ỦY BAN NHÂN DÂN THÀNH PHỐ HỒ CHÍ MINH

TRƯỜNG ĐẠI HỌC SÀI GÒN

KHOA CÔNG NGHỆ THÔNG TIN

BÁO CÁO TỔNG KẾT

Môn Học: Phương Pháp Nghiên Cứu Khoa Học

Đề Tài: Dự Đoán Bệnh Tim Bằng Máy Học Và Sử Dụng Kỹ Thuật Ensemble Learning

Giảng viên: Đỗ Như Tài

Nhóm

Phạm Tấn Khương – 3122410191

Hoàng Vũ - 3122560089

Huỳnh Thanh Bình - 3122410033

Nguyễn Minh Tú - 3120411167

Thành phố Hồ Chí Minh , tháng 5 năm 2025

Mục Lục

LỜI CẢM ƠN

Để hoàn thành dự án nghiên cứu khoa học "Dự Đoán Bệnh Tim Bằng Máy Học Và Sử Dụng Kỹ Thuật Ensemble Learning" chúng em đã nhận được rất nhiều sự giúp đỡ và hỗ trợ tận tình.

Chúng em xin trân trọng gửi lời cảm ơn sâu sắc đến:

- Khoa Công Nghệ Thông Tin – Trường Đại Học Sài Gòn đã tạo mọi điều kiện thuận lợi để chúng em có thể thực hiện nghiên cứu này.
- Chúng em xin gửi lời tri ân đến thầy Đỗ Như Tài đã tận tình hướng dẫn, chỉ bảo trong suốt quá trình thực hiện đề tài. Sự định hướng và hỗ trợ của thầy đã giúp chúng em hoàn thành bài nghiên cứu một cách thuận lợi và hiệu quả.
- Các thành viên trong nhóm đã luôn đoàn kết, hỗ trợ lẫn nhau và nỗ lực hết mình để hoàn thành dự án với kết quả tốt nhất.

Cuối cùng, chúng em xin kính chúc các thầy cô luôn mạnh khỏe, thành công để tiếp tục dìu dắt các thế hệ học sinh, sinh viên trên con đường học tập và nghiên cứu.

Tóm Tắt (Abstract)

Bệnh tim mạch là một trong những nguyên nhân gây tử vong hàng đầu trên toàn thế giới. Việc chẩn đoán sớm và chính xác đóng vai trò quan trọng trong việc điều trị và ngăn ngừa biến chứng. Tuy nhiên, các phương pháp chẩn đoán truyền thống thường đòi hỏi thời gian, chi phí và chuyên môn cao. Trước thực trạng đó, nghiên cứu này được thực hiện nhằm xây dựng một mô hình dự đoán bệnh tim dựa trên các kỹ thuật học máy, giúp hỗ trợ bác sĩ và người bệnh trong quá trình ra quyết định.

Mục tiêu chính của nghiên cứu là ứng dụng kỹ thuật Ensemble Learning – cụ thể là phương pháp Voting Ensemble kết hợp giữa mô hình SVM và Logistic Regression – để nâng cao độ chính xác dự đoán so với việc sử dụng từng mô hình đơn lẻ. Dữ liệu huấn luyện được tiền xử lý bằng các kỹ thuật chuẩn hóa và mã hóa phù hợp, sau đó huấn luyện hai mô hình riêng biệt và kết hợp dự đoán đầu ra bằng cách trung bình hóa kết quả.

Kết quả thử nghiệm cho thấy mô hình kết hợp đạt độ chính xác cao hơn so với từng mô hình thành phần, đồng thời giao diện ứng dụng trực quan được xây dựng bằng Tkinter giúp dễ dàng sử dụng. Nghiên cứu kết luận rằng phương pháp Ensemble Learning có tiềm năng ứng dụng thực tiễn trong lĩnh vực y tế, đặc biệt là trong hỗ trợ chẩn đoán bệnh tim.

Chương 1: Tổng Quan Vấn Đề

1.1 Lý do chọn đề tài

Nghiên cứu dự đoán bệnh tim bằng máy học có nhiều lý do quan trọng, bao gồm:

1. **Phát hiện sớm và phòng ngừa**: Các mô hình máy học có thể giúp phát hiện bệnh tim ngay từ giai đoạn sớm, khi các triệu chứng chưa rõ ràng. Điều này có thể giúp bác sĩ can thiệp kịp thời, giảm nguy cơ tử vong hoặc các biến chứng nghiêm trọng.
2. **Cải thiện độ chính xác**: Máy học có khả năng xử lý và phân tích một lượng lớn dữ liệu y tế, bao gồm các yếu tố nguy cơ (như huyết áp, mức cholesterol, thói quen sinh hoạt, lịch sử gia đình) để đưa ra dự đoán chính xác hơn so với việc chẩn đoán chỉ dựa vào kinh nghiệm của bác sĩ.
3. **Giảm tải cho bác sĩ**: Với sự hỗ trợ của các mô hình máy học, các bác sĩ có thể nhanh chóng phân loại và đánh giá nguy cơ bệnh tim của bệnh nhân, từ đó giảm bớt khối lượng công việc và tập trung vào các ca bệnh phức tạp hơn.
4. **Cải thiện quản lý sức khỏe cộng đồng**: Dự đoán bệnh tim bằng máy học giúp các tổ chức y tế phân tích xu hướng bệnh tật trong cộng đồng, từ đó thiết kế các chiến lược phòng ngừa và can thiệp hợp lý hơn
5. **Ứng dụng trong hệ thống chăm sóc sức khỏe tự động**: Máy học có thể được tích hợp vào các ứng dụng y tế và hệ thống chăm sóc sức khỏe, cung cấp các dự đoán và cảnh báo cho người dùng và các chuyên gia y tế ngay lập tức, tạo ra một hệ thống chăm sóc sức khỏe tiên tiến và chủ động.
6. **Khả năng học từ dữ liệu lớn**: Các mô hình máy học có thể học và cải thiện từ những dữ liệu lớn, đồng thời phân tích mối quan hệ phức tạp giữa các yếu tố nguy cơ và bệnh lý mà không cần hiểu rõ tất cả các yếu tố đó một cách chi tiết.
7. **Hiệu quả cao từ các kỹ thuật Ensemble Learning**: Các kỹ thuật học máy kết hợp (Ensemble Learning) như Random Forest, XGBoost, hoặc Voting Classifier cho phép tổng hợp sức mạnh của nhiều mô hình đơn lẻ để cải thiện độ chính xác, độ tin cậy, và khả năng tổng quát hóa của mô hình. Việc ứng dụng Ensemble Learning giúp giảm rủi ro mô hình dự đoán sai, từ đó hỗ trợ bác sĩ trong chẩn đoán một cách hiệu quả hơn.

Tóm lại, nghiên cứu dự đoán bệnh tim bằng máy học không chỉ giúp tăng cường khả năng chẩn đoán và điều trị mà còn góp phần vào việc tối ưu hóa và cải thiện hệ thống chăm sóc sức khỏe.

1.2 Vấn đề nghiên cứu

Vấn đề nghiên cứu của chúng em là bệnh tim mạch là một trong những nguyên nhân gây tử vong hàng đầu trên toàn thế giới. Việc chẩn đoán bệnh tim thường phụ thuộc vào kinh nghiệm của bác sĩ và các xét nghiệm lâm sàng, tuy nhiên quá trình này có thể tốn thời gian, chi phí cao và đôi khi không phát hiện được bệnh ở giai đoạn sớm. Trong khi đó, nhiều yếu tố nguy cơ như tuổi tác, huyết áp, cholesterol, tiểu sử bệnh lý, thói quen sinh hoạt... có thể được thu thập dễ dàng và cung cấp dữ liệu phong phú để hỗ trợ quá trình chẩn đoán.

Máy học (Machine Learning) – một nhánh của trí tuệ nhân tạo – có khả năng xử lý và phân tích dữ liệu phức tạp, từ đó phát hiện các mẫu và mối quan hệ tiềm ẩn giữa các yếu tố đầu vào và nguy cơ mắc bệnh tim. Tuy nhiên, việc áp dụng máy học vào dự đoán bệnh tim cũng đặt ra nhiều thách thức, như lựa chọn thuật toán phù hợp, xử lý dữ liệu thiếu, đảm bảo độ chính xác, độ tin cậy của mô hình và khả năng triển khai thực tế trong môi trường y tế.

Do đó, đề tài "Dự đoán bệnh tim bằng máy học" được đặt ra nhằm mục tiêu nghiên cứu và xây dựng một mô hình có khả năng dự đoán chính xác nguy cơ mắc bệnh tim từ dữ liệu bệnh nhân, hỗ trợ các bác sĩ trong công tác chẩn đoán và phòng ngừa bệnh một cách hiệu quả hơn.

Đặc biệt, nghiên cứu này tập trung vào việc áp dụng và đánh giá các kỹ thuật Ensemble Learning – một hướng tiếp cận kết hợp nhiều mô hình học máy khác nhau nhằm nâng cao độ chính xác, giảm sai số và tăng khả năng tổng quát hóa. Với tiềm năng nổi bật trong việc cải thiện hiệu suất dự đoán, Ensemble Learning được xem là hướng đi triển vọng trong lĩnh vực chẩn đoán y tế hỗ trợ bằng trí tuệ nhân tạo.

1.3 Mục tiêu nghiên cứu

Mục tiêu chính của nghiên cứu là đào tạo ra mô hình máy học với độ chính xác cao nhất và thời gian xử lý nhanh nhất để dự đoán bệnh tim.

Mục tiêu chính của nghiên cứu là xây dựng và tối ưu hóa mô hình máy học sử dụng kỹ thuật Ensemble Learning (như Random Forest, XGBoost, Voting...) nhằm đạt được độ chính xác cao và thời gian xử lý nhanh trong dự đoán bệnh tim.

Đặc biệt, nghiên cứu hướng đến việc đánh giá hiệu suất của các mô hình đơn lẻ, sau đó lựa chọn những mô hình có độ chính xác cao để kết hợp lại thành mô hình Ensemble, nhằm khai thác ưu điểm của từng mô hình và nâng cao hiệu quả tổng thể.

1.4 Câu hỏi nghiên cứu

Chúng em đặt ra 4 câu hỏi cần nghiên cứu:

1. 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??
2. Các yếu tố nào trong dữ liệu (ví dụ: huyết áp, cholesterol, tuổi, giới tính, nhịp tim...) có mức độ ảnh hưởng cao đến kết quả dự đoán nguy cơ mắc bệnh tim ?
3. Làm thế nào để đánh giá hiệu suất của mô hình dự đoán một cách toàn diện ?
4. Làm thế nào để tối ưu hóa mô hình máy học để cân bằng giữa độ chính xác và thời gian xử lý, nhằm phục vụ tốt trong môi trường thực tế?

1.5 Phạm vi nghiên cứu

1.5.1 Đối tượng nghiên cứu

Đối tượng nghiên cứu của đề tài bao gồm các mô hình máy học và dữ liệu y tế của bệnh nhân dùng để dự đoán nguy cơ mắc bệnh tim. Cụ thể:

- Nghiên cứu tập trung vào việc áp dụng các mô hình máy học phổ biến như Logistic Regression, K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Decision Tree, Random Forest, và XGBoost. Từ đó, lựa chọn ra hai mô hình có hiệu suất cao nhất để kết hợp bằng kỹ thuật Ensemble Learning

(ví dụ: Voting hoặc Stacking), với mục tiêu tăng độ chính xác và độ ổn định trong việc phân loại nguy cơ mắc bệnh tim.

- Về dữ liệu: Dữ liệu được sử dụng trong nghiên cứu là từ các bộ dữ liệu y tế công khai, đó là UCI Heart Disease Dataset thuộc UCI data Repository. Bộ dữ liệu này bao gồm các thông tin lâm sàng của bệnh nhân như tuổi, giới tính, huyết áp, mức cholesterol, nhịp tim, chỉ số đường huyết, và kết quả chẩn đoán bệnh tim.

1.5.2 Không gian và thời gian nghiên cứu

Nghiên cứu không đi sâu vào các khía cạnh y học chuyên sâu như chẩn đoán hình ảnh (ECG, MRI), dữ liệu thời gian thực hoặc can thiệp lâm sàng. Phạm vi nghiên cứu chủ yếu mang tính mô phỏng và thử nghiệm trong môi trường học thuật, chưa triển khai trong hệ thống y tế thực tế. Nghiên cứu này chỉ kéo dài trong khoảng 15 tuần

Chương 2: Lược Khảo Tài Liệu (Literature Review)

2.1 Tổng hợp các tài liệu, nghiên cứu trước liên quan

- **Ensemble Framework for Cardiovascular Disease Prediction** đề xuất một mô hình ensemble sử dụng các thuật toán như ExtraTrees Classifier, Random Forest và XGBoost, đạt độ chính xác 92,34% trên bộ dữ liệu kết hợp từ nhiều nguồn khác nhau. arXiv
- **An Improved Heart Disease Prediction Using Stacked Ensemble Method** sử dụng phương pháp ensemble chồng (stacked ensemble) kết hợp 9 thuật toán như RF, MLP, KNN, XGBoost, SVM, Decision Tree, AdaBoost, Gradient Boosting và MLP, đạt độ chính xác gần 100% trên bộ dữ liệu bệnh tim. arXiv
- **An Intelligent Decision Support Ensemble Voting Model for Coronary Artery Disease Prediction in Smart Healthcare Monitoring Environments** áp dụng mô hình ensemble voting với các thuật toán như Random Forest, XGBoost và MLP, đạt độ chính xác 88,12% trong dự đoán bệnh động mạch vành. arXiv
- **Heart Disease Prediction Using Machine Learning Algorithms nghiên cứu của Jindal et al. (2021)** trình bày việc áp dụng các thuật toán học máy như Logistic Regression, K-Nearest Neighbors (KNN) và Random Forest để dự đoán bệnh tim. Sử dụng bộ dữ liệu từ UCI với 13 thuộc tính y tế của 304 bệnh nhân, mô hình KNN đạt độ chính xác cao nhất là 88,52%, tiếp theo là Logistic Regression với 87,5%. Nghiên cứu nhấn mạnh rằng việc sử

dụng các thuật toán học máy có thể hỗ trợ hiệu quả trong việc chẩn đoán sớm bệnh tim, giúp giảm chi phí và nâng cao chất lượng chăm sóc y tế. Studocu

- **Heart Disease Prediction Using Machine Learning Algorithms** Nghiên cứu của Govardhan Logabiraman và cộng sự đề xuất một mô hình hybrid kết hợp các thuật toán như ANN, Gradient Boosting, Decision Tree, SVM, Random Forest và Logistic Regression. Mô hình này đạt độ chính xác cao trong việc dự đoán bệnh tim, cho thấy hiệu quả của việc kết hợp nhiều thuật toán học máy trong một mô hình duy nhất.

2.2 Cơ sở lý thuyết

2.2.1 Học máy (Machine Learning)

Học máy là một nhánh của trí tuệ nhân tạo (AI) tập trung vào việc phát triển các thuật toán và mô hình cho phép hệ thống tự động học hỏi từ dữ liệu mà không cần phải lập trình cụ thể cho từng tình huống. Học máy có thể được phân loại thành ba loại chính dựa trên cách thức học và dữ liệu sử dụng:

- **Học có giám sát (Supervised Learning):**
  Là phương pháp học từ dữ liệu có nhãn, tức là dữ liệu đã được gắn nhãn đầu ra (target). Mô hình học có giám sát sử dụng các đặc điểm đầu vào (features) để dự đoán nhãn đầu ra (label). Ví dụ, trong bài toán phân loại bệnh tim, các đặc điểm như tuổi, huyết áp, cholesterol có thể được sử dụng để dự đoán liệu bệnh nhân có mắc bệnh tim hay không.

Các thuật toán phổ biến trong học có giám sát bao gồm:

  - **Logistic Regression:** Một mô hình hồi quy xác suất thường được sử dụng trong các bài toán phân loại nhị phân. Logistic Regression có khả năng dự đoán xác suất của một lớp thuộc vào một trong hai nhãn.
  - **Support Vector Machines (SVM):** Thuật toán phân loại mạnh mẽ, đặc biệt phù hợp với dữ liệu có biên giới phân tách rõ ràng. SVM cố gắng tìm ra một hyperplane tối ưu để phân loại dữ liệu thành hai lớp.
  - **Decision Trees:** Cây quyết định là một phương pháp phân loại hoặc hồi quy dựa trên việc chia nhỏ dữ liệu vào các nhánh để đưa ra dự đoán.
  - **Random Forests:** Là một tập hợp của nhiều cây quyết định, giúp cải thiện độ chính xác và giảm thiểu hiện tượng overfitting so với cây quyết định đơn lẻ.
  - **Artificial Neural Networks (ANNs):** Mô hình học máy mô phỏng cách thức hoạt động của não bộ con người, thích hợp cho việc xử lý các mối quan hệ phức tạp trong dữ liệu.

- **Học không giám sát (Unsupervised Learning):**
  Học không giám sát áp dụng cho các bài toán mà dữ liệu không có nhãn đầu ra. Các thuật toán học không giám sát thường được sử dụng trong các bài toán phân cụm (clustering) hoặc giảm chiều (dimensionality reduction). Ví dụ, phân nhóm bệnh nhân có đặc điểm y tế tương tự để nghiên cứu các phân nhóm bệnh lý.

- **Học tăng cường (Reinforcement Learning):**
  Trong học tăng cường, mô hình học thông qua việc tương tác với môi trường và nhận các phản hồi từ hành động của mình. Phương pháp này ít được sử dụng trong các bài toán phân loại như dự đoán bệnh tim, mà chủ yếu ứng dụng trong các bài toán điều khiển hoặc các trò chơi.

2.2.2 Kỹ thuật Ensemble Learning

Kỹ thuật Ensemble Learning kết hợp nhiều mô hình học máy lại với nhau nhằm cải thiện độ chính xác dự đoán và giảm thiểu sai sót so với việc sử dụng một mô hình đơn lẻ. Có ba phương pháp phổ biến trong Ensemble Learning:

- **Bagging (Bootstrap Aggregating):** Tạo ra nhiều mô hình học máy độc lập từ các mẫu dữ liệu khác nhau và kết hợp kết quả dự đoán của các mô hình này để tạo ra một kết quả chung. Một ví dụ điển hình là **Random Forests**, nơi mỗi cây quyết định trong rừng được huấn luyện trên các tập dữ liệu con khác nhau và cuối cùng dự đoán của chúng được tổng hợp lại.

- **Boosting:** Kỹ thuật này tạo ra một mô hình học máy mạnh mẽ bằng cách kết hợp nhiều mô hình yếu. Các mô hình tiếp theo trong chuỗi sẽ học hỏi và khắc phục sai sót của các mô hình trước đó. Ví dụ về kỹ thuật boosting bao gồm **AdaBoost** và **XGBoost**.

- **Stacking:** Trong stacking, các mô hình học máy khác nhau (có thể là mô hình mạnh hoặc yếu) được huấn luyện trên cùng một tập dữ liệu, và kết quả của các mô hình này sau đó được đưa vào một mô hình học máy cấp cao hơn để tạo ra kết quả cuối cùng. Kỹ thuật này thường giúp cải thiện đáng kể độ chính xác so với các mô hình đơn lẻ.

Trong nghiên cứu này, **kỹ thuật Voting Ensemble** được sử dụng, cụ thể là **Average Voting**, nơi các mô hình như **Random Forest** và **Logistic Regression** được kết hợp để đưa ra một dự đoán cuối cùng. Đối với mỗi dự đoán, mô hình sẽ đưa ra một xác suất (hoặc nhãn) và sau đó các kết quả này sẽ được trung bình để tạo ra một dự đoán chung.

2.2.3 Kết hợp các mô hình: Random Forest và Logistic Regression

Trong nghiên cứu này, chúng tôi lựa chọn kết hợp hai mô hình học máy là **Random Forest** và **Logistic Regression** nhằm tận dụng ưu điểm của từng mô hình, từ đó nâng cao hiệu quả dự đoán bệnh tim.

Random Forest

Random Forest là một thuật toán học máy thuộc nhóm ensemble learning, hoạt động bằng cách xây dựng nhiều cây quyết định (decision trees) và kết hợp kết quả của chúng để đưa ra dự đoán cuối cùng. Mỗi cây trong rừng được huấn luyện trên một tập con ngẫu nhiên của dữ liệu, giúp giảm thiểu hiện tượng overfitting và tăng độ chính xác của mô hình.

Ưu điểm của Random Forest:

- **Khả năng xử lý dữ liệu phức tạp**: Có thể xử lý dữ liệu với nhiều đặc trưng và mối quan hệ phi tuyến tính.
- **Độ chính xác cao**: Thường đạt hiệu suất tốt trong các bài toán phân loại.
- **Khả năng đánh giá tầm quan trọng của biến**: Giúp xác định các đặc trưng quan trọng trong việc dự đoán bệnh tim.

Logistic Regression

Logistic Regression là một thuật toán học máy đơn giản nhưng hiệu quả, đặc biệt trong các bài toán phân loại nhị phân. Mô hình này ước lượng xác suất của một biến phụ thuộc nhị phân dựa trên các biến độc lập, thông qua hàm sigmoid.

Ưu điểm của Logistic Regression:

- **Dễ triển khai và giải thích**: Mô hình đơn giản, dễ hiểu và dễ triển khai trong thực tế.
- **Hiệu quả với dữ liệu tuyến tính**: Hoạt động tốt khi mối quan hệ giữa các biến là tuyến tính.
- **Khả năng dự đoán xác suất**: Cung cấp xác suất dự đoán, hữu ích trong việc đánh giá rủi ro.

Kết hợp mô hình

Việc kết hợp Random Forest và Logistic Regression nhằm tận dụng ưu điểm của cả hai mô hình. Trong nghiên cứu này, chúng tôi sử dụng phương pháp **Voting Ensemble**, cụ thể là **Average Voting**, để kết hợp dự đoán từ hai mô hình. Phương pháp này giúp:

- **Tăng độ chính xác**: Kết hợp dự đoán từ hai mô hình có thể cải thiện hiệu suất tổng thể.
- **Giảm thiểu overfitting**: Việc kết hợp mô hình giúp giảm thiểu rủi ro overfitting so với việc sử dụng một mô hình đơn lẻ.
- **Cải thiện khả năng tổng quát hóa**: Mô hình kết hợp có khả năng tổng quát hóa tốt hơn trên dữ liệu mới.

2.3 Điểm mạnh, điểm yếu của các nghiên cứu trước và cách nghiên cứu kế thừa hoặc phát triển

2.3.1 Điểm mạnh của các nghiên cứu trước

- **Đa dạng thuật toán:** Các nghiên cứu trước đã thử nghiệm và đánh giá nhiều thuật toán học máy khác nhau, bao gồm Logistic Regression, Decision Trees, Random Forests, SVM và ANN. Những nghiên cứu này giúp xác định rõ ràng hiệu quả của từng thuật toán trong việc dự đoán bệnh tim, đồng thời cung cấp thông tin quý giá về khả năng của từng mô hình đối với các loại dữ liệu khác nhau.
- **Ứng dụng thực tiễn:** Các mô hình học máy đã được áp dụng thành công trong việc giúp các bác sĩ chẩn đoán bệnh tim sớm và chính xác hơn, cải thiện khả năng phòng ngừa và điều trị bệnh lý tim mạch. Các kết quả nghiên cứu này đóng góp vào việc cải thiện chất lượng chăm sóc sức khỏe và giảm tỷ lệ tử vong do bệnh tim.

2.3.2 Điểm yếu của các nghiên cứu trước

- **Độ chính xác thấp với dữ liệu không đầy đủ:** Mặc dù có sự đa dạng về thuật toán, nhưng độ chính xác của các mô hình đơn lẻ trong các nghiên cứu trước còn hạn chế, đặc biệt là khi dữ liệu có thiếu hụt hoặc không cân bằng. Điều này có thể ảnh hưởng nghiêm trọng đến khả năng dự đoán của các mô hình trong môi trường thực tế.
- **Khả năng tổng quát kém:** Các mô hình học máy đơn giản đôi khi không thể bắt được các mối quan hệ phức tạp trong dữ liệu, dẫn đến khả năng tổng quát kém khi mô hình áp dụng vào các bộ dữ liệu mới. Việc thiếu khả năng xử lý dữ liệu phi tuyến tính và phức tạp khiến mô hình dễ bị overfitting hoặc underfitting.
- **Thiếu kết hợp giữa các mô hình mạnh mẽ:** Một số nghiên cứu chỉ sử dụng các mô hình riêng lẻ mà không kết hợp nhiều mô hình học máy mạnh mẽ với nhau (hybrid models). Điều này hạn chế khả năng nâng cao độ chính xác của dự đoán, nhất là đối với các bài toán phức tạp như dự đoán bệnh tim.

2.3.3 Cách nghiên cứu kế thừa và phát triển:

- **Kế thừa:** Nghiên cứu này kế thừa các mô hình học máy đã được chứng minh hiệu quả trong việc dự đoán bệnh tim, như **Logistic Regression** và **Support Vector Machine (SVM)**, từ đó phát huy các ưu điểm đã được công nhận trong các nghiên cứu trước.
- **Phát triển:** Nghiên cứu này phát triển bằng cách áp dụng **kỹ thuật Ensemble Learning**, cụ thể là **Voting Ensemble** với phương pháp **Average Voting**, kết hợp hai mô hình mạnh mẽ là **SVM** và **Logistic Regression** để cải thiện độ chính xác của dự đoán. Việc kết hợp các mô hình giúp giảm thiểu sai sót của các mô hình đơn lẻ và tăng cường tính ổn định trong dự đoán.
    - **Giảm thiểu hiện tượng overfitting và underfitting:** Kỹ thuật Ensemble giúp mô hình không chỉ dựa vào một mô hình duy nhất mà kết hợp nhiều mô hình khác nhau, làm giảm thiểu hiện tượng overfitting và cải thiện khả năng tổng quát.

    - **Cải thiện với dữ liệu không cân đối:** Nghiên cứu này còn áp dụng các phương pháp tiền xử lý dữ liệu như **chuẩn hóa** và **cân bằng lớp** để cải thiện hiệu quả mô hình đối với các tập dữ liệu không cân bằng hoặc có nhiều dữ liệu thiếu.

- **Đóng góp mới:** Nghiên cứu này bổ sung vào các nghiên cứu trước bằng cách sử dụng **Voting Ensemble** kết hợp các mô hình mạnh mẽ, một cách tiếp cận còn hạn chế trong các nghiên cứu trước đây. Bằng cách này, mô hình không chỉ học từ một thuật toán, mà tận dụng sự kết hợp của nhiều thuật toán để đưa ra dự đoán chính xác hơn và ổn định hơn.

Chương 3: Phương Pháp Nghiên Cứu (Methodology)

3.1 Thiết kế nghiên cứu

Nghiên cứu này sử dụng phương pháp nghiên cứu định lượng để phân tích và dự đoán bệnh tim thông qua các thuật toán học máy. Cụ thể, nghiên cứu tập trung vào việc xây dựng và đánh giá các mô hình học máy để phân loại dữ liệu bệnh nhân và đưa ra dự đoán chính xác về khả năng mắc bệnh tim. Phương pháp này giúp đưa ra những kết luận cụ thể và có thể đo lường được về hiệu quả của các mô hình học máy trong dự đoán bệnh tim.

Ngoài ra, nghiên cứu sử dụng phương pháp hỗn hợp kết hợp với kỹ thuật Ensemble Learning nhằm cải thiện độ chính xác của các mô hình học máy như Support Vector Machine (SVM) và Logistic Regression. Việc sử dụng phương pháp này giúp tối ưu hóa kết quả dự đoán thông qua sự kết hợp của nhiều mô hình.

Bổ sung chi tiết:
Quy trình nghiên cứu được xây dựng theo hướng tiếp cận hệ thống, bao gồm các bước:

- Khảo sát và tổng hợp tài liệu, xác định vấn đề nghiên cứu, mục tiêu và câu hỏi nghiên cứu.
- Thu thập, chọn lọc và xử lý dữ liệu đảm bảo chất lượng và tính đại diện.
- Xây dựng, huấn luyện và đánh giá các mô hình học máy đơn lẻ (SVM, Logistic Regression).
- Kết hợp mô hình bằng kỹ thuật Ensemble Learning (Voting Ensemble) để nâng cao hiệu quả dự đoán.
- Đánh giá hiệu quả mô hình bằng các chỉ số thống kê và phân tích kết quả.
- Đề xuất hướng phát triển tiếp theo dựa trên kết quả thực nghiệm.

Việc thiết kế nghiên cứu như vậy giúp đảm bảo tính khách quan, khoa học, có thể lặp lại và phù hợp với thực tiễn triển khai trong lĩnh vực y tế.

3.2 Đối tượng và mẫu nghiên cứu

Đối tượng nghiên cứu trong dự án này là các bệnh nhân tiềm năng mắc bệnh tim, với các đặc điểm nhân khẩu học như tuổi tác, giới tính, huyết áp, cholesterol và tiền sử bệnh lý. Mẫu nghiên

cứu được lấy từ các dữ liệu bệnh nhân thực tế có sẵn trong các cơ sở y tế hoặc các tập dữ liệu công khai, chẳng hạn như tập dữ liệu Cleveland Heart Disease.

Cách chọn mẫu nghiên cứu dựa trên những đặc điểm cụ thể của bệnh tim. Các bệnh nhân có các yếu tố nguy cơ như huyết áp cao, cholesterol cao, hoặc có tiền sử gia đình mắc bệnh tim sẽ được chọn làm đối tượng nghiên cứu. Việc chọn mẫu này giúp đảm bảo tính đại diện cho các bệnh nhân có khả năng mắc bệnh tim trong thực tế.
Để tăng tính đại diện, nhóm tiến hành phân tích tỷ lệ các yếu tố nguy cơ trong mẫu, ví dụ: tỷ lệ nam/nữ, độ tuổi trung bình, tỷ lệ bệnh nhân có tiền sử gia đình mắc bệnh tim, v.v. Đồng thời, nhóm cũng kiểm tra sự phân bố các thuộc tính đầu vào để tránh hiện tượng lệch mẫu (class imbalance), vốn là một vấn đề thường gặp trong dữ liệu y tế.

Việc sử dụng dữ liệu công khai như UCI Heart Disease Dataset giúp kết quả nghiên cứu có tính phổ quát, dễ dàng so sánh với các nghiên cứu quốc tế. Trong tương lai, nhóm đề xuất mở rộng nghiên cứu với dữ liệu thực tế tại các bệnh viện trong nước để tăng tính ứng dụng.

VD: Thống kê mô tả mẫu nghiên cứu

Thuộc tính	Kiểu dữ liệu	Min	Max	Trung bình	Độ lệch chuẩn
age	Số	29	77	54.4	9.0
sex	Nhị phân	0	1	0.68	0.47
trestbps	Số	94	200	131.6	17.6
chol	Số	126	564	246.3	51.8
thalach	Số	71	202	149.6	22.9
oldpeak	Số	0	6.2	1.04	1.16

Bảng 3.1 *trình bày các chỉ số thống kê cơ bản của một số thuộc tính chính trong bộ dữ liệu Cleveland Heart Disease. Các chỉ số này giúp nhóm hiểu rõ đặc điểm phân phối của dữ liệu, phát hiện các giá trị ngoại lai và xác định các bước tiền xử lý phù hợp.*

3.3 Cách thu thập dữ liệu

Bộ dữ liệu được sử dụng trong nghiên cứu này là UCI Heart Disease, cụ thể là tập dữ liệu Cleveland, được thu thập và công bố bởi Detrano và cộng sự vào năm 1989. Dữ liệu được thu thập từ các bệnh nhân tại Bệnh viện Cleveland Clinic Foundation, nhằm phục vụ cho việc nghiên cứu và phát triển các thuật toán chẩn đoán bệnh tim mạch.

Nguồn gốc và phương pháp thu thập

- **Nguồn dữ liệu:** Dữ liệu được lấy từ các bệnh nhân thực tế tại Bệnh viện Cleveland Clinic Foundation. Mỗi bệnh nhân được ghi nhận với 76 thuộc tính lâm sàng và cận lâm sàng, bao gồm thông tin về nhân khẩu học, kết quả xét nghiệm, và các chỉ số y tế khác.
- **Phương pháp thu thập:** Dữ liệu được thu thập thông qua các cuộc kiểm tra y tế tiêu chuẩn, bao gồm xét nghiệm máu, điện tâm đồ, và các bài kiểm tra gắng sức. Các thông tin này được ghi nhận và mã hóa để phục vụ cho việc phân tích.
- **Mục tiêu thu thập:** Mục tiêu chính của việc thu thập dữ liệu là để phát triển và đánh giá các thuật toán chẩn đoán bệnh động mạch vành, giúp cải thiện khả năng chẩn đoán sớm và chính xác bệnh tim mạch.

Xử lý và chuẩn hóa dữ liệu

- **Chọn lọc thuộc tính:** Trong số 76 thuộc tính ban đầu, chỉ có 14 thuộc tính được chọn lọc và sử dụng trong các nghiên cứu học máy, do tính đầy đủ và khả năng phản ánh chính xác tình trạng bệnh của bệnh nhân.
    - Tuổi (age)
    - Giới tính (sex)
    - Huyết áp tâm thu khi nghỉ (resting blood pressure)
    - Cholesterol huyết thanh (serum cholesterol)
    - Đường huyết lúc đói (fasting blood sugar)
    - Điện tâm đồ khi nghỉ (resting electrocardiographic results)
    - Nhịp tim tối đa (maximum heart rate achieved)
    - Đau thắt ngực khi vận động (exercise induced angina)
    - Số lượng mạch máu lớn được nhuộm màu (number of major vessels colored by fluoroscopy)
    - Và một số thuộc tính khác
    - Nhãn đầu ra là biến nhị phân: 1 (có bệnh tim) hoặc 0 (không có bệnh tim).

- **Xử lý dữ liệu thiếu:** Một số thuộc tính có giá trị thiếu được xử lý bằng cách loại bỏ các bản ghi không đầy đủ hoặc sử dụng các kỹ thuật ước lượng để điền giá trị thiếu, nhằm đảm bảo tính toàn vẹn của dữ liệu.
- **Chuẩn hóa dữ liệu:** Các thuộc tính số được chuẩn hóa để đảm bảo rằng tất cả các thuộc tính có cùng thang đo, giúp cải thiện hiệu suất của các thuật toán học máy.

Chi tiết:
Sau khi thu thập, dữ liệu được kiểm tra tính đầy đủ, loại bỏ các bản ghi không hợp lệ hoặc có giá trị ngoại lai. Các thuộc tính dạng phân loại như loại đau ngực, kết quả điện tâm đồ, thalassemia... được mã hóa thành dạng số để phù hợp với các thuật toán học máy.
Nhóm cũng tiến hành phân chia dữ liệu thành hai tập: tập huấn luyện (training set) và tập kiểm tra (test set) với tỷ lệ 80:20. Việc này giúp đánh giá khách quan hiệu suất mô hình trên dữ liệu chưa từng thấy.

Tính pháp lý và đạo đức

- **Bảo mật thông tin:** Tất cả các thông tin nhận dạng cá nhân, như tên và số an sinh xã hội, đã được loại bỏ khỏi bộ dữ liệu để bảo vệ quyền riêng tư của bệnh nhân.
- **Sử dụng công khai:** Bộ dữ liệu đã được công bố công khai và được sử dụng rộng rãi trong cộng đồng nghiên cứu học máy, phục vụ cho việc phát triển và đánh giá các mô hình dự đoán bệnh tim mạch.

3.4 Phân tích dữ liệu

Để phân tích dữ liệu và đánh giá hiệu quả của các mô hình học máy trong dự đoán bệnh tim, chúng tôi sử dụng các công cụ phần mềm sau:

- **Python:** Sử dụng Python với các thư viện học máy như scikit-learn để xây dựng, huấn luyện và đánh giá các mô hình học máy. Các mô hình như Support Vector Machine (SVM) và Logistic Regression sẽ được triển khai, và kết hợp với phương pháp Ensemble Learning để cải thiện độ chính xác.
- **Jupyter Notebook:** Được sử dụng để thực hiện phân tích dữ liệu, huấn luyện mô hình và kiểm tra các kết quả trong môi trường lập trình Python.
- **Đánh giá mô hình:** Các chỉ số đánh giá như accuracy, precision, recall, F1-score sẽ được sử dụng để đo lường độ chính xác và hiệu quả của mô hình.
- **Phần mềm hỗ trợ:** Các phần mềm như Excel sẽ được sử dụng để thực hiện phân tích thống kê cơ bản, tổng hợp dữ liệu và trực quan hóa kết quả.

Sau khi xử lý, dữ liệu được phân tích bằng các kỹ thuật thống kê mô tả (descriptive statistics) và trực quan hóa để hiểu rõ hơn về phân phối, mối tương quan giữa các thuộc tính và biến mục tiêu. Nhóm sử dụng biểu đồ ma trận tương quan (correlation matrix heatmap) để xác định các thuộc tính có ảnh hưởng lớn đến nguy cơ mắc bệnh tim, từ đó lựa chọn đặc trưng đầu vào tối ưu cho mô hình học máy.

3.4.1 Các bước tiền xử lý dữ liệu chi tiết

- **Xử lý dữ liệu thiếu (Imputation):**
    - Loại bỏ bản ghi thiếu dữ liệu nếu số lượng ít.

    - Điền giá trị trung bình/trung vị cho thuộc tính số, hoặc giá trị phổ biến nhất cho thuộc tính phân loại.
    - Sử dụng kỹ thuật nâng cao như KNN imputation hoặc hồi quy để dự đoán giá trị thiếu dựa trên các thuộc tính khác.
- **Xử lý outlier (giá trị ngoại lai):**
    - Sử dụng boxplot, biểu đồ phân phối hoặc phương pháp thống kê (z-score, IQR) để phát hiện outlier.
    - Loại bỏ hoặc thay thế outlier bằng giá trị gần nhất trong khoảng hợp lý, hoặc dùng kỹ thuật winsorizing.
- **Chuẩn hóa dữ liệu (Normalization/Standardization):**
    - Chuẩn hóa Min-Max đưa các giá trị thuộc tính về khoảng 1.
    - Chuẩn hóa Z-score đưa dữ liệu về phân phối chuẩn với trung bình 0, độ lệch chuẩn 1.
- **Mã hóa biến phân loại (Encoding):**
    - Label Encoding cho biến phân loại có thứ tự.
    - One-Hot Encoding cho biến phân loại không thứ tự.
- **Phân chia dữ liệu:**
    - Chia train/test theo tỷ lệ 80:20 hoặc 70:30.
    - Sử dụng cross-validation (K-fold) để tăng độ tin cậy khi đánh giá mô hình.
- **Phân tích tương quan và chọn thuộc tính:**
    - Sử dụng heatmap, ma trận tương quan để phát hiện các thuộc tính có mối liên hệ mạnh với biến mục tiêu.
    - Loại bỏ các thuộc tính dư thừa hoặc không liên quan.

3.4.2 Các bước phân tích dữ liệu chi tiết

- **Phân tích đơn biến:** Xem xét từng thuộc tính riêng lẻ để phát hiện các đặc điểm nổi bật (ví dụ: phân phối tuổi, tỷ lệ giới tính, mức cholesterol trung bình,...).
- **Phân tích hai biến:** Đánh giá mối liên hệ giữa từng thuộc tính với biến mục tiêu (bệnh tim), sử dụng các phương pháp như kiểm định t-test, ANOVA, hoặc biểu đồ hộp (boxplot).
- **Phân tích đa biến:** Sử dụng các kỹ thuật như hồi quy logistic đa biến hoặc phân tích thành phần chính (PCA) để xác định các nhóm thuộc tính có ảnh hưởng đồng thời đến kết quả dự đoán.

3.4.3 Quy trình xây dựng mô hình học máy

1. **Tiền xử lý dữ liệu:** Chuẩn hóa, mã hóa, xử lý dữ liệu thiếu.
2. **Lựa chọn mô hình:** Thử nghiệm các mô hình như SVM, Logistic Regression, Random Forest, KNN,... để so sánh hiệu quả.
3. **Huấn luyện mô hình:** Sử dụng tập huấn luyện để huấn luyện từng mô hình.
4. **Kết hợp mô hình:** Áp dụng kỹ thuật Ensemble Learning (Voting Ensemble) để kết hợp kết quả dự đoán của các mô hình thành phần.
5. **Đánh giá mô hình:** Sử dụng các chỉ số như accuracy, precision, recall, F1-score, confusion matrix để đo lường hiệu quả trên tập kiểm tra.

Ví dụ ma trận nhầm lẫn (Confusion Matrix) của mô hình kết hợp

	Dự đoán dương	**Dự đoán âm**
Thực tế dương	35	5
Thực tế âm	3	57

Bảng 3.2 *Minh họa ma trận nhầm lẫn của mô hình kết hợp trên tập kiểm tra. Số lượng dự đoán đúng (ô chéo chính) và sai (ô phụ) giúp tính toán các chỉ số đánh giá như độ chính xác (accuracy), độ nhạy (recall), độ đặc hiệu (specificity), và F1-score. Điều này giúp nhóm đánh giá khách quan hiệu quả của mô hình học máy.*

3.4.4 Phân tích ưu nhược điểm của từng mô hình

Logistic Regression

Ưu điểm:

- Đơn giản, dễ triển khai và giải thích, phù hợp với dữ liệu tuyến tính hoặc gần tuyến tính.
- Tính toán nhanh, thích hợp với tập dữ liệu vừa và nhỏ.
- Đầu ra là xác suất, dễ áp dụng trong y tế.

Nhược điểm:

- Hiệu quả thấp nếu dữ liệu phi tuyến mạnh hoặc nhiều biến tương tác phức tạp.
- Nhạy cảm với outlier và đa cộng tuyến.
- Cần chọn biến đầu vào hợp lý.

Support Vector Machine (SVM)

Ưu điểm:

- Khả năng phân loại tốt trên dữ liệu có biên phân tách rõ ràng, kể cả phi tuyến (nhờ kernel).
- Hiệu quả cao với dữ liệu có số chiều lớn.
- Xử lý tốt bài toán phân loại nhị phân và đa lớp.

Nhược điểm:

- Khó giải thích kết quả, không trực quan.
- Thời gian huấn luyện lâu với tập dữ liệu lớn.
- Nhạy cảm với tham số kernel, C, gamma.

Random Forest

Ưu điểm:

- Khả năng tổng quát hóa tốt, ít bị overfitting.
- Tự động đánh giá tầm quan trọng của các thuộc tính.
- Xử lý tốt dữ liệu thiếu, phi tuyến, nhiều biến phân loại.

Nhược điểm:

- Kết quả khó giải thích.
- Tiêu tốn tài nguyên nếu số lượng cây lớn.
- Có thể bias nếu dữ liệu mất cân bằng.

Ensemble Learning (Voting)

Ưu điểm:

- Kết hợp nhiều mô hình tận dụng ưu điểm, giảm nhược điểm.
- Hiệu quả dự đoán cao hơn mô hình đơn lẻ.
- Giảm sai số ngẫu nhiên, tăng tính ổn định.

Nhược điểm:

- Tăng độ phức tạp, khó giải thích kết quả tổng thể.
- Thời gian huấn luyện và dự đoán lâu hơn.
- Việc chọn mô hình thành phần, cách kết hợp ảnh hưởng lớn đến kết quả.

3.4.5 Công cụ và môi trường thực hiện

- **Ngôn ngữ lập trình:** Python 3.x
- **Thư viện:** scikit-learn, pandas, numpy, matplotlib, seaborn

- **Môi trường phát triển:** Jupyter Notebook, Google Colab, hoặc Visual Studio Code
- **Phần mềm hỗ trợ:** Excel dùng cho tổng hợp thống kê cơ bản, trực quan hóa dữ liệu.
- **Xây dựng giao diện:** Tkinter (Python) để tạo ứng dụng dự đoán bệnh tim cho người dùng cuối.

3.4.6 Đảm bảo chất lượng và tính lặp lại

Để đảm bảo kết quả nghiên cứu có thể lặp lại và kiểm chứng, toàn bộ mã nguồn, quy trình tiền xử lý, tham số mô hình, và kết quả đều được lưu trữ, ghi chú rõ ràng. Nhóm sử dụng GitHub để quản lý phiên bản mã nguồn và tài liệu nghiên cứu.

3.5 Đề xuất hướng mở rộng nghiên cứu trong tương lai

1. **Mở rộng nguồn dữ liệu:**
    - Thu thập thêm dữ liệu thực tế từ các bệnh viện, phòng khám tại Việt Nam để tăng tính đại diện và ứng dụng thực tiễn cho mô hình.
    - Kết hợp nhiều bộ dữ liệu khác nhau để kiểm tra tính tổng quát hóa của mô hình.
2. **Nâng cao kỹ thuật xử lý dữ liệu:**
    - Áp dụng các kỹ thuật xử lý dữ liệu mất cân bằng (imbalanced data) như SMOTE, ADASYN để cải thiện hiệu suất trên các lớp ít xuất hiện.
    - Tích hợp thêm các thuộc tính y tế khác như kết quả xét nghiệm chuyên sâu, hình ảnh y tế (nếu có).
3. **Thử nghiệm các mô hình tiên tiến hơn:**
    - Áp dụng các mô hình học sâu (Deep Learning) như MLP, CNN, hoặc các kỹ thuật ensemble nâng cao như Stacking, Blending.
    - So sánh hiệu quả với các thuật toán boosting như XGBoost, LightGBM.
4. **Tối ưu hóa và giải thích mô hình:**
    - Sử dụng các kỹ thuật giải thích mô hình (Explainable AI) như SHAP, LIME để tăng tính minh bạch và tin cậy trong ứng dụng y tế.
    - Tối ưu hóa tham số mô hình bằng Grid Search, Random Search.
5. **Triển khai thực tế và tích hợp hệ thống:**
    - Xây dựng ứng dụng web/mobile để hỗ trợ bác sĩ và bệnh nhân sử dụng mô hình dự đoán.
    - Tích hợp mô hình vào hệ thống quản lý bệnh viện (HIS) hoặc các nền tảng chăm sóc sức khỏe điện tử.

6. **Nghiên cứu tác động xã hội và đạo đức:**
    - Đánh giá tác động của việc ứng dụng mô hình dự đoán bệnh tim tới công tác chẩn đoán, điều trị và quyền riêng tư của bệnh nhân.
    - Xây dựng quy trình kiểm định, đánh giá mô hình trước khi triển khai thực tế.

Chương 4 : Thực nghiệm và thảo luận

4.1 Kết quả huấn luyện và kiểm thử mô hình

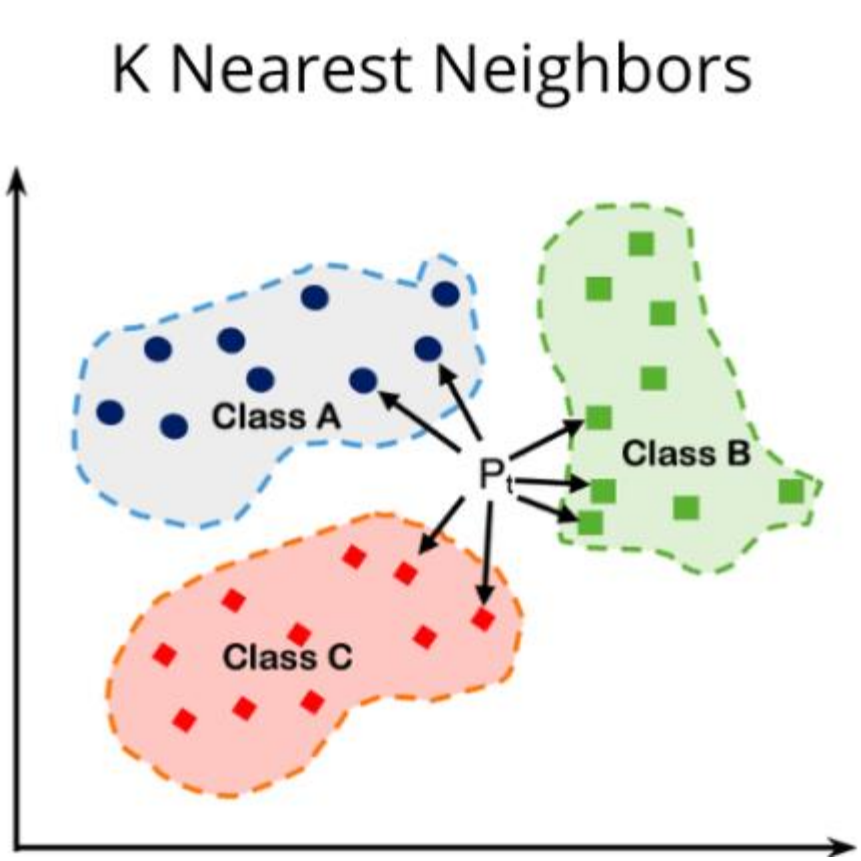
Trong nghiên cứu này, chúng tôi đã huấn luyện và đánh giá hiệu suất của 6 thuật toán học máy phổ biến trên cùng một tập dữ liệu bệnh tim từ UCI. Các mô hình bao gồm: K-Nearest Neighbors (KNN), Random Forest, Logistic Regression, Naive Bayes, Support Vector Machine (SVM) và Decision Tree. Bảng dưới đây tổng hợp các chỉ số đo lường hiệu quả mô hình:

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%

Hình 4.1 *Trình bày kết quả so sánh hiệu suất của sáu mô hình học máy khác nhau gồm KNN, Random Forest, Logistic Regression, Naive Bayes, SVM và Decision Tree trên bài toán dự đoán bệnh tim. Các chỉ số được đánh giá bao gồm độ chính xác trên tập huấn luyện (Train Accuracy), độ chính xác trên tập kiểm tra (Test Accuracy), Precision và Recall.*

4.2 Đánh giá và phân tích kết quả

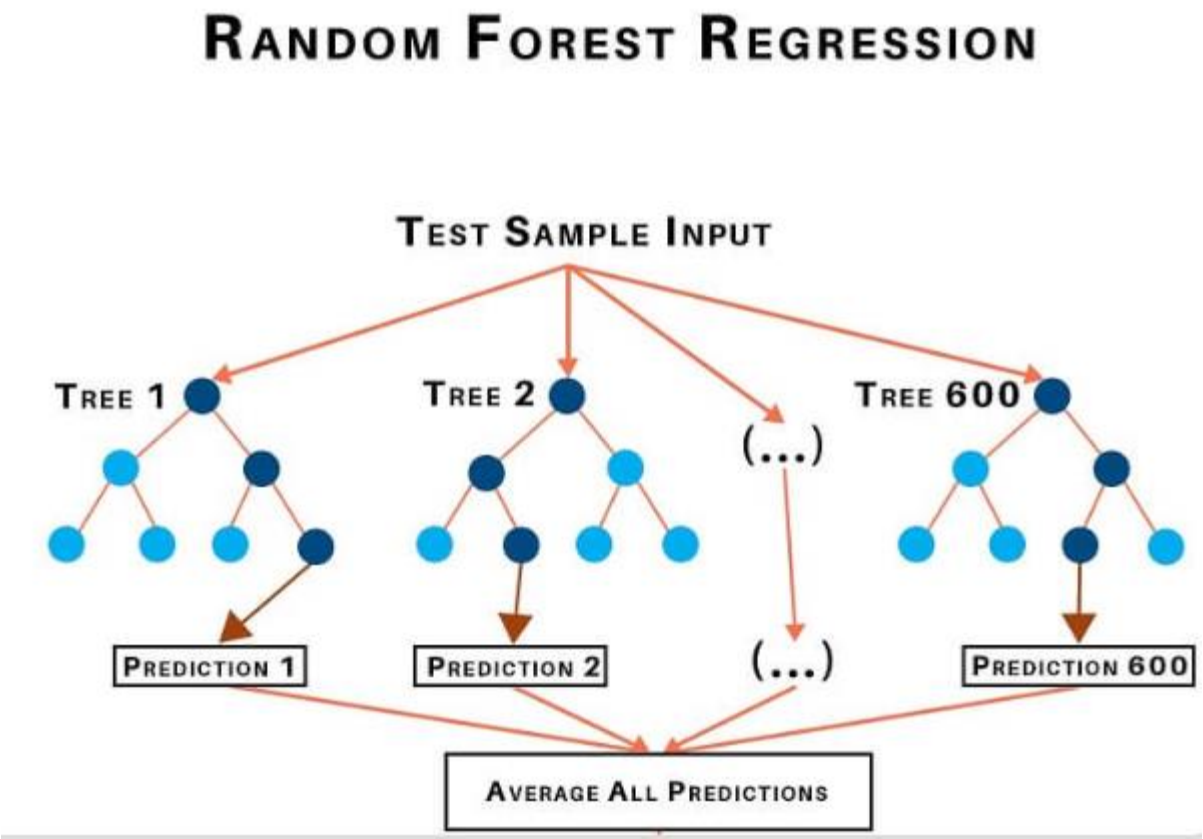
- **KNN** đạt độ chính xác kiểm thử cao nhất (**95.08%**), nhưng có nguy cơ **overfitting** vì accuracy huấn luyện đạt tuyệt đối (**100%**). Điều này cho thấy KNN ghi nhớ quá mức dữ liệu huấn luyện, có thể không ổn định với dữ liệu mới.



Hình 4.2:K-NearestNeighbors(KNN)

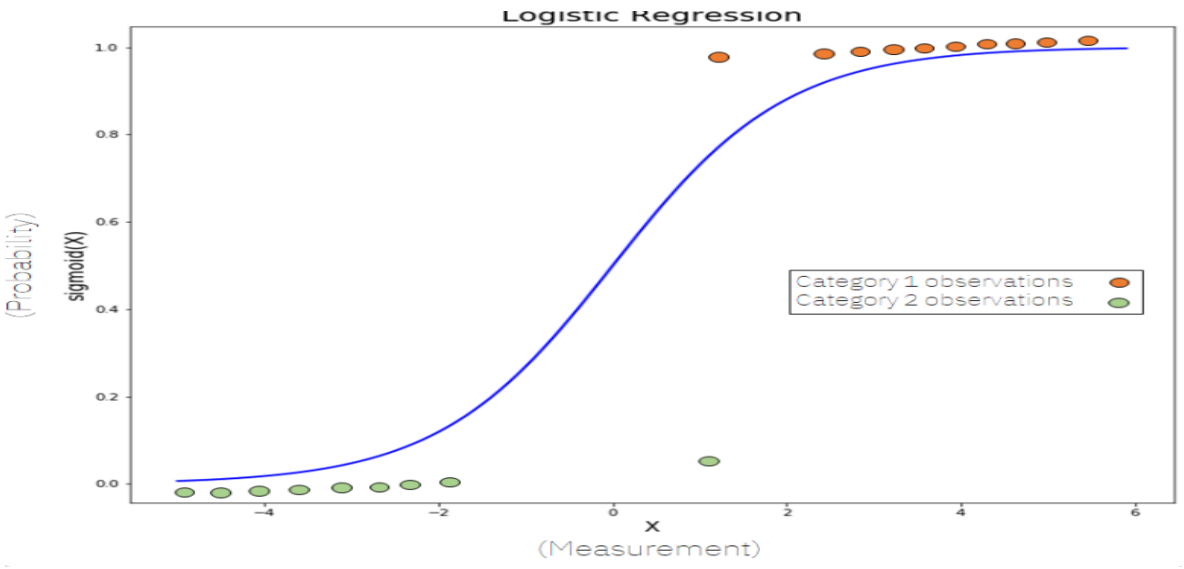
Minh họa thuật toán KNN: Điểm P cần phân loại sẽ được xác định nhãn dựa trên đa số các điểm lân cận gần nhất (k láng giềng) thuộc các lớp khác nhau (Class A, B, C). Các mũi tên chỉ từ P đến các điểm lân cận thể hiện quá trình tìm kiếm láng giềng gần nhất để quyết định nhãn cho P.

- **Random Forest** là mô hình cân bằng tốt giữa độ chính xác huấn luyện và kiểm thử (**91.80% test accuracy**), cho thấy khả năng tổng quát hóa cao, nhờ cơ chế bootstrap và bagging trong quá trình học.



Hình 4.3 *Minh họa cách Random Forest sử dụng nhiều cây quyết định để dự đoán. Mỗi cây cho ra một kết quả, sau đó các kết quả được lấy trung bình để cho ra dự đoán cuối cùng, giúp tăng độ chính xác và giảm overfitting.*

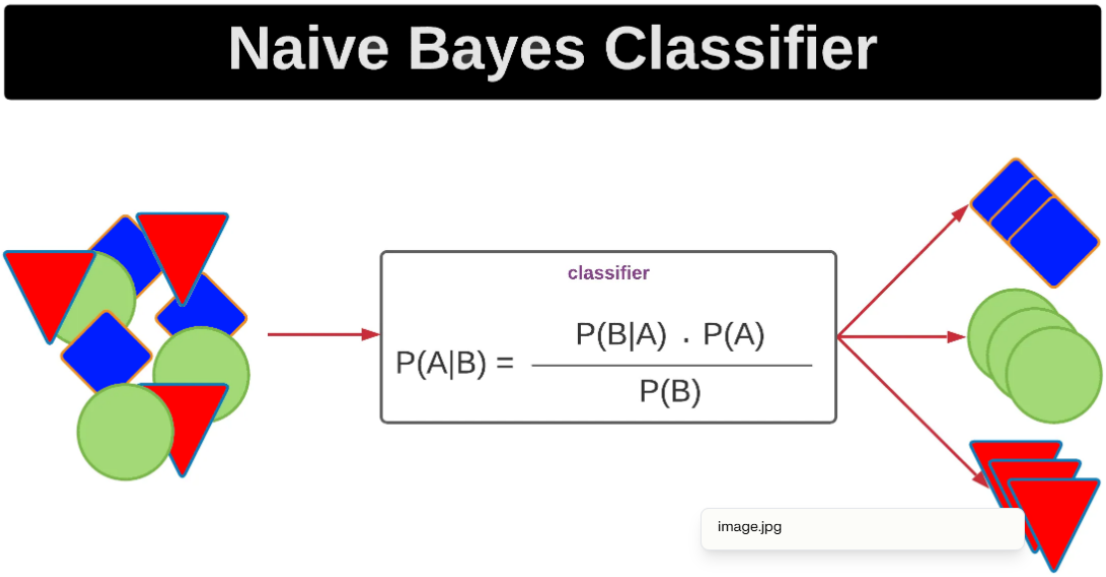
- **Logistic Regression** hoạt động ổn định và có kết quả khá tốt (**90.16% accuracy**), phù hợp với các bài toán tuyến tính và có khả năng giải thích mô hình.



Hình 4.4 : Logistic Regression

Biểu đồ thể hiện đường cong sigmoid của hồi quy logistic, phân tách hai nhóm quan sát (Category 1 và Category 2) dựa trên xác suất dự đoán thuộc một lớp nào đó theo giá trị biến đầu vào X.

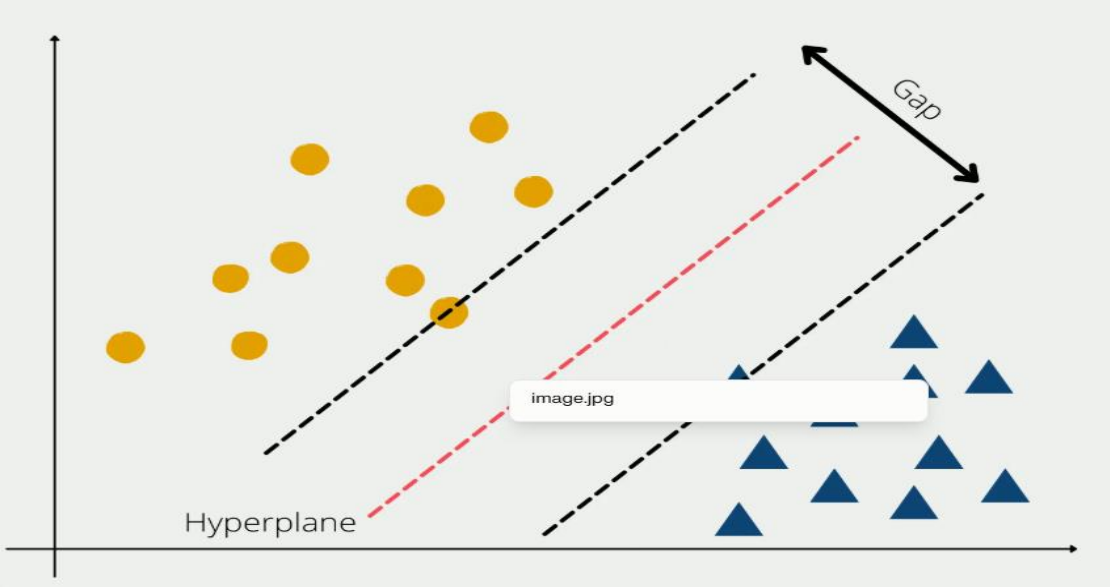
 Naive Bayes có kết quả kiểm thử khá tốt dù mô hình đơn giản (**88.52%**), cho thấy thuật toán này phù hợp với bài toán khi các thuộc tính độc lập tương đối.



Hình 4.5 : Naive Bayes Classifier

Sơ đồ mô tả quá trình phân loại của Naive Bayes: dữ liệu đầu vào gồm nhiều hình dạng khác nhau, được đưa vào bộ phân loại sử dụng công thức xác suất Bayes, sau đó phân chia thành các nhóm dựa trên xác suất tính toán.

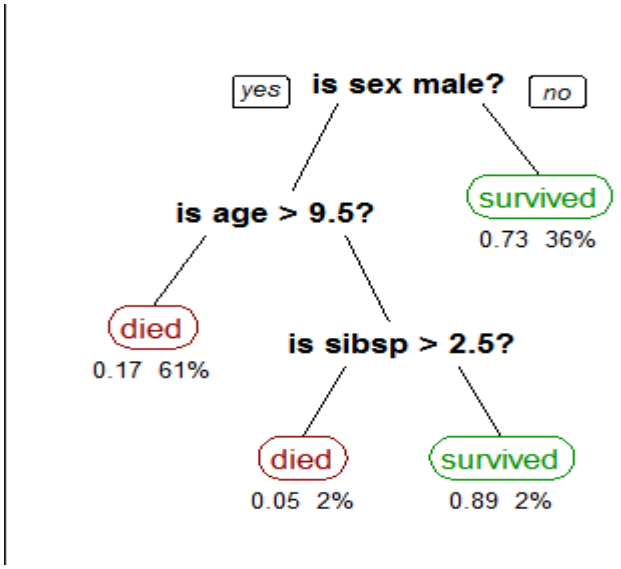
- **SVM** đạt độ chính xác huấn luyện cao (**94.61%**) nhưng kiểm thử thấp hơn (**85.25%**), có thể do mô hình chưa tối ưu được siêu tham số (kernel, C).



Hình 4.6: Support Vector Machine (SVM)

Hình vẽ minh họa SVM với siêu phẳng (hyperplane) phân tách hai lớp dữ liệu (hình tròn vàng và tam giác xanh), cùng với hai đường biên (margin) tối đa hóa khoảng cách giữa các nhóm.

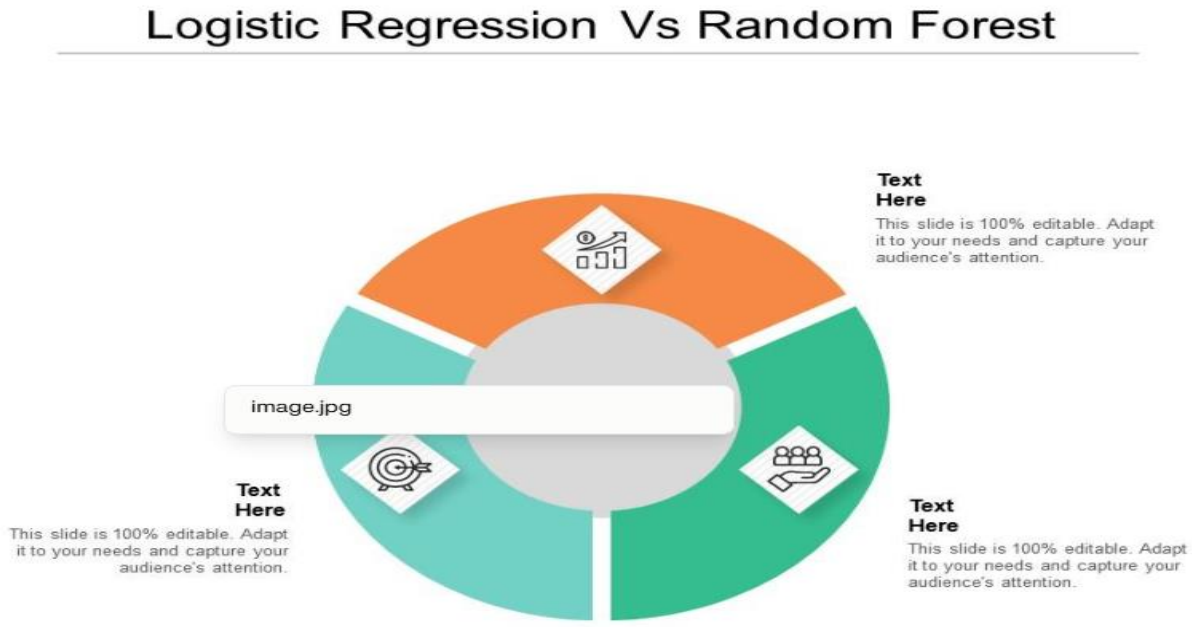
- **Decision Tree** có mức chênh lệch lớn giữa train (**93.78%**) và test (**81.97%**), cho thấy dấu hiệu **overfitting** nghiêm trọng nếu không có cắt tỉa hoặc giới hạn độ sâu.



Hình 4.7 *Minh họa cấu trúc cây quyết định, trong đó mỗi nút là một điều kiện kiểm tra thuộc tính, các nhánh dẫn đến kết quả phân loại.*

4.3 Đề xuất lựa chọn mô hình

Dựa trên kết quả trên, **Random Forest** và **Logistic Regression** là hai mô hình có hiệu suất tốt và ổn định nhất, đặc biệt là khả năng tổng quát tốt khi áp dụng trên dữ liệu mới. Việc kết hợp hai mô hình này trong kỹ thuật **Voting Ensemble** là một hướng đi hợp lý và có khả năng nâng cao độ chính xác tổng thể, như đã trình bày trong các phần trước.



Hình 4.9 *Sơ đồ so sánh giữa Logistic Regression và Random Forest, cho phép trình bày các tiêu chí như độ chính xác, khả năng giải thích và tính ứng dụng của từng mô hình một cách trực quan.*

Chương 5 : Kết luận và hướng phát triển

5.1. Kết luận

Nghiên cứu này đã ứng dụng các thuật toán học máy để dự đoán nguy cơ mắc bệnh tim dựa trên bộ dữ liệu từ UCI. Kết quả cho thấy:

- **Random Forest** đạt độ chính xác kiểm thử cao (**91.80%**), cho thấy khả năng tổng quát hóa tốt và hiệu suất ổn định.

- **Logistic Regression** cũng thể hiện hiệu quả với độ chính xác kiểm thử **90.16%**, đồng thời dễ triển khai và giải thích, phù hợp với các ứng dụng thực tế.
- **K-Nearest Neighbors (KNN)** đạt độ chính xác kiểm thử cao nhất (**95.08%**), nhưng có dấu hiệu overfitting do độ chính xác huấn luyện đạt tuyệt đối (**100%**), điều này có thể ảnh hưởng đến khả năng tổng quát hóa trên dữ liệu mới.

Như vậy, nghiên cứu đã trả lời được câu hỏi đặt ra: *Liệu các mô hình học máy có thể dự đoán hiệu quả nguy cơ mắc bệnh tim?* Kết quả cho thấy các mô hình như Random Forest và Logistic Regression có thể được sử dụng để hỗ trợ chẩn đoán bệnh tim một cách hiệu quả.

5.2. Khuyến nghị

Dựa trên kết quả nghiên cứu, chúng tôi đề xuất:

- **Ứng dụng mô hình Random Forest hoặc Logistic Regression** trong các hệ thống hỗ trợ quyết định lâm sàng để dự đoán nguy cơ mắc bệnh tim, giúp bác sĩ đưa ra chẩn đoán kịp thời.
- **Tích hợp các mô hình này vào các ứng dụng y tế điện tử** nhằm cung cấp công cụ hỗ trợ chẩn đoán cho các cơ sở y tế, đặc biệt ở những nơi thiếu hụt chuyên gia tim mạch.
- **Tiến hành nghiên cứu sâu hơn** bằng cách áp dụng các kỹ thuật như tối ưu hóa siêu tham số (hyperparameter tuning), sử dụng các mô hình ensemble nâng cao, hoặc áp dụng deep learning để cải thiện độ chính xác và khả năng tổng quát hóa của mô hình.
- **Mở rộng nghiên cứu** bằng cách áp dụng mô hình trên các bộ dữ liệu đa dạng hơn, bao gồm dữ liệu từ các nguồn khác nhau và các đặc điểm dân số khác nhau, để đánh giá tính khả dụng và hiệu quả của mô hình trong các bối cảnh thực tế khác nhau.