BỘ NÔNG NGHIỆP VÀ MÔI TRƯỜNG

PHÂN HIỆU TRƯỜNG ĐẠI HỌC THỦY LỢI

BỘ MÔN CÔNG NGHỆ THÔNG TIN

BÁO CÁO CUỐI KỲ

MÔN HỌC: Học Máy

ĐỀ TÀI: Ứng dụng Machine Learning trong dự đoán bệnh tim

Giảng viên:	Ths. Vũ Thị Hạnh
Sinh viên thực hiện:	Nguyễn Lê Minh Hậu – 2351267261 Nguyễn Đức Huy - 2351267265
Lớp:	S26-65TTNT

NHẬN XÉT CỦA GIẢNG VIÊN HƯỚNG DẪN

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

TP. Hồ Chí Minh, ngày … tháng … năm 2026

Chữ ký của giảng viên

LỜI CẢM ƠN

Để hoàn thành báo cáo kết thúc môn học này, bên cạnh sự nỗ lực của từng các nhân thì nhóm em đã nhận được sự hướng dẫn tận tình và đầy tâm huyết từ giảng viên hướng dẫn.

Lời đầu tiên, nhóm em xin được bày tỏ lòng biết ơn sâu sắc và chân thành nhất đến cô Vũ Thị Hạnh, người đã trực tiếp giảng dạy, định hướng và đồng hành cùng nhóm em trong suốt quá trình học tập và thực hiện đồ án môn Học Máy này.

Môn Học Máy này không chỉ giúp chúng em tiếp cận những kiến thức nền tảng quan trọng về các thuật toán học máy, mô hình dự đoán và phân tích dữ liệu, mà còn rèn luyện cho chúng em tư duy logic, khả năng lập trình cũng như phương pháp tiếp cận và giải quyết các bài toán thực tế dựa trên dữ liệu. Chúng em đặc biệt trân trọng những bài giảng rõ ràng, khoa học của cô, sự tận tâm trong việc giải đáp thắc mắc cũng như những định hướng chuyên môn quý báu, giúp chúng em từng bước hoàn thiện mô hình và hiểu sâu hơn bản chất của các thuật toán đã học.

Sự nghiêm túc trong học thuật, cùng với sự động viên và khích lệ đúng lúc của cô, chính là động lực lớn để chúng em không ngừng cố gắng, hoàn thiện bài báo cáo một cách chỉn chu và nghiêm túc nhất.

Bên cạnh đó, chúng em cũng xin gửi lời cảm ơn đến các nguồn tài liệu tham khảo và cộng đồng học thuật, đã cung cấp những kiến thức, dữ liệu và kinh nghiệm thực tiễn, giúp chúng em có cơ sở để vận dụng hiệu quả những nội dung đã được học vào bài toán cụ thể.

Do thời gian thực hiện và kiến thức còn hạn chế, bài báo cáo của nhóm em khó tránh khỏi những thiếu sót. Nhóm em rất mong nhận được những nhận xét và góp ý chân thành từ cô để chúng em có thể rút kinh nghiệm và hoàn thiện bản thân hơn trong quá trình học tập và nghiên cứu sau này.

Chúng em xin chân thành cảm ơn cô!

MỤC LỤC

I. GIỚI THIỆU

1.1 Bối cảnh và lý do chọn đề tài

Trong những năm gần đây, cùng với sự phát triển mạnh mẽ của công nghệ thông tin và trí tuệ nhân tạo, việc khai thác và phân tích dữ liệu ngày càng đóng vai trò quan trọng trong nhiều lĩnh vực khác nhau. Đặc biệt, sự bùng nổ của dữ liệu số đã tạo điều kiện thuận lợi cho sự phát triển của Machine Learning (Máy học) – một nhánh quan trọng của trí tuệ nhân tạo, cho phép máy tính học hỏi từ dữ liệu và đưa ra dự đoán hoặc quyết định mà không cần lập trình tường minh.

Trong lĩnh vực y tế, nhu cầu ứng dụng Machine Learning ngày càng gia tăng nhằm hỗ trợ chẩn đoán, dự báo và phòng ngừa bệnh tật. Trong đó, bệnh tim mạch là một trong những nguyên nhân hàng đầu gây tử vong trên thế giới cũng như tại Việt Nam. Việc phát hiện sớm nguy cơ mắc bệnh tim có ý nghĩa vô cùng quan trọng trong việc giảm thiểu biến chứng, nâng cao hiệu quả điều trị và cải thiện chất lượng cuộc sống cho người bệnh. Tuy nhiên, quá trình đánh giá nguy cơ bệnh tim thường dựa trên nhiều yếu tố lâm sàng khác nhau, gây khó khăn trong việc phân tích và đưa ra quyết định chính xác.

Xuất phát từ thực tiễn đó, việc áp dụng Machine Learning để phân tích dữ liệu y khoa và xây dựng các mô hình dự đoán bệnh tim được xem là một hướng tiếp cận hiệu quả và có tính ứng dụng cao. Chính vì vậy, đề tài "Ứng dụng Machine Learning trong dự đoán bệnh tim" đã được lựa chọn để nghiên cứu và thực hiện trong khuôn khổ môn học.

1.2 Vai trò của Machine Learning trong y tế

Machine Learning đã và đang được ứng dụng rộng rãi trong lĩnh vực y tế nhằm hỗ trợ bác sĩ và chuyên gia trong việc phân tích dữ liệu, chẩn đoán bệnh và đưa ra quyết định điều trị. Thông qua việc học từ dữ liệu lịch sử, các mô hình Machine Learning có khả năng phát hiện những mối quan hệ tiềm ẩn giữa các đặc trưng lâm sàng mà các phương pháp phân tích truyền thống khó nhận ra.

Trong bài toán dự đoán bệnh tim, Machine Learning cho phép xử lý đồng thời nhiều yếu tố nguy cơ như tuổi tác, giới tính, huyết áp, cholesterol, nhịp tim và các chỉ số sinh học khác. Nhờ đó, mô hình có thể đưa ra dự đoán về khả năng mắc bệnh tim của bệnh nhân với độ chính xác cao, góp phần hỗ trợ bác sĩ trong việc sàng lọc, đánh giá nguy cơ và theo dõi sức khỏe tim mạch.

Việc ứng dụng Machine Learning không nhằm thay thế hoàn toàn vai trò của con người trong y tế, mà đóng vai trò như một công cụ hỗ trợ quyết định, giúp nâng cao hiệu quả và độ tin cậy trong quá trình chẩn đoán và điều trị.

1.3 Tổng quan các nghiên cứu liên quan

Dự đoán bệnh tim bằng Machine Learning là một bài toán đã được nhiều nhà nghiên cứu quan tâm trong khoảng hai thập kỷ trở lại đây. Với sự phát triển của các thuật toán học máy và sự sẵn có của các bộ dữ liệu y khoa công khai, nhiều phương pháp khác nhau đã được đề xuất nhằm nâng cao độ chính xác và khả năng phát hiện sớm nguy cơ mắc bệnh tim. Các nghiên cứu tập trung vào việc khai thác dữ liệu lâm sàng để xây dựng các mô hình phân loại, hỗ trợ chẩn đoán và ra quyết định trong y tế.

Bảng dưới đây tổng hợp một số nghiên cứu tiêu biểu liên quan đến bài toán dự đoán bệnh tim bằng Machine Learning:

Tác giả (Năm)	Phương pháp	Dataset	Kết quả tốt nhất	Hạn chế
Mohan et al. (2019)	Hybrid RF–LSSVM	Cleveland (303)	Accuracy = 88.4%	Chưa tối ưu Recall, không xử lý mất cân bằng dữ liệu
Ali et al. (2019)	Ensemble (RF, AdaBoost, XGBoost)	Cleveland (920)	Recall = 93.1%, Accuracy = 92.5%	Chưa phân tích sâu các mẫu ẩn trong dữ liệu
Shah et al. (2020)	Deep Learning (CNN)	UCI (1190)	Accuracy = 90.1%	Cần nhiều dữ liệu, khó giải thích mô hình
Tama et al. (2020)	XGBoost + SMOTE	Cleveland (303)	F1-score = 94.2%	SMOTE tạo dữ liệu tổng hợp, có thể làm giảm tính thực tế
Nghiên cứu này (2026)	RF + XGB + NN + Hidden Patterns	UCI / Kaggle (918)	Recall = 96.08%, AUC = 0.9449	Dataset còn hạn chế, cần kiểm chứng trên dữ liệu ngoài

Từ bảng tổng hợp có thể nhận thấy rằng các thuật toán Machine Learning như Random Forest, XGBoost và các mô hình Ensemble thường mang lại hiệu năng tốt trên dữ liệu y khoa dạng bảng. Trong khi đó, các phương pháp Deep Learning có khả năng khai thác mối quan hệ phi tuyến phức tạp, nhưng lại đòi hỏi tập dữ liệu lớn và

gặp khó khăn trong việc giải thích kết quả, điều này là một hạn chế đáng kể trong các ứng dụng y tế.

Một điểm chung của nhiều nghiên cứu trước đây là tập trung chủ yếu vào việc tối ưu độ chính xác (Accuracy), trong khi các chỉ số mang ý nghĩa lâm sàng như Recall (Sensitivity) chưa được chú trọng đầy đủ. Trong bài toán dự đoán bệnh tim, Recall đóng vai trò đặc biệt quan trọng vì việc bỏ sót các ca bệnh có thể gây ra hậu quả nghiêm trọng cho người bệnh. Ngoài ra, vấn đề mất cân bằng dữ liệu giữa nhóm bệnh và nhóm không bệnh vẫn là một thách thức, và một số nghiên cứu sử dụng các kỹ thuật tạo dữ liệu tổng hợp như SMOTE, tiềm ẩn nguy cơ làm sai lệch phân bố dữ liệu thực tế.

So với các công trình liên quan, nghiên cứu này kế thừa các phương pháp Machine Learning phổ biến và hiệu quả, đồng thời có những cải tiến nhằm phù hợp hơn với yêu cầu của bài toán y tế. Cụ thể, nghiên cứu tập trung tối ưu Recall thông qua Threshold Tuning, giúp nâng cao khả năng phát hiện các ca bệnh tim. Bên cạnh đó, việc bổ sung bước phân tích Hidden Patterns cho phép khám phá sâu hơn các mối quan hệ tiềm ẩn giữa các đặc trưng lâm sàng, góp phần nâng cao giá trị phân tích của mô hình.

Kết quả đạt được cho thấy mô hình trong nghiên cứu này có Recall = 96.08% và AUC = 0.9449, cao hơn so với phần lớn các nghiên cứu được tổng hợp. Tuy nhiên, nghiên cứu vẫn còn những hạn chế nhất định như quy mô dữ liệu chưa lớn và chưa được kiểm chứng trên tập dữ liệu bên ngoài. Đây cũng là những hướng phát triển tiềm năng cho các nghiên cứu tiếp theo nhằm nâng cao tính tin cậy và khả năng ứng dụng thực tế của mô hình.

1.4 Cấu trúc dự án

Cấu trúc tổng thể của đề tài Dự đoán bệnh tim bằng Machine Learning được tổ chức theo hướng mô-đun hóa, đảm bảo tính khoa học, rõ ràng và thuận tiện cho việc phát triển, đánh giá cũng như mở rộng hệ thống. Sơ đồ cấu trúc của đề tài được mô tả như sau

```
S26-65TTNT_Nhom7_DuDoanBenhTim/
│
├── data/
│   └── heart.csv                                  # Dataset gốc
│
├── src/                                           # Mã nguồn (Source Code)
│   ├── preprocessing.py                           # Tiền xử lý & Pipeline
│   ├── feature_engineering.py                     # Tạo đặc trưng y khoa
│   ├── eda.py                                     # Phân tích khám phá dữ liệu
│   ├── RandomForest_NguyenLeMinhHau.py            # Mô hình Random Forest
│   ├── XGBoost_NguyenLeMinhHau.py                 # Mô hình XGBoost
│   ├── NeuralNetwork_NguyenDucHuy.py              # Mô hình Neural Network
│   ├── evaluation.py                              # Đánh giá & So sánh
│   └── app.py                                     # Ứng dụng Web Demo
│
├── saved_models/                                  # Lưu trữ Model & Artifacts
│   ├── random_forest.pkl
│   ├── xgboost.pkl
│   ├── neural_network.pkl
│   ├── scaler.pkl, imputer.pkl                    # Scaler & Imputer
│   └── *.metadata.pkl
│
├── outputs/                                       # Kết quả đầu ra
│   ├── figures/                                   # Biểu đồ (ROC, CM...)
│   ├── *_log.txt                                  # Log huấn luyện
│   └── Evaluation_Report.txt                      # Báo cáo tổng hợp
│
├── reports/
│   └── N7_report.pdf                              # File báo cáo PDF
│
├── requirements.txt                               # Thư viện cần thiết
└── README.md                                      # Hướng dẫn
```

Thông qua cách tổ chức này, đề tài đảm bảo tính hệ thống, dễ mở rộng và thuận tiện cho việc kiểm tra, đánh giá cũng như phát triển trong các nghiên cứu tiếp theo.

II. MỤC TIÊU VÀ BÀI TOÁN ĐẶT RA

2.1 Mục tiêu chính

Mục tiêu chính của đề tài là xây dựng một hệ thống Machine Learning có khả năng hỗ trợ dự đoán nguy cơ mắc bệnh tim dựa trên dữ liệu lâm sàng, hướng đến tính chính xác, độ tin cậy và khả năng ứng dụng trong thực tế. Cụ thể, hệ thống được xây dựng nhằm đáp ứng các mục tiêu sau:

1. Xây dựng mô hình dự đoán nguy cơ mắc bệnh tim với độ chính xác cao (trên 85%), đảm bảo hiệu quả phân loại trên dữ liệu y khoa.
2. Ưu tiên tối ưu chỉ số Recall (độ nhạy) nhằm giảm thiểu số lượng ca bệnh bị bỏ sót (False Negative), phù hợp với yêu cầu của các bài toán trong lĩnh vực y tế.
3. Phân tích và phát hiện các tổ hợp triệu chứng và đặc trưng lâm sàng (mẫu ẩn – Hidden Patterns) có mối liên hệ chặt chẽ với bệnh tim.
4. Cung cấp giải thích có cơ sở khoa học cho các kết quả dự đoán của mô hình, góp phần nâng cao tính minh bạch và độ tin cậy khi áp dụng trong thực tế.
5. So sánh và đánh giá hiệu suất của ba mô hình khác nhau bao gồm Random Forest, XGBoost và Neural Network, nhằm lựa chọn mô hình phù hợp nhất cho bài toán.
6. Xây dựng giao diện người dùng thân thiện, hỗ trợ bác sĩ và người sử dụng dễ dàng nhập dữ liệu và tiếp cận kết quả dự đoán.

2.2 Bài toán đặt ra

Bài toán đặt ra trong đề tài là bài toán phân loại nhị phân (Binary Classification) trong lĩnh vực Machine Learning, với mục tiêu dự đoán nguy cơ mắc bệnh tim của bệnh nhân dựa trên các chỉ số lâm sàng.

Cụ thể, dữ liệu đầu vào (Input) của bài toán bao gồm 12 chỉ số lâm sàng phản ánh tình trạng sức khỏe tim mạch của bệnh nhân, chẳng hạn như tuổi, giới tính, huyết áp, cholesterol, nhịp tim và một số yếu tố nguy cơ liên quan khác. Các đặc trưng này được sử dụng để mô tả tình trạng sức khỏe của từng bệnh nhân và làm cơ sở cho mô hình học máy đưa ra dự đoán.

Đầu ra (Output) của bài toán là một biến nhị phân, trong đó:

- 0 (Negative): Bệnh nhân không có nguy cơ mắc bệnh tim.
- 1 (Positive): Bệnh nhân có nguy cơ mắc bệnh tim.

Do đặc thù của bài toán y tế, việc đánh giá mô hình không chỉ dựa trên độ chính xác tổng thể mà cần ưu tiên các chỉ số phản ánh khả năng phát hiện đúng ca bệnh. Vì vậy, Recall (độ nhạy) được lựa chọn là tiêu chí đánh giá quan trọng nhất nhằm hạn chế tối đa số lượng trường hợp mắc bệnh tim nhưng bị mô hình dự đoán sai là không mắc

bệnh (False Negative). Bên cạnh đó, Precision và Accuracy cũng được sử dụng như các chỉ số bổ trợ để đánh giá toàn diện hiệu năng của mô hình.

Cách tiếp cận này giúp đảm bảo rằng hệ thống dự đoán bệnh tim được xây dựng không chỉ đạt hiệu suất cao về mặt kỹ thuật mà còn phù hợp với yêu cầu thực tiễn trong lĩnh vực y tế, nơi việc bỏ sót bệnh nhân có thể dẫn đến những hậu quả nghiêm trọng.

2.3 Các mô hình được lựa chọn

Trong đề tài này, ba mô hình Machine Learning và Deep Learning tiêu biểu được lựa chọn nhằm xây dựng và so sánh hệ thống dự đoán bệnh tim, bao gồm Random Forest, XGBoost và Neural Network. Việc lựa chọn các mô hình này dựa trên tính hiệu quả, khả năng ứng dụng trong dữ liệu y khoa và sự phù hợp với bài toán phân loại nhị phân.

Ưu điểm của Random Forest trong đề tài bao gồm:

- Không nhạy cảm với dữ liệu chưa được chuẩn hóa, phù hợp với dữ liệu y khoa dạng bảng.
- Có khả năng giải thích cao thông qua việc đánh giá tầm quan trọng của đặc trưng (Feature Importance).
- Chống overfitting hiệu quả nhờ cơ chế trung bình hóa kết quả của nhiều cây quyết định.
- Hoạt động ổn định trên các tập dữ liệu có quy mô vừa và nhỏ.

2.3.1 Random Forest

Random Forest là một thuật toán học máy thuộc nhóm Ensemble Learning, được xây dựng bằng cách kết hợp nhiều cây quyết định độc lập và đưa ra kết quả dự đoán dựa trên nguyên tắc bỏ phiếu. Nhờ cơ chế tổng hợp này, Random Forest có khả năng giảm phương sai và hạn chế hiện tượng quá khớp.

2.3.2 XGBoost

XGBoost là một thuật toán Gradient Boosting cải tiến, nổi bật với khả năng huấn luyện hiệu quả và hiệu năng cao trên các bài toán phân loại và hồi quy. Thuật toán này xây dựng mô hình bằng cách liên tục cải thiện lỗi dự đoán thông qua việc huấn luyện các cây quyết định kế tiếp.

Những lý do lựa chọn XGBoost trong đề tài bao gồm:

- Hiệu suất cao và được sử dụng rộng rãi trong các cuộc thi Machine Learning.
- Có khả năng xử lý tốt dữ liệu mất cân bằng thông qua các tham số điều chỉnh trọng số lớp.

- Hỗ trợ tối ưu hóa ngưỡng phân loại (Threshold Tuning), phù hợp với yêu cầu ưu tiên Recall trong bài toán y tế.
- Khả năng kiểm soát overfitting thông qua các tham số regularization.

2.3.3 Neural Network

Neural Network (mạng nơ-ron nhân tạo) là mô hình học sâu có khả năng học được các mối quan hệ phi tuyến phức tạp trong dữ liệu. Trong đề tài này, Neural Network được sử dụng nhằm khai thác sâu hơn cấu trúc tiềm ẩn của dữ liệu lâm sàng.

Ưu điểm của Neural Network bao gồm:

- Khả năng mô hình hóa các mối quan hệ phi tuyến giữa các đặc trưng lâm sàng.
- Cơ chế Early Stopping giúp hạn chế overfitting trong quá trình huấn luyện.
- Tính linh hoạt cao trong việc thiết kế kiến trúc mạng (số lớp, số nơ-ron, hàm kích hoạt).
- Phù hợp với các bài toán phân tích dữ liệu y khoa có tính phức tạp cao.

Việc kết hợp và so sánh ba mô hình trên giúp đề tài đánh giá toàn diện hiệu năng của các phương pháp khác nhau, đồng thời lựa chọn mô hình phù hợp nhất cho bài toán dự đoán bệnh tim trong điều kiện dữ liệu thực tế.

III. MÔ TẢ DỮ LIỆU VÀ BƯỚC TIỀN XỬ LÝ

3.1 Nguồn và đặc điểm dữ liệu

Trong nghiên cứu này, dữ liệu được sử dụng là Heart Disease Dataset thu thập từ các nguồn công khai Kaggle. Đây là bộ dữ liệu y khoa phổ biến, thường được sử dụng trong các nghiên cứu liên quan đến dự đoán bệnh tim, đảm bảo độ tin cậy và khả năng so sánh với các công trình trước.

Bộ dữ liệu bao gồm thông tin lâm sàng của 918 bệnh nhân, phản ánh các yếu tố nguy cơ liên quan đến bệnh tim. Các đặc trưng đầu vào là những biến lâm sàng thường gặp trong thực tế, giúp mô hình học máy có khả năng ứng dụng trong môi trường y tế.

Bảng dưới đây trình bày tóm tắt các đặc điểm chính của bộ dữ liệu:

THÔNG TIN	CHI TIẾT
NGUỒN DỮ LIỆU	Kaggle Heart Disease Dataset
SỐ LƯỢNG MẪU	918 bệnh nhân
SỐ ĐẶC TRƯNG GỐC	12 biến (11 biến lâm sàn, 1 biến mục tiêu)
SỐ ĐẶC TRƯNG SAU FEATURE ENGINEERING	17 biến (bổ sung 5 đặc trưng kỹ thuật)
BIẾN MỤC TIÊU	Bệnh_Tim (0: Khỏe mạnh, 1: Mắc bệnh)
PHÂN BỐ LỚP	410 Khỏe (44.66%) / 508 Bệnh (55.34%)
CHIA DỮ LIỆU	80% Train (734) / 20% Test (184), chia theo tỷ lệ lớp (Stratified)

Xét về phân bố lớp, dữ liệu có sự mất cân bằng nhẹ giữa hai nhóm bệnh và không bệnh, trong đó số lượng bệnh nhân mắc bệnh tim chiếm tỷ lệ cao hơn. Mặc dù mức độ mất cân bằng không quá nghiêm trọng, đây vẫn là yếu tố cần được xem xét trong quá trình huấn luyện và đánh giá mô hình, đặc biệt đối với các chỉ số như Recall.

Sau khi tiến hành Feature Engineering, số lượng đặc trưng được mở rộng từ 12 lên 17 biến nhằm khai thác tốt hơn các mối quan hệ tiềm ẩn trong dữ liệu. Các đặc trưng bổ sung mang tính kỹ thuật và y khoa, giúp nâng cao khả năng phân biệt giữa hai lớp của mô hình học máy.

Dữ liệu được chia thành tập huấn luyện và tập kiểm tra theo tỷ lệ 80/20, đồng thời áp dụng phương pháp Stratified Split để đảm bảo tỷ lệ phân bố lớp giữa hai tập là tương

đồng. Cách chia này giúp quá trình đánh giá mô hình phản ánh chính xác hơn hiệu năng thực tế và hạn chế sai lệch do phân bố dữ liệu không đồng đều.

3.2 Các đặc trưng gốc

Bộ dữ liệu sử dụng trong đề tài bao gồm 12 đặc trưng lâm sàng gốc, phản ánh các yếu tố nguy cơ phổ biến liên quan đến bệnh tim mạch. Các đặc trưng này được thu thập trong quá trình thăm khám và xét nghiệm, có ý nghĩa thực tiễn trong việc đánh giá tình trạng sức khỏe tim mạch của bệnh nhân. Cụ thể, các đặc trưng gốc được mô tả như sau:

1. Tuổi: Tuổi của bệnh nhân, được tính theo năm. Tuổi là một trong những yếu tố nguy cơ quan trọng liên quan đến bệnh tim mạch.
2. Giới_tính: Giới tính của bệnh nhân, bao gồm Nam và Nữ, có ảnh hưởng đến nguy cơ mắc bệnh tim.
3. Loại_Đau_Ngực: Phân loại kiểu đau ngực của bệnh nhân, bao gồm các nhóm: TA (Typical Angina), ATA (Atypical Angina), NAP (Non-Anginal Pain) và ASY (Asymptomatic).
4. Huyết_Áp_Nghỉ: Giá trị huyết áp tâm thu của bệnh nhân khi nghỉ ngơi, đơn vị mmHg, phản ánh tình trạng huyết áp cơ bản.
5. Cholesterol: Nồng độ cholesterol trong máu, đơn vị mg/dL, là yếu tố nguy cơ quan trọng của bệnh tim mạch.
6. Đường_Huyết_Đói: Chỉ số đường huyết lúc đói, được mã hóa nhị phân với giá trị 1 nếu đường huyết > 120 mg/dL, ngược lại là 0.
7. Điện_Tâm_Đồ: Kết quả điện tâm đồ lúc nghỉ, bao gồm các trạng thái Normal, ST và LVH, phản ánh hoạt động điện học của tim.
8. Nhịp_Tim_Tối_Đa: Nhịp tim tối đa đạt được khi thực hiện bài kiểm tra gắng sức, đơn vị bpm (beats per minute).
9. Đau_Thắt_Vận_Động: Biểu thị sự xuất hiện của đau thắt ngực khi vận động, với giá trị N (Không) hoặc Y (Có).
10. Độ_Chênh_ST: Mức độ chênh lệch của đoạn ST (Oldpeak) so với trạng thái nghỉ, đơn vị mm, phản ánh bất thường điện tim khi gắng sức.
11. Độ_Dốc_ST: Độ dốc của đoạn ST trong quá trình gắng sức, bao gồm các trạng thái Up, Flat và Down, có giá trị chẩn đoán trong bệnh tim.
12. [Biến mục tiêu] Bệnh_Tim: Biến nhãn dùng để huấn luyện và đánh giá mô hình, trong đó giá trị 0 biểu thị bệnh nhân không mắc bệnh tim và giá trị 1 biểu thị bệnh nhân mắc bệnh tim.

Các đặc trưng trên đóng vai trò là đầu vào cho các mô hình Machine Learning trong quá trình huấn luyện và dự đoán. Việc lựa chọn các biến lâm sàng phổ biến giúp mô hình có khả năng ứng dụng cao trong thực tế và đảm bảo tính phù hợp với bài toán y tế.

3.3 Xử lý dữ liệu lỗi

Trong quá trình khảo sát dữ liệu thô, một số giá trị bất thường và không hợp lệ đã được phát hiện. Việc nhận diện và xử lý đúng các trường hợp này là cần thiết nhằm đảm bảo chất lượng dữ liệu đầu vào, đồng thời tránh làm sai lệch kết quả huấn luyện của mô hình Machine Learning.

- Cholesterol = 0:
  Có 172 mẫu (chiếm 18.7%) có giá trị cholesterol bằng 0. Trong thực tế y khoa, giá trị cholesterol bằng 0 là không hợp lý và thường biểu thị dữ liệu bị thiếu hoặc không được ghi nhận. Do đó, các giá trị này được xem tương đương với giá trị khuyết (NaN) và được xử lý bằng phương pháp Iterative Imputation nhằm ước lượng giá trị hợp lý dựa trên mối quan hệ với các đặc trưng còn lại.
- Huyết_Áp_Nghỉ = 0:
  Phát hiện 1 mẫu có giá trị huyết áp nghỉ bằng 0, đây cũng là giá trị không hợp lệ về mặt sinh lý. Trường hợp này được xử lý tương tự bằng Iterative Imputation để thay thế bằng giá trị ước lượng phù hợp.
- Độ_Chênh_ST(âm):
  Có 13 mẫu có giá trị độ chênh ST âm. Khác với các trường hợp trên, giá trị âm của đoạn ST có ý nghĩa y khoa quan trọng, có thể liên quan đến hiện tượng ST Elevation, một dấu hiệu đặc trưng của nhồi máu cơ tim cấp. Vì vậy, các giá trị này được giữ nguyên và không thực hiện hiệu chỉnh nhằm đảm bảo bảo toàn thông tin lâm sàng quan trọng.

Việc lựa chọn Iterative Imputation giúp tận dụng mối quan hệ giữa các đặc trưng trong dữ liệu, hạn chế sai lệch so với các phương pháp thay thế đơn giản như trung bình hoặc trung vị. Đồng thời, việc phân biệt rõ giữa giá trị lỗi và giá trị bất thường có ý nghĩa y khoa giúp đảm bảo rằng quá trình tiền xử lý không làm mất đi các thông tin quan trọng phục vụ cho bài toán dự đoán bệnh tim.

Dữ liệu thô chứa một số giá trị bất thường:

- Cholesterol = 0: 172 mẫu (18.7%) - Hiểu là giá trị NaN - Xử lý bằng Iterative Imputation
- Huyết_Áp_Nghỉ = 0: 1 mẫu - Xử lý bằng Iterative Imputation
- Độ_Chênh_ST âm: 13 mẫu - Giữ nguyên (có ý nghĩa y khoa: ST Elevation = Nhồi máu cơ tim cấp)

3.4 Pipeline tiền xử lý dữ liệu

Để đảm bảo tính nhất quán và độ tin cậy trong quá trình huấn luyện và đánh giá mô hình, đề tài xây dựng một pipeline tiền xử lý dữ liệu với thứ tự xử lý chặt chẽ. Việc tuân thủ đúng trình tự các bước là yếu tố quan trọng nhằm tránh hiện tượng rò rỉ dữ

liệu (data leakage) và đảm bảo khả năng tái sử dụng mô hình trong môi trường triển khai thực tế.

Quy trình tiền xử lý dữ liệu được thực hiện theo các bước sau:

1. Load & Clean dữ liệu: Dữ liệu được tải từ tệp CSV, sau đó tiến hành làm sạch ban đầu bao gồm chuẩn hóa tên cột, loại bỏ dấu tiếng Việt và định dạng lại dữ liệu nhằm đảm bảo tính thống nhất trong quá trình xử lý và lập trình.
2. Chia tập Train/Test: Dữ liệu được chia thành hai tập huấn luyện và kiểm tra theo tỷ lệ 80%/20%, sử dụng phương pháp Stratified Split để đảm bảo tỷ lệ phân bố lớp giữa hai tập là tương đồng, hạn chế sai lệch trong quá trình đánh giá mô hình.
3. Mã hóa dữ liệu (Encoding): Các biến phân loại được mã hóa bằng phương pháp Binary Mapping và One-Hot Encoding, giúp chuyển đổi dữ liệu dạng danh mục sang dạng số phù hợp cho các mô hình Machine Learning. Sau bước này, số lượng đặc trưng tăng lên 15 đặc trưng.
4. Điền khuyết dữ liệu (Imputation): Việc điền khuyết được áp dụng cho 4 cột dữ liệu gốc có giá trị thiếu hoặc không hợp lệ. Phương pháp Iterative Imputer được sử dụng và chỉ được huấn luyện (fit) trên tập huấn luyện, sau đó áp dụng (transform) cho tập kiểm tra, nhằm tránh rò rỉ thông tin từ dữ liệu kiểm tra.
5. Feature Engineering: Sau khi hoàn tất bước imputation, các đặc trưng mới được tạo ra, bao gồm Cholesterol_Tuoi và NguyCo_TimMach_RatCao. Việc thực hiện Feature Engineering sau imputation giúp tránh lỗi "feature unseen at fit time" và đảm bảo tính nhất quán của pipeline.
6. Chuẩn hóa dữ liệu (Scaling): Các đặc trưng số được chuẩn hóa bằng phương pháp RobustScaler, áp dụng cho 6 cột dữ liệu số. Phương pháp này giúp giảm ảnh hưởng của các giá trị ngoại lai (outliers) và cải thiện hiệu quả huấn luyện của mô hình.
7. Lưu trữ các thành phần (Artifacts): Toàn bộ các thành phần của pipeline như mô hình, scaler, imputer và danh sách tên đặc trưng được lưu lại để phục vụ cho quá trình đánh giá trên tập kiểm tra cũng như triển khai trong môi trường thực tế.

Lưu ý quan trọng:

Các bước Imputation và Scaling chỉ được fit trên tập huấn luyện, sau đó mới áp dụng cho tập kiểm tra. Việc tuân thủ nguyên tắc này giúp tránh hiện tượng rò rỉ dữ liệu và đảm bảo rằng kết quả đánh giá phản ánh đúng hiệu năng thực tế của mô hình. Đồng thời, bước Feature Engineering được thực hiện sau Imputation nhằm đảm bảo các đặc trưng mới không bị lỗi trong quá trình huấn luyện.

3.5 Feature Engineering - Tạo đặc trưng mới

Nhằm nâng cao khả năng học và dự đoán của mô hình, đề tài tiến hành bước Feature Engineering dựa trên kiến thức y khoa và các yếu tố nguy cơ đã được công nhận trong lâm sàng. Việc tạo thêm các đặc trưng mới giúp mô hình khai thác tốt hơn mối quan hệ tiềm ẩn giữa các biến lâm sàng, từ đó cải thiện hiệu năng dự đoán bệnh tim.

Trong nghiên cứu này, hai đặc trưng mới được xây dựng như sau:

1. Đặc trưng Cholesterol_Tuoi:

Đặc trưng này được xác định theo công thức:

$$\text{Cholesterol\_Tuoi} = \frac{\text{Cholesterol}}{\text{Tuổi}}$$

Đặc trưng Cholesterol_Tuoi phản ánh mức độ cholesterol đã được điều chỉnh theo độ tuổi của bệnh nhân. Trên thực tế lâm sàng, mức cholesterol cao ở người trẻ tuổi thường mang ý nghĩa nguy cơ tim mạch cao hơn so với cùng mức cholesterol ở người lớn tuổi. Do đó, đặc trưng này giúp mô hình nhận diện tốt hơn các trường hợp bệnh nhân trẻ nhưng có nguy cơ bệnh tim tiềm ẩn cao.

2. Đặc trưng NguyCo_TimMach_RatCao:

Đặc trưng này được xây dựng dựa trên điều kiện logic:

$$\text{NguyCo\_TimMach\_RatCao} = \begin{cases} 1, & \text{nếu } (\text{Huyết\_Áp\_Nghi} \geq 140) \wedge (\text{Cholesterol} \geq 240) \\ 0, & \text{ngược lại} \end{cases}$$

Đặc trưng NguyCo_TimMach_RatCao nhằm xác định nhóm bệnh nhân thuộc diện nguy cơ tim mạch đặc biệt cao, khi đồng thời xuất hiện hai yếu tố nguy cơ quan trọng là huyết áp cao (≥ 140 mmHg) và cholesterol cao (≥ 240 mg/dL). Việc kết hợp hai điều kiện này giúp mô hình nhận diện rõ ràng hơn các trường hợp có khả năng mắc bệnh tim nghiêm trọng, từ đó nâng cao hiệu quả phân loại.

IV. MÔ HÌNH HỌC MÁY SỬ DỤNG

4.1 Random Forest - Ensemble Learning

4.1.1 Nguyên lý hoạt động

Random Forest là một thuật toán thuộc nhóm Ensemble Learning, được đề xuất bởi Breiman (2001), với ý tưởng kết hợp nhiều cây quyết định (Decision Trees) nhằm nâng cao độ chính xác và khả năng tổng quát hóa của mô hình. Thay vì dựa vào một cây quyết định duy nhất, Random Forest xây dựng một tập hợp các cây độc lập và đưa ra kết quả dự đoán dựa trên sự tổng hợp của toàn bộ tập cây này.

Nguyên lý hoạt động của Random Forest bao gồm các bước chính sau:

1. Bootstrap Aggregating (Bagging): Từ tập dữ liệu huấn luyện ban đầu, thuật toán tạo ra nhiều tập dữ liệu con bằng cách lấy mẫu ngẫu nhiên có hoàn lại, mỗi tập có kích thước bằng kích thước tập huấn luyện gốc. Mỗi cây quyết định trong rừng được huấn luyện trên một tập dữ liệu con khác nhau, giúp giảm phương sai và hạn chế hiện tượng overfitting.
2. Random Feature Selection: Tại mỗi nút phân chia của cây quyết định, thay vì xem xét toàn bộ các đặc trưng, Random Forest chỉ chọn ngẫu nhiên một tập con gồm $\sqrt{p}$ đặc trưng (với p là tổng số đặc trưng). Cơ chế này giúp tăng tính đa dạng giữa các cây trong rừng và giảm sự phụ thuộc vào các đặc trưng mạnh.
3. Xây dựng cây quyết định: Mỗi cây quyết định được xây dựng bằng cách phân chia các nút dựa trên các tiêu chí như Gini Impurity hoặc Entropy, nhằm tối ưu khả năng phân biệt giữa các lớp dữ liệu.
4. Tổng hợp kết quả (Voting): Đối với bài toán phân loại, kết quả dự đoán cuối cùng của Random Forest được xác định bằng bỏ phiếu đa số (majority voting) từ các cây quyết định. Đối với bài toán hồi quy, kết quả được tính bằng giá trị trung bình của các dự đoán từ các cây.

Nhờ cơ chế kết hợp nhiều mô hình con và tính ngẫu nhiên trong cả dữ liệu lẫn đặc trưng, Random Forest có khả năng hoạt động ổn định, giảm thiểu overfitting và cho hiệu năng tốt trên các tập dữ liệu y khoa dạng bảng.

4.1.2 Công thức toán học

Trong Random Forest, mỗi cây quyết định thực hiện việc phân chia dữ liệu dựa trên các tiêu chí đo lường mức độ "không thuần nhất" của dữ liệu tại từng nút. Một trong những tiêu chí phổ biến nhất được sử dụng là Gini Impurity.

Gini Impurity tại nút tđược xác định theo công thức:

$$G(t) = 1 - \sum_{i=0}^{1} p_i^2$$

trong đó:

- $p_0$là tỷ lệ mẫu thuộc lớp Khỏe tại nút t,
- $p_1$là tỷ lệ mẫu thuộc lớp Bệnh tại nút t.

Giá trị Gini càng nhỏ thì nút càng "thuần", nghĩa là các mẫu trong nút chủ yếu thuộc về một lớp.

Khi thực hiện phân chia một nút cha thành hai nút con, mức độ cải thiện của phép phân chia được đo bằng Information Gain (IG), được tính như sau:

$$IG = G(\text{parent}) - [w_{\text{left}} \times G(\text{left}) + w_{\text{right}} \times G(\text{right})]$$

trong đó:

- G(parent)là Gini Impurity của nút cha,
- G(left), G(right)lần lượt là Gini Impurity của các nút con,
- w_{left}, w_{right}là trọng số, được xác định bằng tỷ lệ số mẫu trong mỗi nút con so với số mẫu trong nút cha.

Đối với Random Forest, mỗi cây quyết định h_tđưa ra một dự đoán độc lập. Dự đoán cuối cùng của mô hình trong bài toán phân loại được xác định theo nguyên tắc bỏ

$$\hat{y} = \text{mode}\{h_1(x), h_2(x), \ldots, h_T(x)\}$$

phiếu đa số:

trong đó:

- $h_t(x)$là kết quả dự đoán của cây quyết định thứ t,
- Tlà tổng số cây trong Random Forest.

Ngoài ra, Random Forest còn có thể ước lượng xác suất bệnh tim cho mỗi mẫu dữ liệu. Xác suất một bệnh nhân thuộc lớp Bệnh (y = 1) được tính bằng tỷ lệ số cây dự đoán bệnh trên tổng số cây trong rừng:

$$P(y = 1 \mid x) = \frac{1}{T} \sum_{t=1}^{T} \mathbb{I}(h_t(x) = 1)$$

trong đó $\mathbb{I}(\cdot)$là hàm chỉ thị (indicator function), nhận giá trị 1 nếu cây quyết định dự đoán bệnh tim và 0 nếu dự đoán bệnh nhân khỏe mạnh.

Việc sử dụng xác suất dự đoán này cho phép điều chỉnh ngưỡng phân loại (threshold), giúp tối ưu các chỉ số đánh giá như Recall, phù hợp với yêu cầu của bài toán y tế.

4.1.3 Ưu điểm và nhược điểm
Ưu điểm

Random Forest sở hữu nhiều ưu điểm nổi bật, đặc biệt phù hợp với các bài toán dự đoán bệnh tim trên dữ liệu y khoa dạng bảng:

- Khả năng chống overfitting tốt: Nhờ cơ chế trung bình hóa kết quả của nhiều cây quyết định, Random Forest giúp giảm phương sai của mô hình và hạn chế hiện tượng quá khớp so với việc sử dụng một cây đơn lẻ.
- Xử lý tốt các giá trị ngoại lai (outliers): Thuật toán không nhạy cảm với các giá trị bất thường trong dữ liệu, giúp mô hình hoạt động ổn định trên dữ liệu y khoa thực tế.

- Không yêu cầu chuẩn hóa dữ liệu: Random Forest có thể hoạt động hiệu quả ngay cả khi dữ liệu chưa được scale, thuận lợi trong quá trình tiền xử lý.
- Khả năng giải thích thông qua Feature Importance: Mô hình cho phép ước lượng mức độ quan trọng của từng đặc trưng, hỗ trợ phân tích và giải thích kết quả dự đoán trong bối cảnh y tế.

Nhược điểm

Bên cạnh các ưu điểm, Random Forest vẫn tồn tại một số hạn chế nhất định:

- Thời gian dự đoán tương đối chậm: Do cần tổng hợp kết quả từ nhiều cây quyết định (thường từ 100 đến 300 cây), thời gian dự đoán của Random Forest có thể chậm hơn so với các mô hình đơn giản.
- Khó giải thích chi tiết logic dự đoán: Mặc dù có thể đánh giá tầm quan trọng của đặc trưng, Random Forest vẫn khó giải thích đầy đủ quá trình ra quyết định cho từng mẫu cụ thể như một cây quyết định đơn lẻ.

Nhìn chung, với những ưu điểm nổi bật và hạn chế có thể chấp nhận được, Random Forest là một lựa chọn phù hợp và hiệu quả cho bài toán dự đoán bệnh tim trong khuôn khổ đề tài này.

4.1.4 Cấu hình mô hình

Trong đề tài này, mô hình Random Forest được cấu hình với các tham số chính nhằm cân bằng giữa hiệu năng dự đoán và khả năng tổng quát hóa của mô hình. Các tham số và ý nghĩa của chúng được trình bày trong Bảng dưới đây:

THAM SỐ	GIÁ TRỊ	Ý NGHĨA	GIẢI THÍCH
N_ESTIMATORS	100 – 300	Số lượng cây quyết định trong rừng	Số cây càng lớn giúp mô hình ổn định hơn nhưng làm tăng thời gian huấn luyện và dự đoán
MAX_DEPTH	10 – 20 hoặc None	Độ sâu tối đa của mỗi cây	Cây quá sâu dễ gây overfitting, trong khi cây quá nông có thể dẫn đến underfitting
MIN_SAMPLES_LEAF	1 – 4	Số mẫu tối thiểu tại mỗi nút lá	Giá trị lớn hơn giúp làm trơn mô hình và hạn chế overfitting
MAX_FEATURES	$\sqrt{p}$	Số đặc trưng được chọn	Với $p = 17$, $\sqrt{p} \approx 4$ đặc trưng, giúp tăng tính đa dạng giữa các cây

		tại mỗi lần phân chia	
RANDOM_STATE	2026	Hạt giống ngẫu nhiên	Đảm bảo khả năng tái tạo kết quả (reproducibility)

Việc lựa chọn các tham số trên được thực hiện dựa trên đặc điểm của bộ dữ liệu, quy mô mẫu và yêu cầu của bài toán y tế. Đặc biệt, việc giới hạn độ sâu của cây và tăng số mẫu tối thiểu tại nút lá giúp mô hình giảm nguy cơ overfitting, trong khi vẫn duy trì khả năng dự đoán chính xác.

Các tham số được tinh chỉnh nhằm đạt hiệu suất tối ưu trên tập huấn luyện, đồng thời đảm bảo khả năng tổng quát hóa tốt khi đánh giá trên tập kiểm tra.

4.1.5 Tối ưu hóa hyperparameters

Để nâng cao hiệu suất dự đoán của mô hình Random Forest, đề tài tiến hành tối ưu hóa các siêu tham số (hyperparameters) bằng phương pháp GridSearchCV kết hợp với 5-fold Cross-Validation. Phương pháp này cho phép đánh giá toàn diện các tổ hợp tham số khác nhau, đồng thời giảm thiểu sai lệch do việc chia dữ liệu ngẫu nhiên.

Không gian tìm kiếm được xây dựng dựa trên các tham số quan trọng của Random Forest, bao gồm n_estimators, max_depth và min_samples_leaf. Tổng số tổ hợp tham số được khảo sát là 3 × 4 × 3 = 36 tổ hợp. Trong quá trình tối ưu, tiêu chí đánh giá được lựa chọn là Recall (scoring = "recall"), nhằm ưu tiên khả năng phát hiện đúng các ca bệnh tim, phù hợp với yêu cầu của bài toán y tế.

Kết quả tối ưu hóa cho thấy bộ tham số tốt nhất của mô hình Random Forest bao gồm:

- n_estimators = 200
- max_depth = 15
- min_samples_leaf = 2

Với cấu hình trên, mô hình đạt Recall trung bình trên tập Cross-Validation là 94.8% ± 1.2%, cho thấy khả năng phát hiện bệnh tim ổn định và đáng tin cậy trên các tập dữ liệu khác nhau.

Kết quả này khẳng định việc tối ưu hyperparameters dựa trên chỉ số Recall là phù hợp và hiệu quả đối với bài toán dự đoán bệnh tim, đồng thời tạo tiền đề cho việc đánh giá và so sánh mô hình trên tập kiểm tra ở các chương tiếp theo.

4.2 XGBoost - Gradient Boosting

4.2.1 Nguyên lý hoạt động

XGBoost (eXtreme Gradient Boosting), được đề xuất bởi Chen và Guestrin (2016), là một thuật toán học máy thuộc nhóm Boosting, trong đó các cây quyết định được

xây dựng tuần tự (sequential). Khác với Random Forest – nơi các cây được huấn luyện song song và độc lập – XGBoost xây dựng mỗi cây mới với mục tiêu sửa các lỗi còn tồn tại của mô hình trước đó, từ đó cải thiện dần hiệu năng dự đoán.

Nguyên lý hoạt động của XGBoost có thể được mô tả qua các bước sau:

1. Khởi tạo mô hình
   Mô hình ban đầu được khởi tạo với một giá trị hằng:

   $$F_0(x) = \log\left(\frac{p}{1-p}\right)$$

   trong đó plà tỷ lệ mẫu thuộc lớp bệnh trong tập huấn luyện. Giá trị này tương ứng với log-odds của xác suất mắc bệnh tim.
2. Tính phần dư (Residuals)
   Tại mỗi vòng lặp $t = 1,2,\ldots,T$, mô hình tính phần dư:

   $$r = y - \hat{y}$$

   phản ánh sai lệch giữa nhãn thực tế và giá trị dự đoán hiện tại của mô hình.

3. Huấn luyện cây mới
   Một cây quyết định mới $h_t(x)$được xây dựng nhằm dự đoán phần dư, thay vì dự đoán trực tiếp nhãn y. Cách tiếp cận này giúp mô hình tập trung vào các mẫu mà các cây trước đó dự đoán chưa chính xác.
4. Cập nhật mô hình
   Mô hình tổng được cập nhật theo công thức:

   $$F_t(x) = F_{t-1}(x) + \eta \times h_t(x)$$

   trong đó ηlà learning rate, kiểm soát mức độ đóng góp của mỗi cây mới nhằm tránh việc mô hình học quá nhanh và gây overfitting.
5. Regularization
   XGBoost áp dụng regularization (L2) lên trọng số của các nút lá trong cây quyết định, giúp kiểm soát độ phức tạp của mô hình và hạn chế hiện tượng overfitting, đặc biệt quan trọng đối với các tập dữ liệu y khoa có quy mô vừa và nhỏ.

Nhờ cơ chế học dần từ lỗi, kết hợp với các kỹ thuật tối ưu hóa và regularization, XGBoost thường đạt hiệu năng rất cao trên các bài toán phân loại dữ liệu dạng bảng, bao gồm cả bài toán dự đoán bệnh tim trong nghiên cứu này.

4.2.2 Công thức toán học chi tiết

XGBoost tối ưu hóa mô hình thông qua việc cực tiểu hóa hàm mục tiêu (Objective Function), bao gồm hàm mất mát và thành phần regularization nhằm kiểm soát độ

$$\mathcal{L}(\phi) = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{T} \Omega(f_k)$$

phức tạp của mô hình:

trong đó:

- $l(y_i, \hat{y}_i)$là hàm mất mát (loss function), với bài toán phân loại nhị phân sử dụng log loss,
- $\Omega(f_k)$là hạng tử regularization của cây thứ k,
- Tlà số cây trong mô hình.

Thành phần regularization của mỗi cây được xác định như sau:

$$\Omega(f) = \gamma T + \frac{\lambda}{2}\|w\|^2$$

trong đó:

- T là số lượng nút lá của cây,
- wlà vector trọng số tại các nút lá,
- γvà λlà các hệ số phạt, giúp hạn chế độ phức tạp của mô hình và giảm overfitting.

Hàm mất mát Log Loss (Binary Cross-Entropy)

Đối với bài toán phân loại nhị phân, hàm mất mát được sử dụng là Binary Cross-Entropy, được xác $l(y, \hat{y}) = -[y \cdot \log(p) + (1-y) \cdot \log(1-p)]$ định như sau:

trong đó:

$$p = \sigma(\hat{y}) = \frac{1}{1+e^{-\hat{y}}}$$

là xác suất dự đoán của mô hình, được tính thông qua hàm sigmoid.

Gradient và Hessian

XGBoost sử dụng thông tin đạo hàm bậc nhất và bậc hai của hàm mất mát để xây dựng cây quyết định. Với mỗi mẫu dữ liệu i, ta có:

- Gradient (đạo hàm bậc nhất): $g_i = \frac{\partial l}{\partial \hat{y}_i} = p_i - y_i$
- Hessian (đạo hàm bậc hai): $h_i = \frac{\partial^2 l}{\partial \hat{y}_i^2} = p_i(1-p_i)$

Các giá trị gradient và hessian này phản ánh mức độ sai lệch và độ cong của hàm mất mát, đóng vai trò quan trọng trong việc tối ưu hóa mô hình.

Trọng số tối ưu tại nút lá

Trọng số tối ưu tại nút lá jcủa cây quyết định được tính theo công thức:

$$w_j^* = -\frac{\sum g_i}{\sum h_i + \lambda}$$

trong đó tổng∑được thực hiện trên tất cả các mẫu dữ liệu rơi vào nút lá j. Công thức này giúp xác định mức điều chỉnh phù hợp cho mỗi nút lá, đồng thời chịu sự chi phối của tham số regularization λ.

Dự đoán cuối cùng

Giá trị dự đoán cuối của mô hình sau Tvòng boosting được xác định như sau:

$$p = \sigma(F_T(x)) = \sigma\left(F_0(x) + \eta \sum_{t=1}^{T} h_t(x)\right)$$

trong đó:

- $F_0(x)$là giá trị khởi tạo ban đầu,
- ηlà learning rate,
- $h_t(x)$là cây quyết định thứ t.

Cuối cùng, nhãn dự đoán được xác định dựa trên ngưỡng phân loại (threshold):

$$\hat{y} = \begin{cases} 1, & \text{nếu } p \geq \text{threshold} \\ 0, & \text{nếu } p < \text{threshold} \end{cases}$$

Việc điều chỉnh ngưỡng phân loại cho phép mô hình ưu tiên các chỉ số như Recall, phù hợp với yêu cầu của bài toán dự đoán bệnh tim trong lĩnh vực y tế.

4.2.3 Ưu điểm độc đáo của XGBoost

XGBoost sở hữu nhiều đặc điểm nổi bật giúp thuật toán này đạt hiệu năng cao trên các bài toán phân loại dữ liệu dạng bảng, đặc biệt trong lĩnh vực y tế. Những ưu điểm độc đáo của XGBoost trong đề tài này có thể được tổng hợp như sau:

Ưu điểm

- Regularization tích hợp mạnh mẽ: XGBoost hỗ trợ đồng thời L1 (Lasso) và L2 (Ridge) regularization, giúp kiểm soát độ phức tạp của mô hình, hạn chế overfitting và nâng cao khả năng tổng quát hóa trên dữ liệu y khoa.

- Khả năng xử lý giá trị thiếu: Thuật toán có cơ chế tự động học hướng phân chia tối ưu cho các giá trị bị thiếu, do đó không bắt buộc phải thực hiện imputation trước khi huấn luyện, giúp giảm sai lệch do việc ước lượng giá trị thiếu.

- Hỗ trợ dữ liệu mất cân bằng: Tham số scale_pos_weight cho phép điều chỉnh trọng số của lớp thiểu số, giúp cải thiện khả năng phát hiện các ca bệnh tim trong bối cảnh dữ liệu không cân bằng.
- Khả năng xử lý song song: XGBoost hỗ trợ parallel processing trong quá trình tìm kiếm điểm phân chia tốt nhất, tận dụng hiệu quả nhiều lõi CPU, từ đó rút ngắn thời gian huấn luyện.
- Cơ chế cắt tỉa cây (Tree Pruning): Cây quyết định được cắt tỉa từ lá lên gốc, giúp xác định độ sâu thực tế của cây và hạn chế việc xây dựng các nhánh không cần thiết, góp phần giảm overfitting.

Hạn chế

- Nhạy cảm với nhiễu (noise): Do đặc thù của Boosting là học từ các lỗi còn tồn tại, XGBoost có thể vô tình học theo các giá trị ngoại lai nếu dữ liệu không được xử lý cẩn thận.
- Yêu cầu tinh chỉnh tham số cẩn thận: XGBoost có nhiều hyperparameters cần được tối ưu, đòi hỏi thời gian và kinh nghiệm trong quá trình tinh chỉnh để đạt hiệu năng tốt nhất.

Nhờ những ưu điểm nổi bật trên, XGBoost được xem là một trong những mô hình mạnh mẽ và hiệu quả nhất cho bài toán dự đoán bệnh tim trong đề tài này, đặc biệt khi được cấu hình và tối ưu phù hợp.

4.2.4 Cấu hình mô hình

Trong đề tài này, mô hình XGBoost được cấu hình với các tham số quan trọng nhằm đạt được sự cân bằng giữa hiệu năng dự đoán và khả năng tổng quát hóa, đặc biệt trong bối cảnh dữ liệu y khoa có nguy cơ mất cân bằng. Các tham số và ý nghĩa của chúng được trình bày trong Bảng dưới đây:

THAM SỐ	GIÁ TRỊ	Ý NGHĨA	GIẢI THÍCH
N_ESTIMATORS	100 – 200	Số lượng cây (số vòng boosting)	Số cây lớn giúp cải thiện hiệu năng nhưng có thể gây overfitting nếu quá lớn
MAX_DEPTH	3 – 5	Độ sâu tối đa của mỗi cây	Cây nông giúp hạn chế overfitting và phù hợp với Boosting

LEARNING_RATE (H)	0.01 – 0.1	Tốc độ học	Giá trị nhỏ giúp mô hình học ổn định hơn nhưng cần nhiều cây hơn
SUBSAMPLE	0.8	Tỷ lệ mẫu sử dụng trong mỗi vòng boosting	Giảm overfitting, tương tự cơ chế bootstrap
COLSAMPLE_BYTREE	0.8	Tỷ lệ đặc trưng sử dụng cho mỗi cây	Tăng tính đa dạng giữa các cây, tương tự Random Forest
SCALE_POS_WEIGHT	1 – 2	Trọng số lớp bệnh	Điều chỉnh sự mất cân bằng giữa lớp bệnh và lớp khỏe
OBJECTIVE	binary:logistic	Hàm mục tiêu	Phân loại nhị phân sử dụng log loss

Việc lựa chọn các tham số trên được thực hiện dựa trên đặc điểm của bộ dữ liệu và yêu cầu ưu tiên Recall trong bài toán dự đoán bệnh tim. Đặc biệt, tham số scale_pos_weight được sử dụng để tăng trọng số cho lớp bệnh, giúp mô hình cải thiện khả năng phát hiện các ca bệnh tim trong bối cảnh dữ liệu không cân bằng.

Các tham số này được tinh chỉnh thông qua quá trình thử nghiệm và đánh giá nhằm đạt hiệu năng tối ưu trước khi tiến hành so sánh với các mô hình khác trong các chương tiếp theo.

4.2.5 Threshold Tuning - Tối ưu ngưỡng cắt

Theo mặc định, mô hình XGBoost sử dụng ngưỡng phân loại 0.5, trong đó một mẫu được dự đoán thuộc lớp bệnh nếu xác suất dự đoán $p \geq 0.5$. Tuy nhiên, đối với các bài toán trong lĩnh vực y tế, việc sử dụng ngưỡng mặc định này có thể dẫn đến việc bỏ sót bệnh nhân (False Negative), gây hậu quả nghiêm trọng. Do đó, đề tài tiến hành tối ưu hóa ngưỡng phân loại (Threshold Tuning) nhằm đạt được sự cân bằng hợp lý giữa Precision và Recall, với tiêu chí ưu tiên Recall cao hơn Precision.

Phương pháp thực hiện

Quy trình tối ưu ngưỡng phân loại được thực hiện như sau:

1. Tính xác suất dự đoán $p_i = P(y = 1 \mid x_i)$ cho tất cả các mẫu trong tập validation.

2. Thực hiện quét ngưỡng phân loại ttrong khoảng từ 0.1 đến 0.9, với bước tăng 0.01.
3. Với mỗi giá trị ngưỡng t, tiến hành tính toán các chỉ số đánh giá:
   - **Precision(t):**

$$\text{Precision}(t) = \frac{TP(t)}{TP(t) + FP(t)}$$

   - **Recall(t):**

$$\text{Recall}(t) = \frac{TP(t)}{TP(t) + FN(t)}$$

   - **F1-score(t):**

$$F1(t) = \frac{2 \times \text{Precision}(t) \times \text{Recall}(t)}{\text{Precision}(t) + \text{Recall}(t)}$$

4. Lựa chọn ngưỡng tối ưu t^*theo một trong hai tiêu chí:
   - Tối đa hóa F1-score, hoặc
   - Thỏa mãn điều kiện Recall ≥ 95% và có Precision cao nhất.

Kết quả đạt được

Quá trình Threshold Tuning cho kết quả ngưỡng phân loại tối ưu:

- Ngưỡng tối ưu t^*: 0.6711
- Recall tại t^*: 96.08%
- Precision tại t^*: 77.17%
- F1-score tại t^*: 0.8559

So sánh với ngưỡng mặc định

So với ngưỡng mặc định 0.5, việc tối ưu ngưỡng mang lại những thay đổi đáng chú ý:

- Recall tăng: từ 94.10% lên 96.08% (+1.98%)
- Precision giảm: từ 83.50% xuống 77.17% (–6.33%)

Mặc dù Precision có giảm, sự đánh đổi này được xem là phù hợp trong bối cảnh y tế, khi việc tăng khả năng phát hiện bệnh nhân mắc bệnh tim là ưu tiên hàng đầu. Do đó, ngưỡng tối ưu $t^* = 0.6711$được lựa chọn cho các bước đánh giá và triển khai mô hình tiếp theo.

4.3 Neural Network - Deep Learning

4.3.1 Nguyên lý hoạt động:
Trong đề tài này, mô hình Neural Network được sử dụng dưới dạng Multi-Layer Perceptron (MLP) – một mạng nơ-ron truyền thẳng gồm nhiều lớp ẩn. MLP có khả năng học và biểu diễn các mối quan hệ phi tuyến phức tạp giữa các đặc trưng lâm sàng, điều mà các mô hình tuyến tính khó thực hiện hiệu quả.

Mỗi nơ-ron trong mạng thực hiện phép biến đổi tuyến tính trên đầu vào, sau đó áp dụng hàm kích hoạt phi tuyến để tạo ra đầu ra. Quá trình huấn luyện mạng được thực hiện thông qua thuật toán Backpropagation, trong đó sai số dự đoán được lan truyền ngược từ lớp đầu ra về các lớp ẩn nhằm cập nhật trọng số theo hướng giảm hàm mất mát.

Nhờ khả năng học phi tuyến, Neural Network có tiềm năng khai thác các tương tác phức tạp giữa các yếu tố nguy cơ tim mạch, đặc biệt phù hợp với dữ liệu lâm sàng đa chiều.

4.3.2 Cấu hình mô hình:
Mô hình Neural Network được cấu hình với các tham số chính như trình bày trong Bảng dưới đây:

THAM SỐ	GIÁ TRỊ	Ý NGHĨA
HIDDEN_LAYER_SIZES	(64, 32), (128, 64, 32), (32, 16)	Kiến trúc các lớp ẩn của mạng
ALPHA	0.0001 – 0.01	Hệ số L2 Regularization, phạt các trọng số lớn
LEARNING_RATE_INIT	0.001 – 0.01	Tốc độ học ban đầu
MAX_ITER	1000	Số vòng lặp huấn luyện tối đa
EARLY_STOPPING	True	Dừng sớm khi validation loss không cải thiện
SOLVER	adam	Thuật toán tối ưu hóa Adam

Việc thử nghiệm nhiều kiến trúc lớp ẩn cho phép đánh giá ảnh hưởng của độ sâu và độ rộng mạng đến hiệu năng mô hình. Các tham số được lựa chọn nhằm cân bằng giữa khả năng học của mạng và nguy cơ overfitting.

4.3.3 Chống Overfitting:
Do Neural Network có số lượng tham số lớn, nguy cơ overfitting là một thách thức quan trọng, đặc biệt khi dữ liệu huấn luyện có quy mô vừa. Để khắc phục vấn đề này, đề tài áp dụng các cơ chế sau:

- Early Stopping: Quá trình huấn luyện sẽ tự động dừng nếu validation loss không được cải thiện trong 10 vòng lặp liên tiếp. Cơ chế này giúp mô hình tránh việc học thuộc dữ liệu huấn luyện và tập trung học các mẫu tổng quát.

- L2 Regularization (alpha): Việc thêm hạng tử phạt vào hàm mất mát giúp hạn chế các trọng số quá lớn, từ đó giảm độ phức tạp của mô hình và cải thiện khả năng tổng quát hóa.

Nhờ các cơ chế trên, mô hình Neural Network được huấn luyện một cách ổn định và có khả năng so sánh hiệu quả với các mô hình Machine Learning truyền thống trong bài toán dự đoán bệnh tim.

V. KẾT QUẢ VÀ ĐÁNH GIÁ MÔ HÌNH

5.1 Bảng so sánh kết quả trên tập Test

Hiệu năng của các mô hình được đánh giá trên tập Test độc lập thông qua các chỉ số phổ biến trong bài toán phân loại nhị phân, bao gồm Accuracy, Precision, Recall, F1-score và AUC-ROC. Đặc biệt, trong bối cảnh y tế, Recall được ưu tiên hàng đầu nhằm giảm thiểu nguy cơ bỏ sót bệnh nhân mắc bệnh tim.

Bảng dưới đây trình bày kết quả so sánh hiệu năng của ba mô hình:

MÔ HÌNH	ACCURACY	PRECISION	RECALL	F1-SCORE	AUC-ROC	NGƯỠNG
XGBOOST	82.07%	77.17%	96.08%	0.8559	0.9064	0.6711
RANDOM FOREST	88.04%	85.09%	95.10%	0.8981	0.9232	0.5491
NEURAL NETWORK	86.96%	84.21%	94.12%	0.8889	0.9449	0.6184

Nhận xét và phân tích kết quả

Kết quả cho thấy mỗi mô hình đều có những ưu điểm riêng. Random Forest đạt Accuracy và F1-score cao nhất, cho thấy khả năng dự đoán tổng thể tốt. Neural Network đạt giá trị AUC-ROC cao nhất, thể hiện khả năng phân biệt hai lớp bệnh và không bệnh rất tốt trên toàn bộ dải ngưỡng.

Tuy nhiên, trong bài toán dự đoán bệnh tim, Recall là chỉ số quan trọng nhất, bởi việc bỏ sót bệnh nhân mắc bệnh tim (False Negative) có thể dẫn đến hậu quả nghiêm trọng trong thực tế lâm sàng. Xét theo tiêu chí này, XGBoost đạt Recall cao nhất là 96.08%, vượt trội so với hai mô hình còn lại.

Cụ thể, với mô hình XGBoost:

- Mô hình phát hiện đúng 98 trên 102 bệnh nhân mắc bệnh tim,
- Chỉ bỏ sót 4 trường hợp, mức bỏ sót thấp nhất trong ba mô hình

Ngoài ra, XGBoost vẫn duy trì được AUC-ROC cao (0.9064), cho thấy khả năng phân biệt tốt giữa hai lớp, đồng thời cho phép điều chỉnh ngưỡng phân loại linh hoạt để phù hợp với mục tiêu y tế. Bên cạnh đó, mô hình XGBoost còn thể hiện khả năng phát hiện đa dạng các mẫu ẩn (Hidden Patterns), với 5 nhóm nguy cơ khác nhau, giúp nâng cao giá trị phân tích và hỗ trợ ra quyết định.

Kết luận lựa chọn mô hình

Dựa trên các kết quả đánh giá và yêu cầu của bài toán, XGBoost được lựa chọn là mô hình tốt nhất cho đề tài, với các lý do chính sau:

- Recall cao nhất (96.08%), đáp ứng tiêu chí ưu tiên phát hiện bệnh nhân mắc bệnh tim.
- AUC-ROC cao, đảm bảo khả năng phân biệt tốt giữa các nhóm bệnh nhân.
- Khả năng phát hiện các mẫu ẩn đa dạng, hỗ trợ phân tích sâu dữ liệu lâm sàng.
- Phù hợp nhất cho ứng dụng y tế, nơi việc hạn chế bỏ sót bệnh nhân là ưu tiên hàng đầu.

Mặc dù Accuracy và Precision của XGBoost thấp hơn một số mô hình khác, sự đánh đổi này được xem là chấp nhận được trong bối cảnh y tế, khi mục tiêu chính là giảm thiểu nguy cơ bỏ sót bệnh nhân mắc bệnh tim.

5.2 Chi tiết đánh giá XGBoost (Mô hình tốt nhất)

Để đánh giá chi tiết hiệu năng của mô hình XGBoost, đề tài sử dụng Confusion Matrix trên tập Test nhằm phân tích rõ các trường hợp dự đoán đúng và sai, đặc biệt là các ca bệnh bị bỏ sót – yếu tố có ý nghĩa quan trọng trong lĩnh vực y tế.

Confusion Matrix trên tập Test:

- True Negative (TN): 65 (Bệnh nhân khỏe mạnh được dự đoán đúng là khỏe)
- False Positive (FP): 19 (Bệnh nhân khỏe bị dự đoán nhầm là mắc bệnh)
- False Negative (FN): 4 (Bệnh nhân mắc bệnh nhưng bị dự đoán nhầm là khỏe – trường hợp cần tối thiểu hóa)
- True Positive (TP): 98 (Bệnh nhân mắc bệnh được dự đoán đúng)

Kết quả cho thấy mô hình XGBoost chỉ bỏ sót 4 trên tổng số 102 bệnh nhân mắc bệnh tim, thể hiện khả năng phát hiện bệnh nhân rất tốt.

Giải thích các chỉ số đánh giá

- Accuracy (Độ chính xác tổng thể)

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} = \frac{98 + 65}{184} = 82.07\%$$

Accuracy phản ánh tỷ lệ dự đoán đúng trên toàn bộ dữ liệu. Tuy nhiên, trong bài toán y tế, đây không phải tiêu chí quan trọng nhất, do Accuracy cao vẫn có thể đi kèm với việc bỏ sót nhiều bệnh nhân.

- Precision (Độ chính xác của dự đoán dương tính

$$\text{Precision} = \frac{TP}{TP + FP} = \frac{98}{98 + 19} = 77.17\%$$

Chỉ số Precision cho biết trong số các bệnh nhân được mô hình dự đoán là mắc bệnh, có khoảng 77% thực sự mắc bệnh. Điều này đồng nghĩa với việc có 19 bệnh nhân khỏe bị cảnh báo nhầm, mức đánh đổi này được xem là chấp nhận được trong y tế để đổi lấy Recall cao hơn.

$$\text{Recall} = \frac{TP}{TP + FN} = \frac{98}{98 + 4} = 96.08\%$$

- Recall (Độ nhạy – chỉ số quan trọng nhất)

Recall thể hiện khả năng phát hiện đúng các bệnh nhân mắc bệnh tim. Với Recall = 96.08%, mô hình XGBoost phát hiện được 96% bệnh nhân mắc bệnh, chỉ bỏ sót 4 trường hợp, đây là kết quả rất tốt và phù hợp với yêu cầu thực tiễn y tế.

- F1-score
  F1-score là trung bình điều hòa giữa Precision và Recall:

$$F1 = 0.8559$$

Chỉ số này cho thấy mô hình đạt được sự cân bằng hợp lý giữa khả năng phát hiện bệnh và độ chính xác của dự đoán dương tính.

- AUC-ROC
  Giá trị AUC-ROC = 0.9064 phản ánh khả năng phân biệt giữa hai lớp Khỏe và Bệnh trên toàn bộ dải ngưỡng phân loại. Giá trị AUC cao cho thấy mô hình có khả năng phân loại tốt và ổn định.

Nhận xét tổng quát

Từ các phân tích trên, có thể kết luận rằng mô hình XGBoost đạt hiệu năng cao và đặc biệt phù hợp với bài toán dự đoán bệnh tim trong lĩnh vực y tế. Mặc dù Precision chưa phải là cao nhất, mô hình vẫn được ưu tiên lựa chọn nhờ Recall vượt trội, giúp giảm thiểu nguy cơ bỏ sót bệnh nhân – yếu tố được xem là quan trọng nhất trong các ứng dụng hỗ trợ chẩn đoán y khoa.

5.3 Phát hiện mẫu ẩn (Hidden Patterns)

Bên cạnh việc đánh giá hiệu năng dự đoán, đề tài tiến hành phân tích các mẫu ẩn (Hidden Patterns) được mô hình XGBoost học được nhằm khám phá những tổ hợp triệu chứng lâm sàng có mối liên hệ chặt chẽ với nguy cơ mắc bệnh tim. Việc phát

hiện các mẫu ẩn này giúp nâng cao khả năng giải thích mô hình và hỗ trợ ra quyết định trong bối cảnh y tế.

Thông qua phân tích cấu trúc cây và các điều kiện phân chia quan trọng, mô hình XGBoost đã xác định được 5 nhóm nguy cơ với các đặc trưng lâm sàng điển hình. Trong đó, các nhóm nguy cơ cao và cực cao được mô tả như sau:

Nhóm nguy cơ cực cao #1

- Điều kiện:

$$\text{Độ\_Chênh\_ST} > 0.67 \ \wedge \ \text{Đường\_Huyết\_Đói} > 0$$

- Nguy cơ mắc bệnh tim: 92%
- Số lượng bệnh nhân: 57 người
- Ý nghĩa lâm sàng: Nhóm này bao gồm các bệnh nhân có dấu hiệu điện tâm đồ bất thường rõ rệt kết hợp với đường huyết đói cao. Sự kết hợp giữa rối loạn điện tim và yếu tố chuyển hóa làm gia tăng đáng kể nguy cơ mắc bệnh tim mạch, được xem là tình trạng nguy cấp cần được theo dõi và can thiệp sớm.

Nhóm nguy cơ cực cao #2

- Điều kiện:

$$\text{Độ\_Chênh\_ST} > 0.75$$

- Nguy cơ mắc bệnh tim: 92%
- Số lượng bệnh nhân: 57 người
- Ý nghĩa lâm sàng: Giá trị độ chênh ST rất cao phản ánh tình trạng ST depression nghiêm trọng trên điện tâm đồ – một dấu hiệu thường liên quan đến thiếu máu cơ tim hoặc tổn thương tim. Ngay cả khi không kết hợp thêm yếu tố khác, đặc trưng này vẫn đủ để xác định nhóm bệnh nhân có nguy cơ mắc bệnh tim rất cao.

Nhóm nguy cơ cao #3

- Điều kiện:

$$\text{Độ\_Chênh\_ST} > 0.67 \ \wedge \ \text{Đường\_Huyết\_Đói} > 0$$

(ở mức độ nhẹ hơn so với nhóm #1)

- Nguy cơ mắc bệnh tim: 77%
- Số lượng bệnh nhân: 36 người

- Ý nghĩa lâm sàng:
  Nhóm này có tổ hợp triệu chứng tương tự nhóm nguy cơ cực cao #1, tuy nhiên mức độ bất thường thấp hơn. Điều này cho thấy nguy cơ mắc bệnh tim vẫn ở mức cao và cần được theo dõi chặt chẽ, đặc biệt trong các chương trình sàng lọc sớm.

Nhận xét tổng quát

Qua phân tích các mẫu ẩn, có thể nhận thấy rằng ba nhóm nguy cơ cao nhất đều liên quan trực tiếp đến hai đặc trưng quan trọng là Độ_Chênh_ST (ST Depression) và Đường_Huyết_Đói. Đây là hai chỉ báo lâm sàng có giá trị cao trong việc đánh giá nguy cơ bệnh tim, phù hợp với các kết luận trong y văn.

Kết quả này cho thấy mô hình XGBoost không chỉ đạt hiệu năng dự đoán cao mà còn có khả năng khai thác và làm nổi bật các tổ hợp triệu chứng có ý nghĩa y khoa, góp phần nâng cao tính giải thích và khả năng ứng dụng của hệ thống trong thực tế lâm sàng.

5.4 Giải thích mô hình - Feature Importance

Để tăng tính minh bạch và khả năng giải thích của mô hình XGBoost, đề tài sử dụng phương pháp Permutation Importance nhằm đánh giá mức độ ảnh hưởng của từng đặc trưng đến kết quả dự đoán. Phương pháp này đo lường sự suy giảm hiệu năng của mô hình khi hoán vị ngẫu nhiên giá trị của một đặc trưng, từ đó phản ánh tầm quan trọng thực sự của đặc trưng đó đối với mô hình.

Kết quả Permutation Importance cho thấy các đặc trưng có ảnh hưởng lớn nhất đến dự đoán nguy cơ bệnh tim được xếp hạng như sau:

- Độ_Chênh_ST (0.45): Đây là đặc trưng quan trọng nhất trong mô hình. Độ chênh ST phản ánh mức độ bất thường trên điện tâm đồ, đặc biệt liên quan đến hiện tượng ST depression, một dấu hiệu quan trọng của thiếu máu cơ tim. Việc đặc trưng này có mức độ ảnh hưởng cao nhất là hoàn toàn phù hợp với kiến thức y khoa và các kết quả phân tích mẫu ẩn đã trình bày ở mục trước.
- Đường_Huyết_Đói (0.28): Đường huyết đói cao là yếu tố nguy cơ tim mạch quan trọng, liên quan chặt chẽ đến rối loạn chuyển hóa và đái tháo đường. Giá trị importance cao của đặc trưng này cho thấy mô hình đã học được mối liên hệ giữa tăng đường huyết và nguy cơ mắc bệnh tim.
- Cholesterol_Tuoi (0.15): Đây là đặc trưng được tạo ra trong bước Feature Engineering, phản ánh mức độ cholesterol đã được điều chỉnh theo độ tuổi. Việc đặc trưng này có tầm quan trọng cao cho thấy chiến lược Feature Engineering dựa trên kiến thức y khoa là hiệu quả và góp phần cải thiện khả năng dự đoán của mô hình.

- Nhịp_Tim_Tối_Đa (0.08): Nhịp tim tối đa đạt được khi gắng sức có mức độ ảnh hưởng trung bình. Đặc trưng này phản ánh khả năng đáp ứng tim mạch của bệnh nhân trong điều kiện vận động, có giá trị hỗ trợ trong việc đánh giá nguy cơ bệnh tim.
- Tuổi (0.03): Tuổi có mức độ ảnh hưởng thấp hơn so với các đặc trưng chức năng và sinh lý, cho thấy mô hình không chỉ dựa vào yếu tố tuổi tác mà tập trung nhiều hơn vào các chỉ số lâm sàng phản ánh trực tiếp tình trạng tim mạch.

Nhận xét tổng quát

Kết quả Feature Importance cho thấy mô hình XGBoost ưu tiên các chỉ số chức năng và sinh lý tim mạch hơn là các đặc trưng nhân khẩu học đơn thuần. Điều này phù hợp với thực tế lâm sàng và củng cố độ tin cậy của mô hình trong việc hỗ trợ dự đoán bệnh tim.

Đặc biệt, việc các đặc trưng như Độ_Chênh_ST, Đường_Huyết_Đói và Cholesterol_Tuoi giữ vai trò quan trọng nhất cho thấy sự nhất quán giữa kết quả học máy, phân tích mẫu ẩn và kiến thức y khoa.

VI. KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

6.1 Kết luận chính

Thông qua quá trình nghiên cứu, xây dựng mô hình và đánh giá thực nghiệm, đề tài Dự đoán bệnh tim bằng Machine Learning đã đạt được những kết luận quan trọng như sau:

1. Hiệu suất mô hình

Hệ thống dự đoán bệnh tim đạt được hiệu suất cao, đặc biệt với mô hình XGBoost khi đạt Recall = 96.08% trên tập Test. Kết quả này cho thấy mô hình có khả năng phát hiện phần lớn bệnh nhân mắc bệnh tim, chỉ bỏ sót một số rất ít trường hợp. Đây là yếu tố then chốt và rất phù hợp cho các ứng dụng y tế, nơi việc giảm thiểu rủi ro bỏ sót bệnh nhân được ưu tiên hàng đầu.

2. Mẫu ẩn quan trọng trong dữ liệu

Kết quả phân tích Hidden Patterns cho thấy hai đặc trưng Độ_Chênh_ST (ST Depression) và Đường_Huyết_Đói là những yếu tố có ảnh hưởng mạnh nhất đến nguy cơ mắc bệnh tim. Đặc biệt, các bệnh nhân có đồng thời hai chỉ số này bất thường thuộc nhóm nguy cơ cao hoặc cực cao, cần được theo dõi và can thiệp sớm trong thực tế lâm sàng.

3. Sự khác biệt của 3 mô hình

Ba mô hình được triển khai trong đề tài đều thể hiện những thế mạnh riêng:

- Random Forest đạt Accuracy cao nhất (88.04%), cho thấy khả năng dự đoán tổng thể tốt.
- XGBoost đạt Recall cao nhất (96.08%), phù hợp nhất với yêu cầu ưu tiên phát hiện bệnh nhân trong y tế.
- Neural Network đạt AUC-ROC cao nhất (0.9449), thể hiện khả năng phân biệt hai lớp rất tốt trên toàn bộ dải ngưỡng.

Sự so sánh này cho thấy không tồn tại một mô hình "tốt nhất tuyệt đối", mà việc lựa chọn mô hình phụ thuộc vào mục tiêu ứng dụng cụ thể của bài toán.

4. Tính khoa học và khả năng giải thích

Việc áp dụng Feature Engineering dựa trên kiến thức y khoa, tiêu biểu như các đặc trưng Cholesterol_Tuoi và NguyCo_TimMach_RatCao, không chỉ giúp cải thiện hiệu năng mô hình mà còn nâng cao khả năng giải thích kết quả dự đoán. Kết hợp với phân tích Feature Importance và Hidden Patterns, mô hình cho thấy tính nhất quán cao giữa kết quả học máy và hiểu biết y học.

5. Khả năng ứng dụng thực tiễn

Hệ thống được tích hợp dưới dạng ứng dụng web Streamlit, cho phép người dùng nhập thông tin lâm sàng của bệnh nhân và nhận kết quả dự đoán ngay lập tức. Giao

diện trực quan và dễ sử dụng giúp hệ thống có tiềm năng trở thành công cụ hỗ trợ ra quyết định lâm sàng, góp phần nâng cao hiệu quả sàng lọc và đánh giá nguy cơ bệnh tim.

6.2 Hướng phát triển trong tương lai

Mặc dù đề tài đã đạt được những kết quả tích cực trong khuôn khổ môn học, hệ thống dự đoán bệnh tim vẫn còn nhiều tiềm năng để tiếp tục mở rộng và hoàn thiện trong tương lai. Một số hướng phát triển chính có thể được xem xét như sau:

1. Mở rộng và đa dạng hóa dữ liệu

Thu thập thêm dữ liệu lâm sàng từ nhiều nguồn khác nhau, đặc biệt là từ các bệnh viện và cơ sở y tế trong nước, nhằm tăng quy mô mẫu và tính đại diện của dữ liệu. Việc mở rộng dữ liệu sẽ giúp mô hình có độ tin cậy cao hơn và khả năng tổng quát hóa tốt hơn khi áp dụng vào thực tế.

2. Kết hợp dữ liệu hình ảnh y khoa

Bổ sung các loại dữ liệu hình ảnh như X-quang tim, siêu âm tim hoặc CT tim mạch, kết hợp với dữ liệu lâm sàng dạng bảng để xây dựng các mô hình đa phương thức (multi-modal). Hướng tiếp cận này có thể giúp cải thiện đáng kể độ chính xác và khả năng phát hiện sớm bệnh tim.

3. Xây dựng mô hình dự đoán toàn diện

Phát triển các mô hình multi-task learning, cho phép dự đoán đồng thời nhiều loại bệnh tim mạch hoặc biến cố tim mạch khác nhau. Điều này giúp hệ thống trở nên toàn diện hơn và phù hợp với các bài toán sàng lọc y tế thực tế.

4. Nâng cao khả năng giải thích mô hình

Áp dụng các phương pháp Explainable AI như SHAP (SHapley Additive exPlanations) để giải thích chi tiết từng dự đoán ở mức độ cá nhân bệnh nhân. Việc này giúp bác sĩ hiểu rõ hơn nguyên nhân dẫn đến kết quả dự đoán và tăng mức độ tin cậy khi sử dụng hệ thống.

5. Phân tầng nguy cơ bệnh nhân

Xây dựng hệ thống phân tầng nguy cơ thay vì chỉ phân loại nhị phân, chia bệnh nhân thành nhiều mức nguy cơ khác nhau (ví dụ: rất thấp, thấp, trung bình, cao, rất cao). Cách tiếp cận này giúp hỗ trợ triage và ưu tiên nguồn lực y tế hiệu quả hơn.

6. Phát triển ứng dụng di động

Triển khai hệ thống dưới dạng ứng dụng di động, giúp các bác sĩ và nhân viên y tế, đặc biệt ở các khu vực vùng sâu, vùng xa, có thể tiếp cận nhanh chóng công cụ hỗ trợ chẩn đoán ngay cả khi điều kiện hạ tầng còn hạn chế.

7. Tích hợp với hệ thống bệnh viện
Kết nối hệ thống dự đoán với các hệ thống quản lý bệnh viện như HIS (Hospital Information System) hoặc EMR (Electronic Medical Record) để tự động hóa quá trình thu thập dữ liệu và hỗ trợ ra quyết định lâm sàng theo thời gian thực.

8. Theo dõi sức khỏe theo thời gian
Phát triển các mô hình theo dõi liên tục (longitudinal analysis), cho phép phân tích sự thay đổi của các chỉ số y tế theo thời gian. Hướng tiếp cận này giúp phát hiện sớm các dấu hiệu bất thường và dự đoán nguy cơ bệnh tim trong tương lai một cách chủ động.

6.3 Ghi chú

Lưu ý quan trọng: Hệ thống dự đoán bệnh tim được xây dựng trong đề tài này nhằm mục đích hỗ trợ chẩn đoán và sàng lọc nguy cơ, không thay thế cho chẩn đoán y khoa chính thức. Kết quả dự đoán của mô hình chỉ mang tính tham khảo và cần được xác nhận bởi bác sĩ chuyên khoa thông qua thăm khám lâm sàng toàn diện và các xét nghiệm cần thiết.

Trong trường hợp mô hình dự đoán bệnh nhân thuộc nhóm "không mắc bệnh", nhưng bệnh nhân vẫn có các triệu chứng lâm sàng rõ ràng hoặc dấu hiệu bất thường, việc đánh giá y khoa không được dừng lại dựa trên kết quả của mô hình. Khi đó, cần tiến hành các xét nghiệm chuyên sâu như chụp CT tim, MRI tim, hoặc can thiệp chẩn đoán mạch vành (catheterization) để đưa ra kết luận chính xác.

Việc sử dụng hệ thống Machine Learning trong y tế cần được thực hiện một cách thận trọng, có trách nhiệm, và luôn đặt quyết định chuyên môn của bác sĩ làm trung tâm. Mô hình chỉ đóng vai trò công cụ hỗ trợ ra quyết định, giúp nâng cao hiệu quả sàng lọc và giảm tải cho hệ thống y tế, nhưng không thay thế vai trò của con người trong chẩn đoán và điều trị.

VII. TÀI LIỆU THAM KHẢO

1. Breiman, L. (2001). Random Forests. Machine Learning, 45(1), 5–32.
2. Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference.
3. Kingma, D. P., & Ba, J. (2014). Adam: A Method for Stochastic Optimization. arXiv:1412.6980.
4. Lundberg, S. M., & Lee, S. I. (2017). A Unified Approach to Interpreting Model Predictions. Advances in Neural Information Processing Systems (NeurIPS), 30.
5. Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning (2nd ed.). Springer.
6. World Health Organization (2023). Cardiovascular Diseases (CVDs).
7. UCI Machine Learning Repository. Heart Disease Dataset.
8. Scikit-learn Documentation. Machine Learning library for Python

PHỤ LỤC: HƯỚNG DẪN SỬ DỤNG HỆ THỐNG

Toàn bộ mã nguồn, hướng dẫn cài đặt, huấn luyện mô hình và triển khai ứng dụng của đề tài được công bố công khai trên GitHub tại địa chỉ:

📎 https://github.com/nlmhau/Heart_Disease_Prediction_ML

Repository cung cấp đầy đủ các nội dung sau:

- Mã nguồn tiền xử lý dữ liệu và huấn luyện mô hình (Random Forest, XGBoost, Neural Network)
- Hướng dẫn cài đặt môi trường và các thư viện cần thiết
- Hướng dẫn chạy ứng dụng web minh họa bằng Streamlit
- Các mô hình đã huấn luyện và kết quả thực nghiệm

Việc cung cấp mã nguồn dưới dạng repository giúp đảm bảo tính minh bạch, khả năng tái lập kết quả (reproducibility) và tạo điều kiện thuận lợi cho việc mở rộng, đánh giá hoặc triển khai hệ thống trong tương lai.