_split: 2-10  
 - min\_samples\_leaf: 1-10  
 - max\_features: sqrt, log2, None  
   
 Logistic Regression parameters:  
 - C: 0.001-10  
 - penalty: l1, l2, elasticnet  
 - solver: saga

## 6.3. Kết quả đánh giá mô hình

Kết quả ROC AUC trên tập validation:  
   
 - XGBoost: 0.9156  
 - Random Forest: 0.9104  
 - Logistic Regression: 0.8987  
   
 Mô hình Ensemble (Voting Classifier):  
 - Kết hợp 3 mô hình trên với voting='soft'  
 - ROC AUC: 0.9201  
 - Precision: 0.89  
 - Recall: 0.91  
 - F1-Score: 0.90  
   
 Tối ưu threshold: 0.4873 (để cân bằng precision và recall)

## 6.4. Phân tích Feature Importance

Top 10 đặc trưng quan trọng nhất:  
   
 1. ST\_Slope: Quan trọng nhất trong dự đoán  
 2. ChestPainType: Loại đau ngực  
 3. Oldpeak: ST depression  
 4. MaxHR: Nhịp tim tối đa  
 5. ExerciseAngina: Đau thắt ngực khi tập thể dục  
 6. Age\_MaxHR\_Ratio: Đặc trưng kỹ thuật mới  
 7. BP\_Stress\_Diff: Đặc trưng kỹ thuật mới  
 8. Cholesterol: Mức cholesterol  
 9. Age: Tuổi  
 10. RestingBP: Huyết áp tâm trương

# 7. TRIỂN KHAI ỨNG DỤNG WEB

## 7.1. Kiến trúc hệ thống

Ứng dụng web được xây dựng với FastAPI framework:  
   
 Backend:  
 - FastAPI: Web framework hiện đại, hiệu suất cao  
 - Joblib: Lưu trữ và load mô hình machine learning  
 - Pandas: Xử lý dữ liệu đầu vào  
 - Pydantic: Validation dữ liệu  
   
 Frontend:  
 - HTML5/CSS3: Giao diện người dùng  
 - Jinja2: Template engine  
 - Responsive design: Tương thích mobile  
   
 API Endpoints:  
 - GET /: Trang chủ với form nhập liệu  
 - POST /predict: API dự đoán JSON  
 - POST /chat: Form-based prediction  
 - GET /docs: Tài liệu API tự động

## 7.2. Tính năng ứng dụng

Các tính năng chính:  
   
 1. Giao diện thân thiện:  
 - Form nhập liệu trực quan  
 - Validation dữ liệu đầu vào  
 - Responsive design  
   
 2. Dự đoán thông minh:  
 - Sử dụng mô hình ensemble tốt nhất  
 - Áp dụng feature engineering tự động  
 - Phân loại mức độ nguy cơ (Thấp/Trung bình/Cao)  
   
 3. Kết quả chi tiết:  
 - Xác suất dự đoán chính xác  
 - Phân loại mức độ nguy cơ  
 - Khuyến nghị cho người dùng  
   
 4. API Documentation:  
 - Tài liệu API tự động với FastAPI  
 - Swagger UI tương tác  
 - Dễ dàng tích hợp với hệ thống khác

## 7.3. Xử lý dữ liệu đầu vào

Quy trình xử lý dữ liệu từ người dùng:  
   
 1. Validation: Kiểm tra tính hợp lệ của dữ liệu đầu vào  
 2. Feature Engineering: Tạo các đặc trưng mới tự động  
 3. Scaling: Chuẩn hóa dữ liệu theo cùng phương pháp training  
 4. Prediction: Sử dụng mô hình ensemble để dự đoán  
 5. Post-processing: Chuyển đổi kết quả thành mức độ nguy cơ

# 8. KẾT LUẬN VÀ KIẾN NGHỊ

## 8.1. Kết quả đạt được

Đồ án đã đạt được những kết quả sau:  
   
 1. Xây dựng thành công pipeline xử lý dữ liệu hoàn chỉnh:  
 - Xử lý missing values bằng model-based imputation  
 - Áp dụng các kỹ thuật feature engineering hiệu quả  
 - Chuẩn hóa và làm sạch dữ liệu chuyên nghiệp  
   
 2. Phát triển mô hình machine learning chính xác cao:  
 - ROC AUC đạt 0.9201 trên tập test  
 - Sử dụng ensemble method để tối ưu hiệu suất  
 - Tối ưu threshold để cân bằng precision-recall  
   
 3. Triển khai ứng dụng web hoàn chỉnh:  
 - Giao diện người dùng thân thiện và chuyên nghiệp  
 - API RESTful cho việc tích hợp  
 - Xử lý dữ liệu real-time  
   
 4. Đảm bảo tính thực tiễn:  
 - Validation dữ liệu đầu vào nghiêm ngặt  
 - Khuyến nghị phù hợp với từng mức độ nguy cơ  
 - Thiết kế responsive cho nhiều thiết bị

## 8.2. Hạn chế và kiến nghị

Hạn chế của đồ án:  
   
 1. Kích thước dataset tương đối nhỏ (918 mẫu)  
 2. Thiếu một số đặc trưng y học quan trọng khác  
 3. Chưa có validation từ chuyên gia y tế  
 4. Chưa test trên dữ liệu real-world  
   
 Kiến nghị phát triển:  
   
 1. Mở rộng dataset với nhiều nguồn dữ liệu khác  
 2. Tích hợp thêm các đặc trưng y học khác  
 3. Collaboration với các chuyên gia tim mạch  
 4. Áp dụng deep learning cho dữ liệu phức tạp hơn  
 5. Triển khai trên cloud để mở rộng quy mô  
 6. Thêm tính năng theo dõi lịch sử bệnh nhân

## 8.3. Ý nghĩa thực tiễn

Đồ án có ý nghĩa thực tiễn như sau:  
   
 1. Hỗ trợ sàng lọc sơ bộ: Giúp phát hiện sớm nguy cơ tim mạch  
 2. Công cụ hỗ trợ quyết định: Cung cấp thông tin cho bác sĩ  
 3. Giáo dục sức khỏe: Nâng cao nhận thức về yếu tố nguy cơ  
 4. Tiết kiệm chi phí: Giảm số lượng xét nghiệm không cần thiết  
 5. Accessibility: Có thể truy cập từ bất kỳ đâu có internet

# 9. TÀI LIỆU THAM KHẢO

1. Heart Failure Prediction Dataset, Kaggle  
 URL: https://www.kaggle.com/datasets/fedesoriano/heart-failure-prediction  
   
 2. Scikit-learn Documentation  
 URL: https://scikit-learn.org/stable/  
   
 3. XGBoost Documentation  
 URL: https://xgboost.readthedocs.io/  
   
 4. FastAPI Documentation  
 URL: https://fastapi.tiangolo.com/  
   
 5. Github bài làm  
 URL: [MHP0920/Heart-Disease-Failure-Kaggle](https://github.com/MHP0920/Heart-Disease-Failure-Kaggle)  
   
 6. Python Data Science Handbook by Jake VanderPlas  
   
 7. Hands-On Machine Learning by Aurélien Géron  
   
 8. Pattern Recognition and Machine Learning by Christopher Bishop