**Tóm Tắt Bài Báo**

**Heart disease prediction using machine learning algorithms**

**Harshit Jindal, Sarthak Agrawal, Rishabh Khera, Rachna Jain and Preeti Nagrath**

Họ và tên: Phạm Tấn Khương

MSSV: 3122410191

Tóm tắt (Abstract) của bài báo nêu rằng tỷ lệ mắc bệnh tim đang gia tăng nhanh chóng, và việc dự đoán sớm các bệnh này là rất quan trọng và đáng lo ngại. Chẩn đoán bệnh tim là một nhiệm vụ khó khăn, đòi hỏi sự chính xác và hiệu quả. Nghiên cứu này tập trung vào việc xác định bệnh nhân nào có nhiều khả năng mắc bệnh tim dựa trên các thuộc tính y tế khác nhau . Các tác giả đã xây dựng một hệ thống dự đoán bệnh tim để dự đoán khả năng bệnh nhân được chẩn đoán mắc bệnh tim hay không bằng cách sử dụng tiền sử bệnh của bệnh nhân. Họ đã sử dụng các thuật toán học máy khác nhau như **logistic regression** và **KNN** để dự đoán và phân loại bệnh nhân mắc bệnh tim. Một phương pháp hữu ích đã được sử dụng để điều chỉnh cách mô hình có thể được sử dụng để cải thiện độ chính xác của việc dự đoán nhồi máu cơ tim ở bất kỳ cá nhân nào. Mô hình được đề xuất cho thấy sức mạnh khá thỏa mãn và có khả năng dự đoán bằng chứng mắc bệnh tim ở một cá nhân cụ thể bằng cách sử dụng KNN và Logistic Regression, cho thấy độ chính xác tốt hơn so với các bộ phân loại được sử dụng trước đây như naive bayes. Mô hình này giúp giảm bớt áp lực trong việc tìm ra xác suất để bộ phân loại xác định chính xác bệnh tim. Hệ thống dự đoán bệnh tim được đưa ra giúp nâng cao chăm sóc y tế và giảm chi phí. Dự án này cung cấp kiến thức quan trọng có thể giúp dự đoán bệnh nhân mắc bệnh tim và được triển khai ở định dạng .pynb.

Trang thứ hai của bài báo tiếp tục giới thiệu về tầm quan trọng của việc dự đoán bệnh tim. Các bệnh tim mạch rất phổ biến hiện nay và mô tả một loạt các tình trạng có thể ảnh hưởng đến tim1 . Tổ chức Y tế Thế giới (WHO) ước tính có **17,9 triệu ca tử vong trên toàn cầu** là do các bệnh tim mạch (CVDs). Đây là nguyên nhân hàng đầu gây tử vong ở người lớn. Dự án này có thể giúp dự đoán những người có khả năng được chẩn đoán mắc bệnh tim dựa trên tiền sử bệnh của họ.... Nó giúp nhận biết những người có bất kỳ triệu chứng nào của bệnh tim như đau ngực hoặc huyết áp cao và có thể giúp chẩn đoán bệnh với ít xét nghiệm y tế hơn và điều trị hiệu quả hơn, để họ có thể được chữa trị phù hợp .

Dự án này tập trung chủ yếu vào **ba kỹ thuật khai thác dữ liệu**: (1) **Logistic regression**, (2) **KNN**, và (3) **Random Forest Classifier** . Độ chính xác của dự án là **87,5%**, cao hơn so với các hệ thống trước đây chỉ sử dụng một kỹ thuật khai thác dữ liệu2 . Việc sử dụng nhiều kỹ thuật khai thác dữ liệu đã làm tăng độ chính xác và hiệu quả của Hệ thống dự đoán bệnh tim (HDPS) . **Logistic regression** thuộc loại học có giám sát và chỉ sử dụng các giá trị rời rạc .

Trang tiếp theo của bài báo, bắt đầu từ trang 4, bao gồm các phần **1. Giới thiệu** và tiếp tục với **2. Các công việc liên quan**1 ....

Phần **1. Giới thiệu** định nghĩa **Machine Learning** là một cách thao tác và trích xuất thông tin ẩn, đã biết/chưa biết trước đó và hữu ích tiềm năng về dữ liệu1 . Machine Learning là một lĩnh vực rất rộng lớn và đa dạng, phạm vi và ứng dụng của nó ngày càng tăng. Nó kết hợp nhiều bộ phân loại thuộc Học có giám sát, Học không giám sát và Học tập kết hợp để dự đoán và tìm độ chính xác của tập dữ liệu đã cho. Kiến thức này có thể được sử dụng trong dự án Hệ thống dự đoán bệnh tim (HDPS)1 .

Phần **2. Các công việc liên quan** đề cập đến một lượng đáng kể các công việc liên quan đến chẩn đoán bệnh tim mạch bằng cách sử dụng các thuật toán Học máy đã thúc đẩy nghiên cứu này . Phần này bao gồm một tổng quan ngắn gọn về các tài liệu nghiên cứu trước đó. Một hệ thống dự đoán bệnh tim mạch hiệu quả đã được thực hiện bằng cách sử dụng nhiều thuật toán khác nhau, bao gồm **Logistic Regression, KNN, Random Forest Classifier**, v.v. . Kết quả cho thấy mỗi thuật toán đều có những điểm mạnh riêng để đạt được các mục tiêu đã xác định .

Mô hình kết hợp IHDPS có khả năng tính toán ranh giới quyết định bằng cách sử dụng mô hình học máy và học sâu trước đây và mới. Nó tạo điều kiện thuận lợi cho các yếu tố/kiến thức quan trọng và cơ bản nhất, chẳng hạn như tiền sử gia đình liên quan đến bất kỳ bệnh tim nào. Tuy nhiên, độ chính xác thu được trong mô hình IHDPS thấp hơn nhiều so với các mô hình mới hơn như phát hiện bệnh động mạch vành bằng mạng nơ-ron nhân tạo và các thuật toán học máy và học sâu khác . Các yếu tố rủi ro của bệnh động mạch vành hoặc xơ vữa động mạch đã được McPherson et al. xác định bằng cách sử dụng thuật toán triển khai sẵn có, sử dụng một số kỹ thuật của Mạng nơ-ron và chỉ có thể dự đoán chính xác liệu bệnh nhân thử nghiệm có mắc bệnh hay không .

Phần **2. Các công việc liên quan** tiếp tục đề cập đến việc chẩn đoán và dự đoán Bệnh tim và Huyết áp cùng với các thuộc tính khác bằng sự hỗ trợ của mạng nơ-ron đã được giới thiệu bởi R. Subramanian et al. . Một mạng nơ-ron sâu đã được xây dựng, kết hợp các thuộc tính liên quan đến bệnh, có khả năng tạo ra kết quả được thực hiện bởi perceptron đầu ra và bao gồm gần 120 lớp ẩn, đây là kỹ thuật cơ bản và phù hợp nhất để đảm bảo kết quả chính xác về khả năng mắc bệnh tim nếu mô hình được sử dụng cho Tập dữ liệu thử nghiệm . Mạng giám sát đã được khuyến nghị để chẩn đoán bệnh tim. Khi mô hình được bác sĩ kiểm tra bằng dữ liệu lạ, mô hình đã sử dụng và huấn luyện từ dữ liệu đã học trước đó và dự đoán kết quả, từ đó tính toán độ chính xác của mô hình đã cho.

Trang tiếp theo của bài báo, trang 5, tiếp tục phần **2. Các công việc liên quan** và bắt đầu phần **3. Nguồn dữ liệu**.

Trong phần **2. Các công việc liên quan**, các tác giả đề cập đến việc mạng giám sát đã được khuyến nghị để chẩn đoán bệnh tim1 . Khi mô hình được bác sĩ kiểm tra bằng một bộ dữ liệu chưa quen, mô hình đã sử dụng và huấn luyện từ dữ liệu đã học trước đó để dự đoán kết quả và tính toán độ chính xác của mô hình1 .

Phần **3. Nguồn dữ liệu** mô tả rằng một **bộ dữ liệu có tổ chức về các cá nhân đã được chọn** dựa trên tiền sử bệnh tim và các tình trạng y tế khác của họ. Bệnh tim là các tình trạng đa dạng ảnh hưởng đến tim . Theo Tổ chức Y tế Thế giới (WHO), nguyên nhân gây tử vong hàng đầu ở người trung niên là các bệnh tim mạch. Các tác giả đã sử dụng một nguồn dữ liệu bao gồm **tiền sử y tế của 304 bệnh nhân khác nhau thuộc các nhóm tuổi khác nha** . Bộ dữ liệu này cung cấp thông tin cần thiết, cụ thể là các **thuộc tính y tế như tuổi, huyết áp lúc nghỉ ngơi, mức đường huyết lúc đói, v.v.** của bệnh nhân, giúp phát hiện bệnh nhân được chẩn đoán mắc bệnh tim hay không . Bộ dữ liệu này chứa **13 thuộc tính y tế của 304 bệnh nhân**, giúp xác định xem bệnh nhân có nguy cơ mắc bệnh tim hay không và phân loại bệnh nhân thuộc nhóm nguy cơ hay không3 . Bộ dữ liệu về bệnh tim này được lấy từ kho lưu trữ **UCI repository**. Theo bộ dữ liệu này, các mẫu dẫn đến việc phát hiện bệnh nhân có khả năng mắc bệnh tim sẽ được trích xuất. Các bản ghi này được chia thành hai phần: **Huấn luyện (Training) và Kiểm tra (Testing)**. Bộ dữ liệu này chứa **303 hàng và 14 cột**, trong đó mỗi hàng tương ứng với một bản ghi duy nhất....

Trang tiếp theo của bài báo, trang 6, tiếp tục phần **3. Nguồn dữ liệu** và giới thiệu phần **4. Phương pháp luận**.

Phần **3. Nguồn dữ liệu** tiếp tục đề cập đến việc **bộ dữ liệu chứa 303 hàng và 14 cột, với mỗi hàng tương ứng với một bản ghi duy nhất**, và tất cả các thuộc tính được liệt kê trong ‘Bảng 1’.

Phần **4. Phương pháp luận** trình bày phân tích về các thuật toán học máy khác nhau được sử dụng trong bài báo này, bao gồm **K nearest neighbors (KNN), Logistic Regression và Random Forest Classifiers**, những thuật toán này có thể hữu ích cho các nhà thực hành hoặc nhà phân tích y tế để chẩn đoán chính xác bệnh tim. Nghiên cứu này bao gồm việc xem xét các tạp chí, bài báo đã xuất bản và dữ liệu về bệnh tim mạch trong thời gian gần đây. Phương pháp luận cung cấp một khuôn khổ cho mô hình được đề xuất. Phương pháp luận là một quy trình bao gồm các bước chuyển đổi dữ liệu đã cho thành các mẫu dữ liệu được nhận dạng cho người dùng. Phương pháp luận được đề xuất (Hình 1) bao gồm các bước, trong đó bước đầu tiên là **thu thập dữ liệu**, sau đó ở giai đoạn thứ hai là **trích xuất các giá trị quan trọng**, tiếp theo là giai đoạn **tiền xử lý** để khám phá dữ liệu.

**Tiền xử lý dữ liệu** bao gồm việc xử lý các giá trị bị thiếu, làm sạch dữ liệu và chuẩn hóa tùy thuộc vào các thuật toán được sử dụng. Sau khi tiền xử lý dữ liệu, bộ phân loại được sử dụng để phân loại dữ liệu đã được tiền xử lý. Các bộ phân loại được sử dụng trong mô hình đề xuất là **KNN, Logistic Regression và Random Forest Classifier**. Cuối cùng, mô hình đề xuất được thực hiện, trong đó mô hình được đánh giá dựa trên độ chính xác và hiệu suất bằng cách sử dụng các số liệu hiệu suất khác nhau. Trong mô hình này, một **Hệ thống dự đoán bệnh tim hiệu quả (EHDPS)** đã được phát triển bằng cách sử dụng các bộ phân loại khác nhau. Mô hình này sử dụng **13 thông số y tế** như đau ngực, đường huyết lúc đói, huyết áp, cholesterol, tuổi, giới tính, v.v. để dự đoán.

Trang này cũng chứa **Hình 1. Mô hình đề xuất**, minh họa các bước trong quy trình:

* Thu thập bộ dữ liệu bệnh tim
* Trích xuất các biến số quan trọng
* Tiền xử lý dữ liệu
  + Tách dữ liệu (Training Data và Testing Data)
* Huấn luyện (KNN, Logistic Regression, Random Forest Classifier)
* Phân loại (Heart Disease / Normal)

Trang tiếp theo của bài báo, trang 7, bao gồm phần **5. Kết quả & Thảo luận** và các Hình 2, 3, 4 và 5 minh họa cho các kết quả này.

Trong phần **5. Kết quả & Thảo luận**, các tác giả chỉ ra rằng mặc dù nhiều nhà nghiên cứu sử dụng các thuật toán khác nhau như SVC, Cây quyết định để phát hiện bệnh nhân mắc bệnh tim, nhưng **KNN, Random Forest Classifier và Logistic regression mang lại kết quả tốt hơn**. Các thuật toán được sử dụng trong nghiên cứu này chính xác hơn, tiết kiệm chi phí và nhanh hơn so với các thuật toán được sử dụng bởi các nhà nghiên cứu trước đây. Hơn nữa, **độ chính xác tối đa thu được bởi KNN và Logistic Regression là 88,5%**, bằng hoặc gần bằng độ chính xác thu được từ các nghiên cứu trước đây. Các tác giả kết luận rằng độ chính xác của họ được cải thiện do việc sử dụng nhiều thuộc tính y tế hơn từ bộ dữ liệu đã chọn.

Dự án này cũng chỉ ra rằng **Logistic Regression và KNN vượt trội hơn Random Forest Classifier** trong việc dự đoán bệnh nhân được chẩn đoán mắc bệnh tim. Điều này chứng minh rằng KNN và Logistic Regression tốt hơn trong việc chẩn đoán bệnh tim.

Trang này cũng bao gồm các hình sau:

* **Hình 2**: Thể hiện **Nguy cơ đau tim dựa trên độ tuổi**. Có hai chú thích: "Risk of Heart Attack" (Nguy cơ đau tim) và "No Risk of Heart Attack" (Không có nguy cơ đau tim).
* **Hình 3**: Thể hiện **Nguy cơ đau tim dựa trên Huyết áp lúc nghỉ ngơi**. Tương tự, có hai chú thích: "Risk of Heart Attack" và "No Risk of Heart Attack".
* **Hình 4**: Thể hiện số lượng bệnh nhân **có hoặc không mắc bệnh tim dựa trên Giới tính**. Các chú thích là "Female Popularity" (Số lượng nữ giới) và "Male Popularity" (Số lượng nam giới).
* **Hình 5**: Thể hiện số lượng bệnh nhân **có hoặc không mắc bệnh tim dựa trên loại Đau ngực**. Các chú thích bao gồm "Asymptomatic Popularity" (Số lượng không có triệu chứng), "Non-Anginal Pain Popularity" (Số lượng đau thắt ngực không điển hình), "Atypical Angina Popularity" (Số lượng đau thắt ngực không ổn định) và "Typical Angina Popularity" (Số lượng đau thắt ngực điển hình).

Trang tiếp theo của bài báo, trang 8, bao gồm phần **6. Kết luận** và **Hình 6**.

Trong phần **6. Kết luận**, các tác giả tóm tắt rằng một mô hình phát hiện bệnh tim mạch đã được phát triển bằng cách sử dụng ba kỹ thuật mô hình hóa phân loại ML. Dự án này dự đoán những người mắc bệnh tim mạch bằng cách trích xuất tiền sử bệnh của bệnh nhân dẫn đến bệnh tim gây tử vong từ một bộ dữ liệu bao gồm tiền sử bệnh của bệnh nhân như đau ngực, lượng đường trong máu, huyết áp, v.v.. Hệ thống phát hiện bệnh tim này hỗ trợ bệnh nhân dựa trên thông tin lâm sàng về việc họ đã được chẩn đoán mắc bệnh tim trước đó. Các thuật toán được sử dụng để xây dựng mô hình này là **Logistic regression, Random Forest Classifier và KNN**. **Độ chính xác của mô hình là 87,5%**. Việc sử dụng nhiều dữ liệu huấn luyện hơn đảm bảo cơ hội cao hơn để mô hình dự đoán chính xác liệu người được kiểm tra có mắc bệnh tim hay không. Bằng cách sử dụng các kỹ thuật hỗ trợ máy tính này, có thể dự đoán bệnh nhân nhanh hơn và tốt hơn, đồng thời giảm đáng kể chi phí. Có nhiều cơ sở dữ liệu y tế có thể được khai thác vì các kỹ thuật Học máy này tốt hơn và có thể dự đoán tốt hơn con người, điều này giúp ích cho cả bệnh nhân và bác sĩ. Do đó, kết luận lại rằng dự án này giúp dự đoán những bệnh nhân được chẩn đoán mắc bệnh tim bằng cách làm sạch bộ dữ liệu và áp dụng hồi quy logistic và KNN để đạt được độ chính xác trung bình là **87,5%** trên mô hình, cao hơn so với các mô hình trước đó có độ chính xác là 85%. Ngoài ra, kết luận rằng **độ chính xác của KNN là cao nhất** trong số ba thuật toán đã sử dụng, cụ thể là **88,52%**.

**Hình 6** thể hiện **tổng số bệnh nhân có hoặc không mắc bệnh tim**. Hình này cho thấy **44% số người được liệt kê trong bộ dữ liệu đang mắc bệnh tim**.