TRƯỜNG ĐẠI HỌC MỞ THÀNH PHỐ HỒ CHÍ MINH

**KHOA CÔNG NGHỆ THÔNG TIN**

---🙡 🕮 🙣---



**BÁO CÁO MÔN HỌC**

**PHÂN TÍCH DỮ LIỆU**

**Đề tài**

**MÔ HÌNH PHÂN TÍCH**

**VÀ DỰ ĐOÁN BỆNH TIM MẠCH**

**Sinh viên thực hiện: Đỗ Thị Thảo Nguyên**

**Mã số sinh viên: 2251010064**

**Sinh viên thực hiện: Nguyễn Thanh Nhi**

**Mã số sinh viên: 2251010067**

**Lớp: DH22CS01**

**Giảng viên hướng dẫn: GV Hồ Hướng Thiên**

**Tháng 04 năm 2025**

**MỤC LỤC**

**CHƯƠNG I: MỤC TIÊU**

**CHƯƠNG II: TÍNH CẤP THIẾT**

**CHƯƠNG III: PHƯƠNG PHÁP ÁP DỤNG**

**CHƯƠNG IV: MÔ TẢ BỘ DỮ LIỆU**

**CHƯƠNG V: PHÂN TÍCH**

**CHƯƠNG I: MỤC TIÊU**

1. **Giới thiệu:**

Mô hình phân tích và dự đoán là sự kết hợp giữa khoa học dữ liệu và học máy giúp người dùng nghiên cứu và xây dựng các dự đoán có tỷ lệ chính xác cao, phục vụ cho các nhu cầu thiết yếu của con người trong mọi lĩnh vực. Chúng tôi đã sử dụng mô hình trên để xây dựng một dự án trong lĩnh vực y tế. Mô hình phân tích và dự đoán bệnh tim mạch được xây dựng dựa trên dữ liệu y tế được thu thập trên nền tảng **Kaggle,** kết hợp với các thuật toán máy học (**Machine Learning**).

**2. Mục tiêu xây dựng mô hình:**

**- Mục tiêu chính**: Mục tiêu của việc xây dựng mô hình máy học là dự đoán khả năng mắc bệnh tim mạch dựa trên các yếu tố y tế như tuổi, giới tính, huyết áp, cholesterol, thói quen sống và các chỉ số sức khỏe khác. Mô hình này sẽ giúp xác định những cá nhân có nguy cơ cao mắc bệnh tim, từ đó hỗ trợ phát hiện sớm và đưa ra các biện pháp phòng ngừa hoặc điều trị phù hợp. Nhờ đó, bệnh nhân có thể thực hiện các can thiệp kịp thời, như điều chỉnh lối sống và nhận được sự chăm sóc y tế sớm, giảm thiểu nguy cơ phát triển các biến chứng nghiêm trọng.

**- Lợi ích**: Sử dụng mô hình máy học trong việc dự đoán nguy cơ mắc bệnh tim mang lại nhiều lợi ích rõ rệt. Trước tiên, nó giúp nhận diện nhanh chóng những người có nguy mắc bệnh, từ đó tạo điều kiện để thực hiện các biện pháp can thiệp sớm. Điều này không chỉ giảm tỷ lệ tử vong mà còn giúp tiết kiệm chi phí điều trị, vì các biện pháp phòng ngừa sớm có thể ngăn chặn bệnh phát triển nghiêm trọng. Bên cạnh đó, việc phát hiện nguy cơ bệnh tim từ sớm sẽ khuyến khích bệnh nhân thay đổi lối sống, cải thiện chế độ ăn uống và tăng cường tập luyện, giúp giảm thiểu nguy cơ mắc bệnh và duy trì sức khỏe lâu dài.

**CHƯƠNG II: TÍNH CẤP THIẾT**

1. **Thực trạng:**

Theo Tổ chức Y tế Thế giới (WHO), mỗi năm có khoảng 17,5 triệu người tử vong do mắc các bệnh lý về tim và mạch máu. Cứ mỗi 2 giây lại có một người chết vì bệnh tim mạch, và cứ 5 giây sẽ có một người bị nhồi máu cơ tim. Tỷ lệ bệnh tim mạch đang ngày càng tăng cao ở các nước đang phát triển, trong đó có Việt Nam. Bên cạnh đó, chi phí khám và điều trị bệnh tim mạch cũng là một gánh nặng kinh tế, với tổng chi phí lên đến hàng trăm tỷ đô la mỗi năm [1].

1. **Mô hình phân tích và dự đoán bệnh tim có vai trò gì trong việc chăm sóc sức khỏe?**

**-** Khả năng dự đoán sớm: giúp người dùng có thể nhận biết về khả năng mắc bệnh tim của bệnh nhân từ sớm, từ đó có thể giảm thiểu các vấn đề nghiêm trọng ảnh hưởng đến sức khỏe.

**-** Hỗ trợ hệ thống y tế: Với thực trạng hiện nay, bệnh tim ngày càng phổ biến, tỷ lệ mắc bệnh ngày càng tăng, việc áp dụng mô hình vào hỗ trợ y tế sẽ giúp đẩy nhanh tiến độ khám chữa bệnh ở các bệnh viện, phân loại nguy cơ mắc bệnh theo các mức độ từ đó đưa ra những biện pháp chữa trị thích hợp.

**-** Ứng dụng khoa học công nghệ vào lĩnh vực y tế: Sự phát triển của công nghệ 4.0 mang lại tiềm năng lớn về độ chính xác cho máy móc, thiết bị khám chữa bệnh. Mô hình dự đoán bệnh tim được xây dựng dựa trên công nghệ học máy mang lại độ tin cậy cao, hỗ trợ y tế trong việc chuẩn đoán và đưa ra lời khuyên về sức khỏe cho người sử dụng dịch vụ.

**-** Cải thiện chất lượng cuộc sống: Mô hình dự đoán bệnh tim không chỉ giúp người bệnh phát hiện sớm tình trạng bệnh mà còn nâng cao chất lượng đời sống cho người bệnh. Độ chính xác trong mô hình sễ giúp cho bệnh nhân có những thay đổi về lối sống hàng ngày để giảm thiểu các nguy cơ gây bệnh từ đó có cuộc sống chất lượng hơn.

**CHƯƠNG III: PHƯƠNG PHÁP ÁP DỤNG**

**1: Dữ liệu ban đầu:**

1.1 Chia dữ liệu: Dữ liệu ban đầu đưa chia làm 2 phần: dữ liệu số và dữ liệu phân loại.

1.2 Thống kê giá trị cột số: Giá trị lớn nhất, giá trị nhỏ nhất, giá trị trung bình, phương sai, độ lệch chuẩn, mode, số lần xuất hiện mode.

1.3 Thống kê các giá trị cột phân loại.

**2: Các thư viện được sử dụng:**

**3: Tiền xử lý dữ liệu:**

3.1 Chuẩn hóa dữ liệu, đưa bộ dữ liệu về 1 thang đo MinMaxScaler.

3.2 Áp dụng thuật toán KNN để điền giá trị thiếu

3.3 Xóa duplicate và outlier

3.4 Trích xuất đặc trưng

**4: Huấn luyện mô hình**

4.1 Huấn luyện 4 mô hình trên tập dữ liệu với số cột như ban đầu và tập dữ liệu với số cột sau khi đã trích xuất và so sánh.

4.2 Huấn luyện mô hình trên tập dữ liệu cuối cùng để tìm ra mô hình tối ưu nhất.

4.3 Tăng tỷ lệ dự đoán nhãn 1(có bệnh)

4.4 Lưu mô hình cuối cùng để sử dụng lâu dài về sau.

**CHƯƠNG IV: MÔ TẢ DỮ LIỆU**

1. **Thu thập dữ liệu y tế:** Dữ liệu y tế hỗ trợ nghiên cứu mô hình phân tích và dự đoán bệnh tim được tìm hiểu và thu thập dựa trên nền tảng Kaggle. Nguồn dữ liệu thu thập được 10000 dòng với các chỉ số sức khỏe liên quan đến bệnh tim mạch, gồm 21 cột: Age, Gender, Blood Pressure, Cholesterol Level, Exercise Habits, Smoking, Family Heart Disease, Diabetes, BMI, High Blood Pressure, Low HDL Cholesterol, High LDL Cholesterol, Alcohol Comsumption, Stress Level, Sleep Hours, Sugar Comsumption, Triglyceride Level, Fasting Blood Sugar, CRP Level, Homocysteine Level, Heart Disease Status. Dựa trên các thông tin nói trên ta có thể xây dựng mô hình dự đoán bệnh tim để hỗ trợ lĩnh vực y tế trong tương lai.
2. **Mô tả dữ liệu:**

- Age: Tuổi

- Gender: Giới tính (Nam/Nữ)

- Blood Pressure: Chỉ số huyết áp

- Cholesterol Level: Mức độ cholesterol

- Exercise Habits: Thói quen tập thể dục (Cao, Thấp, Trung Bình)

- Smoking: Hút thuốc (Có, Không)

- Family Heart Disease: Gia đình có tiền xử bệnh tim (Có, Không)

- BMI: Chỉ số khối cơ thể

- High Blood Pressure: Huyết áp cao (Có, Không)

- Low HDL Cholesterol: Mức cholesterol HDL thấp (Có, Không)

- High LDL Cholesterol: Mức cholesterol LDL cao (Có, Không)

- Alcohol Consumption: Mức độ tiêu thụ rượu (Cao, Thấp, Trung Bình)

- Stress Level: Mức độ căng thẳng (Cao, Thấp, Trung Bình)

- Sleep Hours: Số giờ ngủ

- Sugar Consumption: Mức độ tiêu thụ đường (Cao, Thấp, Trung Bình)

- Triglyceride Level: Mức độ triglyceride

- Fasting Blood Sugar: Đường huyết lúc đói

- CRP Level: Mức độ protein C-reactive

- Homocysteine Level: Mức đọ Homocysteine

- Heart Disease Status: Tình trạng bệnh tim cá nhân (Có, Không)

**CHƯƠNG V: PHÂN TÍCH**

[Tại sao tỷ lệ mắc bệnh tim mạch ở Việt Nam ngày càng báo động?](https://tamanhhospital.vn/ty-le-mac-benh-tim-mach-o-viet-nam/#:~:text=Trong%20nh%E1%BB%AFng%20n%C4%83m%20g%E1%BA%A7n%20%C4%91%C3%A2y%20t%E1%BB%B7%20l%E1%BB%87%20m%E1%BA%AFc,m%E1%BA%A1nh%2C%20%C3%A1p%20l%E1%BB%B1c%20c%C3%B4ng%20vi%E1%BB%87c%20ng%C3%A0y%20c%C3%A0ng%20cao.)

[Các yếu tố nguy cơ dẫn đến bệnh Tim Mạch - Chương trình mục tiêu quốc gia - Cổng thông tin Bộ Y tế](https://moh.gov.vn/chuong-trinh-muc-tieu-quoc-gia/-/asset_publisher/7ng11fEWgASC/content/cac-yeu-to-nguy-co-dan-en-benh-tim-mach?inheritRedirect=false)