|  |
| --- |
| **BỘ GIÁO DỤC VÀ ĐÀO TẠO**  **TRƯỜNG ĐẠI HỌC NGUYỄN TẤT THÀNH**  **KHOA CÔNG NGHỆ THÔNG TIN**  Description: C:\Documents and Settings\Administrator\Desktop\logo dai hoc_khong nen.png  **TIỂU LUẬN MÔN HỌC**  **KHAI THÁC DỮ LIỆU VÀ ỨNG DỤNG**  **DỰ ĐOÁN VỀ VIỆC TÌNH TRẠNG TIM**  Giảng viên giảng dạy: TS. VÕ THỊ HÔNG THẮM  Sinh viên thực hiện: LÊ XUÂN TRỌNG  MSSV : 2100009483  Chuyên ngành : KHOA HỌC DỮ LIỆU  Môn học : KHAI THÁC DỮ LIỆU VÀ ỨNG DỤNG  Khóa : 2021    **Tp.HCM, tháng 05 Năm 2024** |
| **BỘ GIÁO DỤC VÀ ĐÀO TẠO**  **TRƯỜNG ĐẠI HỌC NGUYỄN TẤT THÀNH**  **KHOA CÔNG NGHỆ THÔNG TIN**  Description: C:\Documents and Settings\Administrator\Desktop\logo dai hoc_khong nen.png  **TIỂU LUẬN MÔN HỌC**  **KHAI THÁC DỮ LIỆU VÀ ỨNG DỤNG**  **DỰ ĐOÁN VỀ VIỆC TÌNH TRẠNG TIM**  Giảng viên giảng dạy: TS. VÕ THỊ HÔNG THẮM  Sinh viên thực hiện: LÊ XUÂN TRỌNG  MSSV : 2100009483  Chuyên ngành : KHOA HỌC DỮ LIỆU  Môn học : KHAI THÁC DỮ LIỆU VÀ ỨNG DỤNG  Khóa : 2021    **Tp.HCM, tháng 05 Năm 2024** |

|  |  |
| --- | --- |
| Trường Đại học Nguyễn Tất Thành  **Khoa Công Nghệ Thông Tin**  🙜 🙜 🙝 🙝 | CỘNG HÒA XÃ HỘI CHỦ NGHĨA VIỆT NAM  **Độc lập – Tự do – Hạnh phúc**  🙜 🙜 🙝 🙝 |

**NHIỆM VỤ ĐỒ ÁN**

Họ và tên: Lê Xuân Trọng MSSV: 2100009483

Chuyên ngành: KHOA HỌC DỮ LIỆU Lớp: 21DTH3A

Email: xuantrongle987@gmail.com SĐT: 0366394982

Tên đề tài: DỰ ĐOÁN VỀ TÌNH TRẠNG TIM MẠCH

Giảng viên giảng dạy:

Thời gian thực hiện:22/02/2024 đến 25/04/2024.

Nhiệm vụ/nội dung (mô tả chi tiết nội dung, yêu cầu, phương pháp):

Nội dung và yêu cầu đã được thông qua bộ môn.

*TP.HCM, ngày tháng 05 năm 2024*

|  |  |
| --- | --- |
| **TRƯỞNG BỘ MÔN**  *(Ký và ghi rõ họ tên)* | **GIÁO VIÊN HƯỚNG DẪN**  *(Ký và ghi rõ họ tên)* |

**LỜI MỞ ĐẦU**

Bệnh tim mạch hiện nay đã trở thành một trong những vấn đề sức khỏe quan trọng và nguy hiểm trên toàn cầu. Với sự gia tăng của các yếu tố nguy cơ như lối sống không lành mạnh, stress, chế độ ăn uống không cân đối, và di truyền, việc dự đoán nguy cơ mắc bệnh tim mạch đã trở thành một phần quan trọng trong việc xác định những người có nguy cơ cao và áp dụng các biện pháp phòng ngừa hiệu quả.

Nghiên cứu này tập trung vào việc áp dụng các kỹ thuật học máy để xây dựng mô hình dự đoán nguy cơ mắc bệnh tim mạch dựa trên các yếu tố nguy cơ y tế. Sử dụng dữ liệu lâm sàng và các biến đầu vào như tuổi, giới tính, huyết áp, cholesterol, lịch sử bệnh tiểu đường, và thông tin lối sống, chúng tôi nhằm mục tiêu tạo ra một công cụ dự đoán chính xác và đáng tin cậy để xác định những người có nguy cơ cao mắc bệnh tim mạch.

Nghiên cứu này không chỉ nhằm mục tiêu cải thiện khả năng dự đoán mà còn hướng tới ứng dụng thực tiễn trong việc đưa ra các biện pháp can thiệp sớm, từ đó giảm thiểu nguy cơ mắc bệnh tim mạch và cải thiện chất lượng cuộc sống của cộng đồng. Mục tiêu cuối cùng của nghiên cứu là tạo ra một công cụ hữu ích để hỗ trợ quyết định lâm sàng và chăm sóc sức khỏe cá nhân.

**LỜI CẢM ƠN**

Nếu như phải gửi một lời cảm ơn dành cho các thầy cô đã trợ giúp cho mình thì cho em xin phép mượn câu nói của ngài William a Warrrd, ông từng nói rằng:

“Một người thầy trung bình chỉ biết nói

Một người thầy giỏi biết giải thích

Một người thầy chúng biết minh họa

Một người thầy vĩ đại biết truyền cảm hứng”

Nếu phải dành lời cảm ơn em muốn mượn câu nói này để dành cho các thầy cô phụ trách, cũng như là cô Võ Thị Hồng Thắm giảng viên phụ trách môn khai thác dữ liệu và ứng dụng đã giúp cũng như hỗ trợ chúng em hoàn thành được đồ án về chủ đề dự đoán về vấn đề suy tim. Cũng từ đồ án này em nhận ra mình cần phải có được tội lỗi phàm ăn để khơi lên cơn đói khát về kiến thức của mình từ đó ăn bớt những hạn chế của mình trong con sông kiến thức cũng như tránh xa các tội lỗi kiêu ngạo, lười biếng và đố kị để làm tròn nên những kinh nghiệm mà mình cần phải học hỏi từ mọi người để bước gần hơn tới thành công. Em cũng mong nhận được những lời gớp ý giúp em cải thiện cũng hoàn thiện đồ án hơn.

Cuối cùng em xin kính chúc quí thầy cô có một ngày tốt lạnh.

Em chân thành cảm ơn

Sinh viên thực hiện

LÊ XUÂN TRỌNG

|  |  |
| --- | --- |
| TRƯỜNG ĐẠI HỌC NGUYỄN TẤT THÀNH  **TRUNG TÂM KHẢO THÍ** | **KỲ THI KẾT THÚC HỌC PHẦN**  **HỌC KỲ II NĂM HỌC 2023 - 2024** |

**PHIẾU CHẤM THI TIỂU LUẬN/ĐỒ ÁN**

BM-ChT-11

Môn thi: Khai thác dữ liệu và ứng dụng Lớp học phần:010107579503

Nhóm sinh viên thực hiện:

1. Châu Thế Kiệt Tham gia đóng góp:

2. Đặng Gia Tuân Tham gia đóng góp:

3. Phạm Nguyễn Nhật Trường Tham gia đóng góp:

4. Lê Xuân Trọng Tham gia đóng góp:

5. Tham gia đóng góp:

6. Tham gia đóng góp:

7. Tham gia đóng góp:

8. Tham gia đóng góp:

Ngày thi: 27/05/2024 Phòng thi: L.712

Đề tài tiểu luận/báo cáo của sinh viên : Dự án về dự đoán bệnh suy tim

|  |  |  |  |
| --- | --- | --- | --- |
| **Tiêu chí (theo CĐR HP)** | **Đánh giá của GV** | **Điểm tối đa** | **Điểm đạt được** |
| Cấu trúc của báo cáo | Gồm 3 chương | 1 |  |
| Nội dung |  |  |  |
| * Các nội dung thành phần | Chương 2 và 3 | 5 |  |
| * Lập luận | Chương 1 | 2 |  |
| * Kết luận |  | 1 |  |
| Trình bày | Theo chuẩn format luận văn font chữ 13, canh trái, phải,… | 1 |  |
| **TỔNG ĐIỂM** |  | **10** |  |

|  |  |
| --- | --- |
|  | **Giảng viên chấm thi**  *(ký, ghi rõ họ tên)* |

**NHẬN XÉT CỦA GIẢNG VIÊN GIẢNG DẠY**

*Tp.HCM, Ngày 28 tháng 12 năm 2022*

**Giảng viên giảng dạy**

(Ký tên và ghi rõ họ tên)

**MỤC LỤC**

[CHƯƠNG 1: GIỚI THIỆU 1](#_Toc149293818)

[1.1 Giới thiệu đề tài. 1](#_Toc149293819)

[1.2 Lý do chọn đề tài 2](#_Toc149293820)

[1.3 Mục tiêu đề tài. 3](#_Toc149293821)

[1.4 Phương pháp đề tài. 4](#_Toc149293822)

[1.5 Đối tượng và phạm vi nghiên cứu. 4](#_Toc149293823)

[1.5.1 Đối Tượng Nghiên Cứu 5](#_Toc139821567)

[1.5.2 Phạm Vi Nghiên Cứu 5](#_Toc139821567)

[CHƯƠNG 2.ỨNG DỤNG VÀ THUẬT TOÁN 6](#_Toc149293824)

[2.1Giới thiệu về ngôn ngữ lập trình python. 6](#_Toc149293825)

[2.2 Mô tả thuật toán. 7](#_Toc149293826)

[2.2.1 Thuật toán Logistic Regression 8](#_Toc139821567)

[2.2.2 Thuật toán Random Forest 8](#_Toc139821567)

[2.2.3 Thuật toán K-Nearest Neighbors 9](#_Toc139821567)

[2.3 Xây dựng bộ dữ liệu. 8](#_Toc149293827)

[2.3.1. Bộ dữ liệu gồm các cột 9](#_Toc139821567)

[2.3.2. Tiền xử lý dữ liệu 10](#_Toc139821567)

[2.3.3. Outlier 10](#_Toc139821567)

[2.4 Áp dụng thuật toán và bài toán. 11](#_Toc149293827)

[2.5 Thử nghiệm với thư viên. 12](#_Toc149293827)

[CHƯƠNG 3.XÂY DỰNG ỨNG DỤNG NGÔN NGỮ PYTHON 15](#_Toc149293828)

[3.1. Xây dựng ứng dụng và giải thích 15](#_Toc149293829)

[3.1.1. Logistic Regression 15](#_Toc139821567)

[3.1.2. Decision Tree Classifier 17](#_Toc139821567)

[3.1.3. Support Vector Classifier 18](#_Toc139821567)

[3.1.4. K-Nearest Neighbors 20](#_Toc139821567)

[3.1.5. Random Forest 22](#_Toc139821567)

[3.1.6. XGBoost 24](#_Toc139821567)

[3.1.7. So sánh các phương pháp dự đoán 25](#_Toc139821567)

[3.2. Một số ảnh trực quan hóa được phân tích trong dự án 26](#_Toc149293829)

[3.2.1. Trực quan hóa về độ tuổi 26](#_Toc139821567)

[3.2.2. Trực quan hóa về khả năng mất về độ tuổi 27](#_Toc139821567)

[3.2.3. Trực quan hóa về nồng độ enzyme trong máu 27](#_Toc139821567)

[3.2.4. Trực quan hóa về nồng độ enzyme ảnh hưởng tới khả năng sống sót 27](#_Toc139821567)

[3.2.5. Trực quan hóa về khả năng co bớp của tim 28](#_Toc139821567)

[3.2.6. Trực quan hóa về giới tính 29](#_Toc139821567)

[3.2.7. Trực quan hóa về bệnh tiểu đường ảnh hướng đến bệnh tim 29](#_Toc139821567)

[3.2.8. Tương quan bệnh tiểu đường với tuổi tác 30](#_Toc139821567)

[KẾT LUẬN 21](#_Toc149293831)

[Kết quả đạt được. 21](#_Toc149293832)

[Hạn chế và hướng phát triển . 21](#_Toc149293833)

[TÀI LIỆU THAM KHẢO 23](#_Toc149293834)

**DANH MỤC HÌNH**

[Hình 1: Bảng dữ liệu 9](#_17dp8vu)

[Hình 2: Bảng dữ liệu thông tin dữ liệu 10](#_1ksv4uv)

[Hình 3: Tìm các outlier bằng biểu đồ boxplot 11](#_44sinio)

[Hình 4: Xóa các outlier 11](#_z337ya)

[Hình 5: Các thư viện đã được sử dụng 13](#_3j2qqm3)

[Hình 6: Dữ liệu chuẩn bị cho phần học máy 15](#_1y810tw)

[Hình 7: Khởi tạo mô hình Logistic Regression 17](#_3as4poj)

[Hình 8: Kết quả mô hình Logistic Regression 17](#_2p2csry)

[Hình 9: Khởi tạo mô hình Decision Tree Classifier 18](#_ihv636)

[Hình 10: Kết quả mô hình Decision Tree Classifier 18](#_1hmsyys)

[Hình 11: Khởi tạo mô hình Support Vector Classifier 19](#_2grqrue)

[Hình 12: Kết quả mô hình Support Vector Classifier 19](#_3fwokq0)

[Hình 13: Khởi tạo mô hình K-Nearest Neighbors 20](#_4f1mdlm)

[Hình 14: Kết quả mô hình K-Nearest Neighbors 21](#_3fwokq0)

[Hình 15: Khởi tạo mô hình Random Forest 22](#_3fwokq0)

[Hình 16: Kết quả mô hình Random Forest 23](#_3fwokq0)

[Hình 17: Kết quả mô hình XGBoost 24](#_3fwokq0)

[Hình 18: Kết quả mô hình XGBoost 25](#_3fwokq0)

[Hình 19: Xuất ra phần trăm sai lệch của các dự báo 25](#_3fwokq0)

[Hình 20: Trực quan hóa về phần trăm sai lệch của các dự báo 26](#_3fwokq0)

[Hình 21: Trực quan hóa về tuổi 26](#_3fwokq0)

[Hình 22: Trực quan hóa về tuổi ảnh hướng tới khả năng sống sót 27](#_3fwokq0)

[Hình 23: Trực quan hóa về nồng độ enzyme trong máu 27](#_3fwokq0)

[Hình 24: Trực quan hóa về nồng độ enzyme trong máu tới khả năng sống sót 28](#_3fwokq0)

[Hình 25: Trực quan hóa về tình trạng co bớp của tim 28](#_3fwokq0)

[Hình 26: Trực quan hóa về giới tính ảnh hưởng tới bệnh tim 29](#_3fwokq0)

[Hình 27: Trực quan hóa về bệnh tiểu đường và bệnh tim 30](#_3fwokq0)

[Hình 28: Trực quan hóa về bệnh tiểu đường và tuổi tác 30](#_3fwokq0)

**KÍ HIỆU CÁC CỤM TỪ VIẾT TẮT**

|  |  |
| --- | --- |
| Chữ viết tắt | Ý nghĩa |
| ANN | Artificial Neural Network |
| RNN | Recurrent Neural Network |

# CHƯƠNG 1: GIỚI THIỆU

## Giới thiệu đề tài.

## Đã là một con người sống trên thế giới này không ít thì nhiều chúng ta đều sẽ quan tâm tới vấn đề sức khỏe của mình không sớm thì muốn khi chúng ta cảm nhận được sự thay đổi đó của cơ thể chúng sẽ bắt đầu để tâm đến vấn đề đó. Chính vì nhìn được vấn đề đó nhóm chúng em mới quyết định chọn một đề tài liên quan tới sức khỏe đến sức khỏe thứ ăn sâu vào mỗi chúng ta nhưng để quyết định chọn một loại bệnh để bắt tay vào chủ đề nghiên cứu nó lại là một vấn đề khác. Suy nghĩ thì đã có rồi nhưng giữa vô vàng chủ đề để lựa chọn muốn quyết định thì lại là một vấn đề chính vì thế bọn em đã tính chọn làm về covid-19 thứ rất nổi hồi 5 năm đổ lại đây nhưng nó lại quá cũ cũng vì suy nghĩ này em chợt nhớ tới một câu nói rất hay mà em đã nghe trước đây: “Trong bảng thành tích toàn trường bạn có tự hỏi tại sao mọi người ai cũng sẽ chỉ nhớ người đứng nhất chẳng ai nhớ kẻ đứng nhì là ai cả. Vì dù có về nhì thì kẻ đó cũng chỉ là kẻ thua cuộc.” Nhưng mình không nghĩ vậy để giới thiệu cho mọi người về chủ đề này mình đã thử đi tìm hiểu vào thời gian đại dịch xảy ra vào những năm 2018 và 2019 mọi người có biết không số lượng người mất của covid-19 vào năm đó chiếm 11.5% số người mất trên toàn thế giới trong khi đó số người mất vì bệnh tim chiếm 7.3%. Có thể trong thi đua người ta có thể lãng quên người về sau nhưng khi mắc sai lầm chẳng ai bỏ qua cho kẻ gây ít tội cả. Chưa kể tới vào năm 2023 bệnh tim mạch là nguyên nhân gây tử vong cho gần 36% người trên thế giới đó là lý do mình muốn giới thiệu về bệnh tim mạch một trong những nguyên nhân hàng gây tử vong trên toàn cầu. Cũng vì như vậy việc dự đoán nguy cơ mắc bệnh tim mạch đã trở thành một trong những lĩnh vực nghiên cứu y học quan trọng và khoa học dữ liệu nói riêng, vựa trên mục tiêu đó là xác định những yếu tố dẫn đến nguy cơ và đồng thời cung cấp dự đoán chính xác về việc ai có thể mắc phải căn bệnh này. Chủ đề này của bọn em sẽ tập trung vào việc sử dụng dữ liệu y tế và đồng thời áp dụng những kiến thức đã học để vận dụng vào việc phân tích dữ liệu nhầm xác định những người có nguy cơ mắc bệnh tim mạch. Để giúp cho mọi người hiểu hơn về tình trạng nguy hiểm của bệnh nay bọn em sẽ áp dụng bằng cách sử dụng các phương pháp máy học, học sâu và phân tích số liệu đã có để điều tra, xác định mối tương quan giữa các đặc điểm có thể dẫn đến nguy cơ xuất hiện bệnh cũng như là những người đã mắc mệnh có khả năng sống sót. Với các đặc điểm được xem xét có thể bao gồm tuổi tác, lượng hồng cầu có trong máu, enzyme trong máu, bệnh tiểu đường, số lần co bớp của tim, lối sống, và các chỉ số y khoa như huyết áp, cholesterol, hồng cầu, bạch cầu, vv… Bằng cách kết hợp các thông tin và mô hình dự đoán, từ đó giúp nhận diện nhóm người có nguy cơ cao mắc bệnh tim mạch cũng như những ai có khả năng sống sót cao. Mục tiêu của đồ án mà bọn em đem tới đó là cung cấp các thông tin hữu ích giúp hỗ trợ những người đang bệnh có cái nhìn tích cực cũng như là giúp họ từ hiểu thêm về bệnh tình của mình cũng như là lời cảnh tỉnh những người chưa bệnh cho mình biết được tầm quan trong của sức khỏe cũng như là hạn chế những người chưa bệnh tránh các nguy cơ thậm chí thay đổi lối sống để giảm thiểu nguy cơ mắc bệnh tim mạch. Với việc kết hợp giữa y học và khoa học dữ liệu bọn em hy vọng có thể mang lại nhiều lợi ích cho việc chăm sóc sức khỏe của mọi người cũng như nâng cao ý thức phòng ngừa bệnh tim mạch của cộng đồng.

## 1.2.Lý do chọn đề tài

Nếu phải nói lý do nào để khiến bọn em chọn đề tài "Dự báo về tình trạng những người mắc bệnh tim mạch" thì cự thể cũng không có nhiều lý do nhưng có thể kể đến những lý do như sau đó là:

* Tầm quan trọng y tế công cộng: Bệnh tim mạch là một trong những nguyên nhân hàng đầu gây tử vong trên toàn cầu. Việc hiểu rõ và dự đoán nguy cơ mắc bệnh tim mạch không chỉ giúp cá nhân mà còn hỗ trợ cho các chương trình y tế công cộng trong việc đưa ra các biện pháp phòng ngừa và can thiệp kịp thời.
* Áp dụng của khoa học dữ liệu trong y học: Đây là một cơ hội để kết hợp các nguyên tắc của khoa học dữ liệu và thông tin y tế để tạo ra những mô hình dự đoán chính xác. Sử dụng dữ liệu lớn và các thuật toán máy học, chúng ta có thể hiểu rõ hơn về mối liên hệ giữa các yếu tố nguy cơ và bệnh tim mạch.
* Hỗ trợ quyết định lâm sàng và chăm sóc sức khỏe cá nhân: Mô hình dự đoán nguy cơ mắc bệnh tim mạch có thể cung cấp thông tin hữu ích cho bác sĩ và người bệnh. Điều này giúp quyết định lâm sàng được tốt hơn và hỗ trợ cá nhân trong việc theo dõi sức khỏe cũng như thay đổi lối sống để giảm thiểu nguy cơ mắc bệnh.
* Tiềm năng trong việc phòng ngừa: Việc dự đoán nguy cơ mắc bệnh tim mạch không chỉ dừng lại ở việc chẩn đoán, mà còn mở ra cơ hội cho việc phòng ngừa. Bằng cách nhận diện nhóm người có nguy cơ cao, chúng ta có thể thúc đẩy các biện pháp phòng ngừa, từ thay đổi lối sống đến quản lý y tế cá nhân.

Tóm lại, việc lựa chọn đề tài này không chỉ đáp ứng nhu cầu y tế quan trọng mà còn tạo điều kiện để khám phá sức mạnh của khoa học dữ liệu trong lĩnh vực y học và chăm sóc sức khỏe cộng đồng.

## 1.3.Mục tiêu đề tài.

Mục tiêu mà em đề ra cho đề tài "Dự đoán tình trạng bệnh tim mạch" bao gôm các mục tiêu sau:

* Xây dựng mô hình dự đoán chính xác: Tạo ra các mô hình dự đoán sử dụng dữ liệu y tế và công nghệ dữ liệu để định lượng và đánh giá nguy cơ mắc bệnh tim mạch. Mục tiêu là phát triển các công cụ có khả năng dự đoán chính xác nguy cơ mắc bệnh này dựa trên các yếu tố nguy cơ được xác định.
* Đánh giá yếu tố nguy cơ quan trọng: Xác định và đánh giá sự ảnh hưởng của các yếu tố nguy cơ khác nhau đối với bệnh tim mạch. Loại bỏ hoặc đề xuất cải thiện các yếu tố nguy cơ có thể giúp giảm nguy cơ mắc bệnh.
* Hỗ trợ quyết định lâm sàng: Cung cấp thông tin hữu ích cho các nhà điều trị và bệnh nhân để họ có thể thực hiện các quyết định lâm sàng thông minh. Việc dự đoán nguy cơ mắc bệnh tim mạch có thể giúp tăng cường sự hiểu biết và chủ động trong việc quản lý sức khỏe.
* Phòng ngừa và can thiệp sớm: Sử dụng thông tin từ mô hình dự đoán để phòng ngừa bệnh tim mạch bằng cách đề xuất các biện pháp thay đổi lối sống hoặc can thiệp y tế sớm để giảm thiểu nguy cơ mắc bệnh.
* Nghiên cứu đối tượng đa dạng: Nghiên cứu có thể tiếp cận và đánh giá ảnh hưởng của các yếu tố nguy cơ đối với các nhóm dân số khác nhau, từ người trẻ tuổi đến người cao tuổi, từ các nhóm dân tộc đến các khu vực địa lý khác nhau.

Tóm lại, mục tiêu của đề tài là tạo ra các công cụ dự đoán hiệu quả, cung cấp thông tin hữu ích và hỗ trợ quyết định cho cả bác sĩ và người bệnh, từ đó giúp giảm thiểu nguy cơ mắc bệnh tim mạch và tăng cường sức khỏe tim mạch trong cộng đồng.

## 1.4.Phương pháp đề tài.

Có một số phương pháp tiếp cận đề tài "Dự đoán Nguy cơ Mắc Bệnh Tim Mạch" mà em sử dụng để đạt được mục tiêu của nó

Thu thập và tiền xử lý dữ liệu: Thu thập dữ liệu từ các nguồn đa dạng như bệnh viện, các nghiên cứu y tế, hoặc các thiết bị đeo thông minh. Dữ liệu này bao gồm thông tin về tuổi, giới tính, chỉ số y khoa (huyết áp, cholesterol, đường huyết, v.v.), tiền sử y tế và các yếu tố nguy cơ khác. Tiền xử lý dữ liệu là bước quan trọng để chuẩn bị dữ liệu cho việc phân tích, bao gồm làm sạch dữ liệu, xử lý các giá trị thiếu, và chuẩn hóa dữ liệu.

Phân tích dữ liệu và mô hình hóa: Sử dụng các phương pháp máy học như học máy, học sâu, hay các mô hình thống kê để phân tích dữ liệu và xây dựng mô hình dự đoán. Các mô hình này là logistic regression, support vector machines, decision trees, random forests, neural networks, hoặc các mô hình dự đoán khác.

Đánh giá và tinh chỉnh mô hình: Sử dụng các phương pháp đánh giá mô hình như cross-validation, confusion matrix, ROC curve để đánh giá hiệu suất của mô hình. Tinh chỉnh mô hình bằng cách điều chỉnh các siêu tham số và sử dụng kỹ thuật feature selection để cải thiện chất lượng của mô hình.

Xác định yếu tố nguy cơ quan trọng: Sử dụng các kỹ thuật như feature importance, SHAP values, hay permutation importance để xác định yếu tố nguy cơ có ảnh hưởng lớn nhất đến dự đoán nguy cơ mắc bệnh tim mạch.

Triển khai và ứng dụng: Sau khi có mô hình dự đoán tốt, áp dụng mô hình này để dự đoán nguy cơ mắc bệnh cho các bệnh nhân mới và cung cấp thông tin hữu ích cho quyết định lâm sàng và can thiệp sớm.

Từ đó kết hợp các phương pháp này một cách hợp lý để tạo ra một quy trình phân tích toàn diện và đáng tin cậy trong việc dự đoán nguy cơ mắc bệnh tim mạch.

## 1.5.Đối tượng và phạm vi nghiên cứu.

Đối tượng và phạm vi nghiên cứu trong dự án về dự đoán nguy cơ mắc bệnh tim mạch có thể được xác định như sau:

### **1.5.1. Đối Tượng Nghiên Cứu**

Các cá nhân có nguy cơ mắc bệnh tim mạch cũng như cái đối tượng cho triệu chứng bao gồm những người trong cộng đồng, gia đình hoặc yếu tố nguy cơ cá nhân đặc biệt.

Dữ liệu y tế gồm những thông tin về tuổi, giới tính, huyết áp, cholesterol, lịch sử bệnh tiểu đường và các chỉ số y tế khác được bộ

### **1.5.2. Phạm Vi Nghiên Cứu**

Thu thập và Sử dụng Dữ liệu: Thu thập dữ liệu từ các nguồn khác nhau như cơ sở y tế, bệnh viện, nghiên cứu trước đó hoặc các nguồn dữ liệu y tế khác. Sử dụng dữ liệu này để huấn luyện và đánh giá mô hình dự đoán.

Xây dựng mô hình dự đoán: Tập trung vào việc xây dựng mô hình dự đoán nguy cơ mắc bệnh tim mạch từ dữ liệu thu thập được. Các mô hình này có thể dựa trên các thuật toán học máy như Logistic Regression, Random Forest, KNN, Neural Networks, v.v.

Đánh giá và tinh chỉnh mô hình: Đánh giá hiệu suất của mô hình và tinh chỉnh các siêu tham số để cải thiện khả năng dự đoán.

Ứng dụng trong y học thực tế: Khảo sát khả năng ứng dụng của mô hình dự đoán trong thực tế y học, bao gồm việc áp dụng vào quyết định lâm sàng và hỗ trợ trong quá trình chẩn đoán và can thiệp y tế.

Đối tượng nghiên cứu là những cá nhân có liên quan đến nguy cơ mắc bệnh tim mạch và phạm vi nghiên cứu tập trung vào việc xử lý, phân tích dữ liệu y tế và xây dựng mô hình dự đoán để cải thiện khả năng dự đoán và ứng dụng thực tế trong lĩnh vực y học và chăm sóc sức khỏe.

# CHƯƠNG 2 ỨNG DỤNG VÀ THUẬT TOÁN

## 2.1 Giới thiệu bài toán.

Bệnh tim mạch đang là một trong những vấn đề y tế quan trọng trên toàn cầu, gây tử vong và tăng cường gánh nặng y tế cộng đồng. Việc dự đoán nguy cơ mắc bệnh tim mạch có thể giúp ngăn ngừa và can thiệp kịp thời, từ đó giảm thiểu tỷ lệ mắc bệnh và cải thiện chăm sóc sức khỏe.

Bài toán tập trung vào việc sử dụng khoa học dữ liệu để phân tích và dự đoán nguy cơ mắc bệnh tim mạch của một cá nhân dựa trên thông tin y tế và các yếu tố nguy cơ. Bằng cách sử dụng dữ liệu lâm sàng và các kỹ thuật máy học, mục tiêu là xây dựng mô hình có khả năng ước tính xác suất mắc bệnh tim mạch cho một người dựa trên thông tin đa dạng như tuổi, giới tính, chỉ số y khoa (huyết áp, cholesterol, tiểu đường), lối sống và tiền sử y tế.

## 2.2 Mô tả thuật toán.

### **2.2.1. Thuật toán Logistic Regression**

Logistic Regression được sử dụng để ước tính xác suất một sự kiện xảy ra hoặc không xảy ra, trong trường hợp này là xác suất một người có nguy cơ mắc bệnh tim mạch dựa trên các yếu tố nguy cơ y tế. Thuật toán này sẽ tạo ra một mô hình có khả năng dự đoán xác suất mà một người có thể mắc bệnh tim mạch dựa trên dữ liệu đầu vào như tuổi, giới tính, huyết áp, cholesterol, tiểu đường, hút thuốc, và các yếu tố khác.

Thuật toán này cũng cung cấp thông tin về ảnh hưởng của mỗi yếu tố nguy cơ đối với nguy cơ mắc bệnh. Bằng cách xem xét các hệ số trong mô hình Logistic Regression, chúng ta có thể hiểu được mức độ ảnh hưởng của mỗi biến đầu vào đến xác suất mắc bệnh tim mạch.

Logistic Regression cũng cho phép điều chỉnh ngưỡng quyết định để phù hợp với yêu cầu cụ thể của bài toán. Điều này có thể thực hiện bằng cách điều chỉnh giá trị ngưỡng để tối ưu hoá độ chính xác, độ nhạy hoặc độ cụ thể tùy thuộc vào yêu cầu cụ thể của việc dự đoán nguy cơ mắc bệnh tim mạch.

Những ứng dụng của Logistic Regression trong dự đoán nguy cơ mắc bệnh tim mạch giúp cung cấp thông tin hữu ích về xác suất mắc bệnh và tác động của các yếu tố nguy cơ, từ đó hỗ trợ trong việc đưa ra quyết định lâm sàng và các biện pháp can thiệp sớm để giảm thiểu nguy cơ mắc bệnh cho cá nhân và cộng đồng.

### **2.2.2. Thuật toán Random Forest**

Random Forest là một phương pháp học máy ensemble, kết hợp nhiều cây quyết định (Decision Trees) với nhau để tạo ra một mô hình mạnh mẽ hơn. Trong trường hợp này, nó được sử dụng để dự đoán nguy cơ mắc bệnh tim mạch dựa trên các yếu tố nguy cơ y tế.

Random Forest có khả năng làm việc với các tập dữ liệu lớn và đa chiều, và có khả năng tránh hiện tượng overfitting (quá khớp) thông qua việc sử dụng nhiều cây quyết định khác nhau và kết hợp kết quả từ chúng.

Random Forest cung cấp thông tin về mức độ quan trọng của các biến đầu vào trong việc dự đoán nguy cơ mắc bệnh tim mạch. Bằng cách xem xét độ quan trọng của các biến, chúng ta có thể hiểu được đóng góp của từng yếu tố nguy cơ vào việc dự đoán kết quả.

Random Forest cho phép đánh giá hiệu suất của mô hình thông qua các kỹ thuật như Out-of-Bag Evaluation và cung cấp cơ hội để điều chỉnh các siêu tham số để tối ưu hoá hiệu suất.

Tùy thuộc vào mục tiêu cụ thể của bài toán, chúng ta có thể điều chỉnh ngưỡng quyết định để phân loại người dùng vào các nhóm nguy cơ khác nhau (ví dụ: nguy cơ cao, trung bình, thấp) dựa trên xác suất dự đoán từ Random Forest.

Sử dụng Random Forest trong dự đoán nguy cơ mắc bệnh tim mạch giúp cải thiện khả năng dự đoán và cung cấp thông tin quan trọng về mức độ quan trọng của các yếu tố nguy cơ, từ đó hỗ trợ trong việc đưa ra các quyết định lâm sàng và các biện pháp phòng ngừa hiệu quả.

### **2.2.3. Thuật toán K-Nearest Neighbors**

KNN được sử dụng để phân loại các cá nhân dựa trên các điểm dữ liệu gần nhất trong không gian đặc trưng. Trong trường hợp này, KNN có thể dùng để dự đoán nguy cơ mắc bệnh tim mạch của một cá nhân dựa trên thông tin của các người có đặc tính tương tự.

Thuật toán KNN xác định các điểm dữ liệu "gần" nhất với điểm dữ liệu cần dự đoán dựa trên độ đo khoảng cách (thường sử dụng khoảng cách Euclid). Các điểm dữ liệu gần nhất này sẽ được sử dụng để quyết định nhãn hoặc giá trị dự đoán cho điểm dữ liệu mới.

Mặc dù KNN không trực tiếp cung cấp xác suất, nhưng có thể tính toán xác suất dự đoán dựa trên số lượng láng giềng gần nhất thuộc vào mỗi lớp. Xác suất có thể được ước lượng bằng cách đếm số lượng láng giềng thuộc vào mỗi lớp và chia cho tổng số láng giềng.

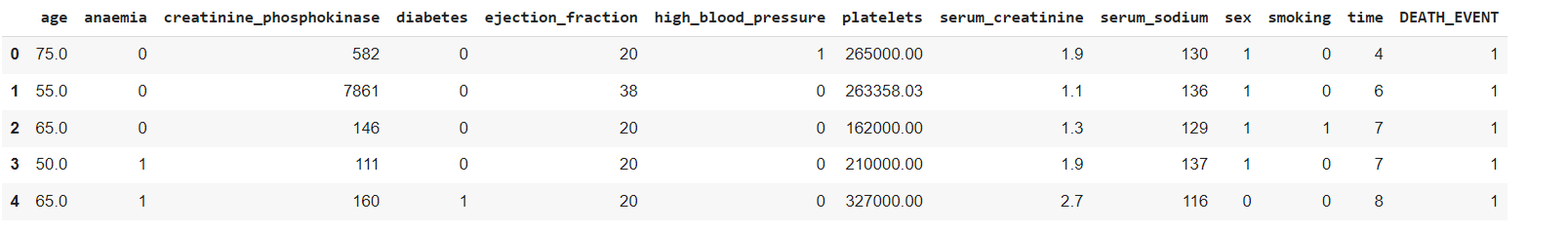
KNN có một siêu tham số quan trọng là số láng giềng (K) được sử dụng để dự đoán. Việc lựa chọn K thích hợp có thể ảnh hưởng đến hiệu suất của mô hình. Quá ít láng giềng có thể dẫn đến overfitting, trong khi quá nhiều láng giềng có thể dẫn đến underfitting.

KNN cho phép đánh giá hiệu suất mô hình bằng cách sử dụng các phương pháp như cross-validation, confusion matrix, độ chính xác, độ nhạy và độ cụ thể để đánh giá chính xác của mô hình.

KNN có thể được sử dụng trong dự đoán nguy cơ mắc bệnh tim mạch để phân loại và dự đoán nguy cơ dựa trên dữ liệu từ các người có đặc điểm tương tự, tuy nhiên, việc lựa chọn tham số K cần được quan tâm để đảm bảo hiệu suất tốt nhất cho mô hình.

## 2.3. Xây dựng bộ dữ liệu.

### **2.3.1. Bộ dữ liệu gồm các cột**



**Hình 1: Bảng dữ liệu**

1. age: Tuổi

2. anaemia: Tình trạng thiếu máu hay còn gọi là giảm hồng cầu

* 0: Lượng hồng cầu bình thường
* 1: Lượng hồng cầu có trong máu khá thấp

3. creatinine phosphokinase (CPK): Nồng độ enzyme CPK trong máu (mcg/L)

4. diabetes: Những bệnh nhân mắc bệnh tiểu đường

* 0: Là những bệnh nhân không có mắc bệnh tiểu đường
* 1: Là những bệnh nhân mắc bệnh tiểu đường

5. ejection fraction: Phần trăm máu rời khỏi tim trong mỗi cơn co thắt (phần trăm)

6. high blood pressure: Những bệnh nhân mắc bệnh cao huyết áp

* 0: Là những bệnh nhân không có mắc bệnh cao huyết áp
* 1: Là những bệnh nhân mắc bệnh cao huyết áp

7. platelets: Số lương tiểu cầu trong máu (kiloplatelets/ mL)

8. sex: Giới tính

* 0: Là nữ
* 1: Là nam

9. serum creatinine: Mức creatinine huyết thanh trong máu (mg/dL)

10. serum sodium:Mức natri huyết thanh trong máu (mEq/L)

11. smoking: Nếu bệnh nhân hút thuốc hay không

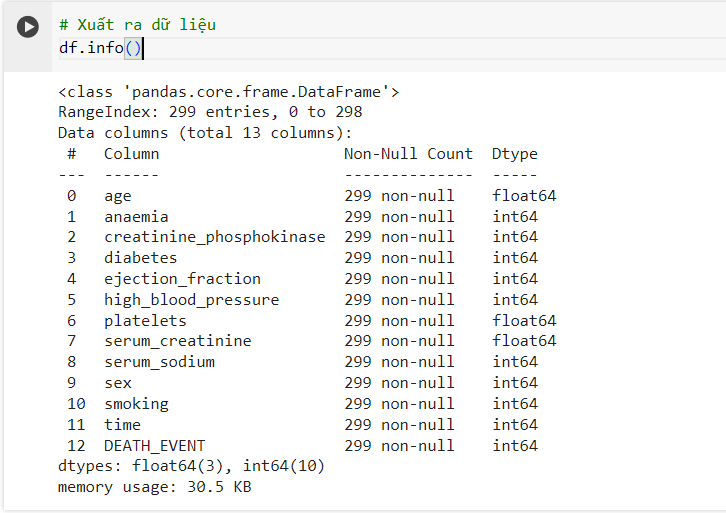
* 0: không có hút thuốc
* 1: Có hút thuốc

12. time: Thời gian theo dõi (ngày)

13. [target] death event: Nếu bệnh nhân chết trong thời gian theo dõi

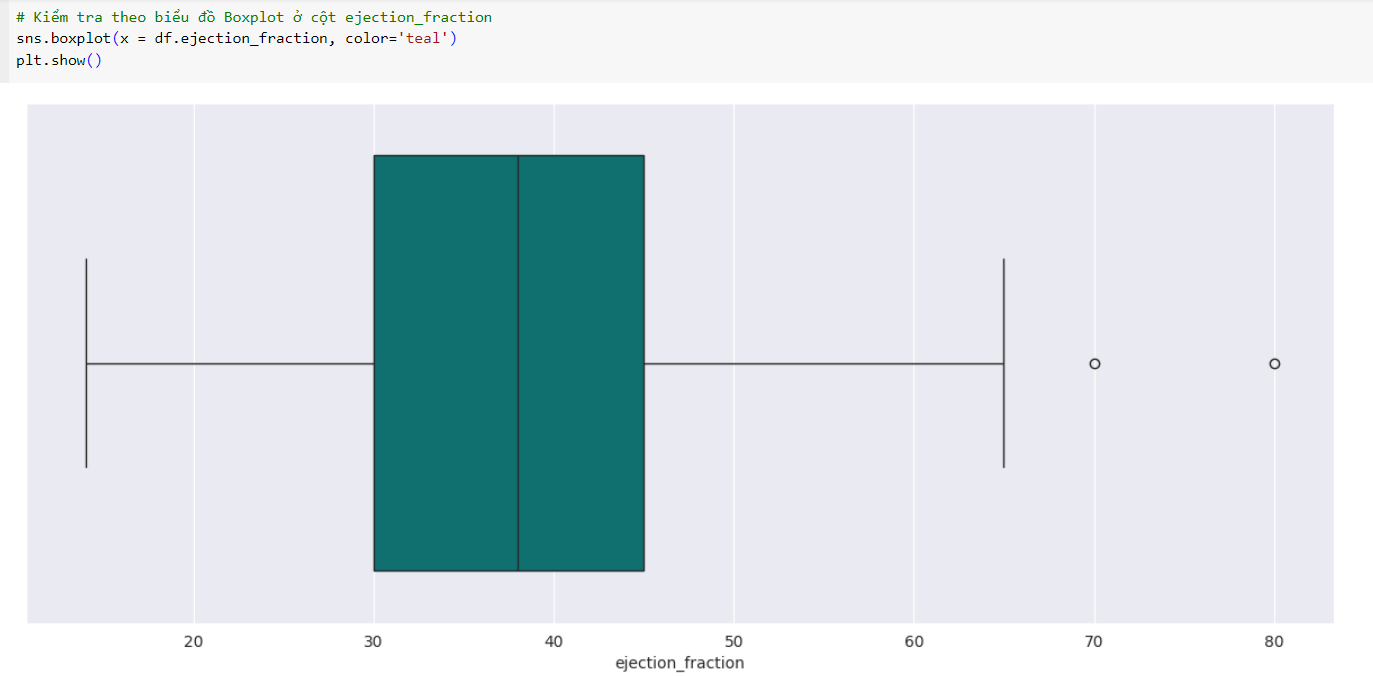
* 0: Còn sống
* 1: Đã mất

### **2.3.2. Tiền xử lý dữ liệu**

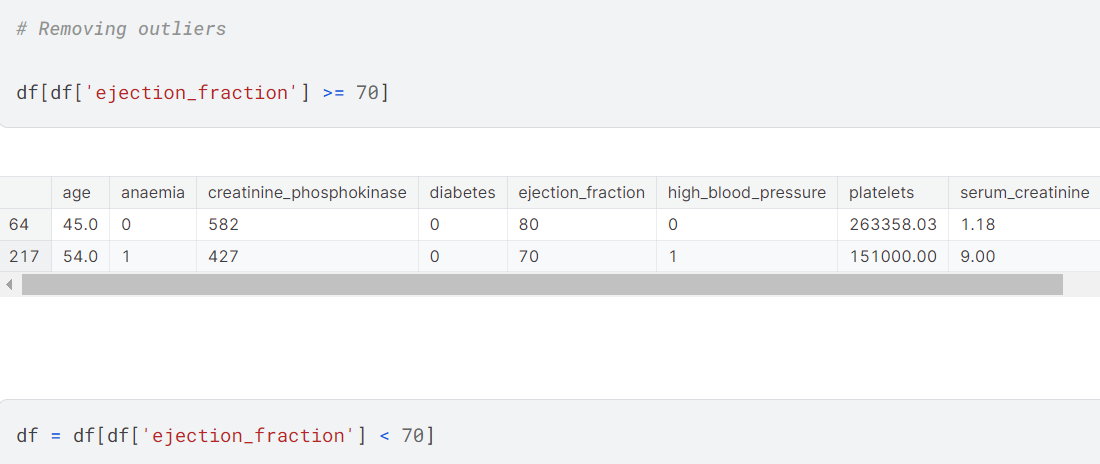


**Hình 2: Bảng thông tin dữ liệu**

**2.3.3. Outlier**

****

**Hình 3: Tìm các outlier bằng biểu đồ boxplot**

****

**Hình 4: Xóa các outlier**

## 2.4. Áp dụng thuật toán vào bài toán.

Trong đề tài “dự đoán nguy cơ mắc bệnh tim mạch”, có nhiều thuật toán machine learning và deep learning có thể được áp dụng để xây dựng mô hình dự đoán. Dưới đây là một số thuật toán mà em đã áp dụng trong dự án này:

**Logistic Regression**: Một thuật toán phân loại thường được sử dụng để dự đoán kết quả nhị phân dựa trên các biến đầu vào. Trong trường hợp này, nó có thể được sử dụng để ước tính xác suất mắc bệnh tim mạch dựa trên các yếu tố nguy cơ.

**Decision Trees và Random Forests**: Decision Trees và Random Forests là các phương pháp phân loại linh hoạt, có khả năng xử lý dữ liệu không tuyến tính và làm việc tốt với các biến đầu vào đa chiều. Random Forests cũng có thể giúp giảm overfitting.

**Support Vector Machines (SVM):** SVM là một phương pháp phân loại mạnh mẽ, có khả năng tạo ra ranh giới quyết định phức tạp hơn giữa các lớp dữ liệu.

**Neural Networks (Deep Learning):** Mạng nơ-ron và học sâu là các phương pháp mạnh mẽ trong việc học từ dữ liệu phức tạp và rất phù hợp trong việc dự đoán nguy cơ mắc bệnh tim mạch.

**Gradient Boosting Algorithms**: Các thuật toán này tập trung vào việc xây dựng một loạt các mô hình yếu và kết hợp chúng để tạo ra một mô hình dự đoán tốt hơn. K-nearest neighbors (KNN): KNN dựa trên nguyên tắc gần nhất trong không gian đặc trưng và có thể được sử dụng để dự đoán nguy cơ mắc bệnh tim mạch dựa trên các điểm dữ liệu tương tự. Ensemble Learning: Kết hợp các mô hình để tạo ra một mô hình tổ hợp mạnh mẽ hơn, như Voting Classifier, Stacking.

## 2.5. Thực nghiệm với thư viện.

Chắc hẳng bất kì ai học về lập trình đều phải biết đến một thuật ngữ “Modules” đó là thuật ngữ chỉ một dạng thư viện trong lập trình chưa các đoạn mã thường được sử dụng do các nhà developer phát triển nên khi lập trình. Cũng vì vậy để chuẩn bị cho đề tài “dự đoán tình trạng bệnh tim mạch” thư viện là một trong những thứ không thể thiếu khi xây dựng đề tài vì thứ tiếp theo mà em muốn giới thiệu tới mọi người một chút về những thư viện mà em đã dùng để hoàn thành đồ án này.



**Hình 5: Các thư viện đã được sử dụng**

**1.pandas (pd):**

Chúng ta sẽ bắt đầu với thư viện gần gũi với những người mới bắt đầu học về python nói tới đây chắc mọi người cũng biết bọn mình đang nhắc tới thư viện gì rồi nhỉ. Bingo đoán đúng rồi đó thư viện mà mình muốn nói tới ở đầy đó là Pandas. Một thư viện mạnh mẽ cho xử lý và phân tích dữ liệu trong Python bên cạnh đó còn cung cấp các cấu trúc dữ liệu mạnh mẽ như DataFrame, Series thuận tiện cho việc dễ dàng trực quan hóa.

**2.matplotlib.pyplot (plt):**

Tiếp theo em xin được phép giới thiệu về thư viện thứ hai mà em sử dụng đó là thư viện matplotlib là một thư viện vẽ đồ thị trong Python.

Pyplot là một module của matplotlib cung cấp các chức năng để tạo và hiển thị đồ thị bằng các dụng cụ như Tkinter, wxPython.

**3.seaborn (sns):**

Seaborn là một thư viện trực quan hóa dữ liệu xây trên matplotlib.

Cung cấp các giao diện cao cấp và mặc định giúp tạo ra các biểu đồ thống kê có trực quan hóa tốt.

**4.numpy (np):**

Numpy là một thư viện quan trọng cho tính toán số học trong Python.

Cung cấp các đối tượng mảng và nhiều hàm toán học mạnh mẽ giúp thao tác dữ liệu nhanh chóng và hiệu quả.

**5.scikit-learn**

(sklearn là tên ngắn gọn) là một thư viện Python phổ biến trong lĩnh vực học máy và thống kê. Nó cung cấp một loạt các công cụ để xây dựng và đánh giá các mô hình học máy.

Thư viện được cấp phép bản quyền chuẩn FreeBSD và chạy được trên nhiều nền tảng Linux.

**6. plotly**

Plotly là một thư viện đồ họa tương tác, mã nguồn mở, và dựa trên nền tảng trình duyệt để tạo ra các biểu đồ tương tác, trực quan.

# CHƯƠNG 3 XÂY DỰNG ỨNG DỤNG NGÔN NGỮ PYTHON

## 3.1 Xây dựng ứng dụng và giải thích

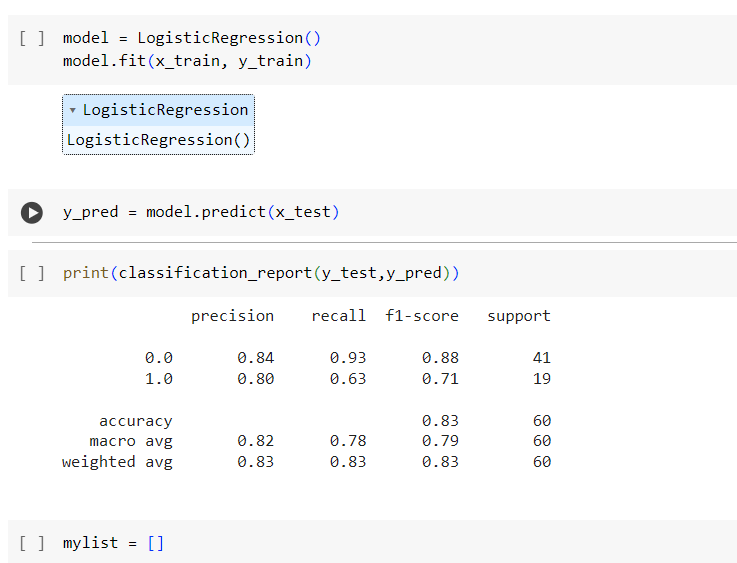
Trước khi đi vào các thuật toán dự đoán thì em xin phép giải thích một ít về phần dữ liệu mà em sẽ dự đoán như sau:



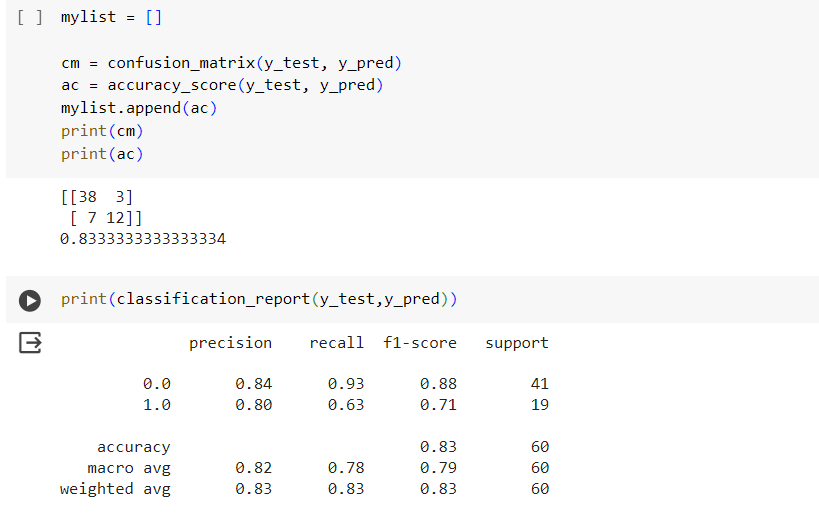
**Hình 6: Dữ liệu chuẩn bị cho phần học máy**

### **3.1.1. Logistic Regression**

Thuật toán Logistic Regression là một phương pháp thống kê và học máy được sử dụng chủ yếu trong bài toán phân loại, nơi chúng ta muốn dự đoán lớp (class) của một đối tượng dựa trên các đặc trưng (features) của nó. Trong trường hợp của đề tài “dự đoán tình trạng bệnh tim mạch” chúng ta có thể sử dụng Logistic Regression để dự đoán những người có khả năng qua đời trong thời gian điều trị



**Hình 7: Khởi tạo mô hình Logistic Regression**

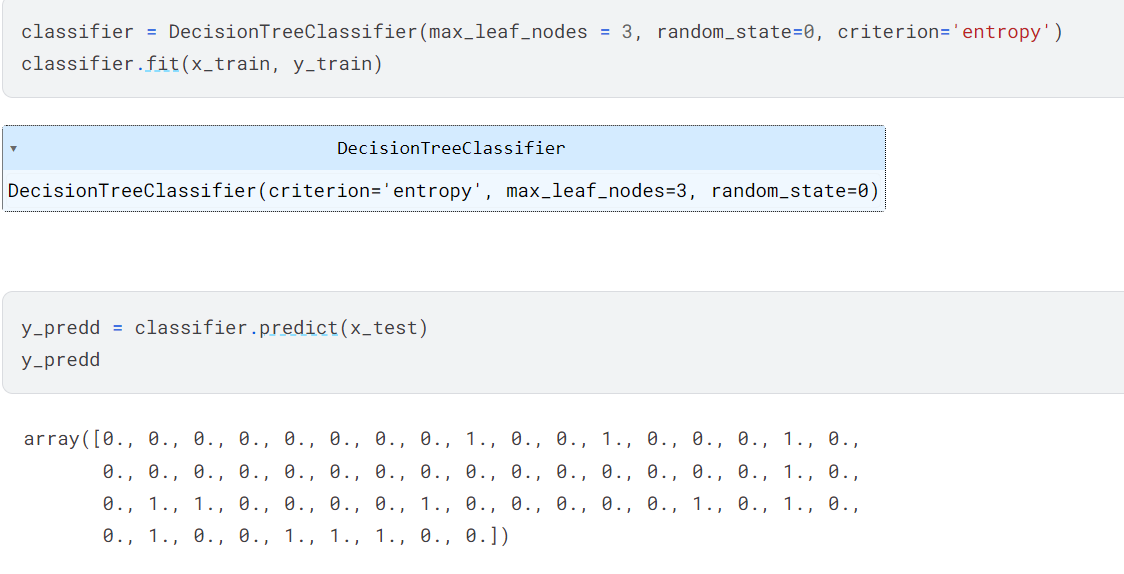


**Hình 8: Kết quả mô hình Logistic Regreession**

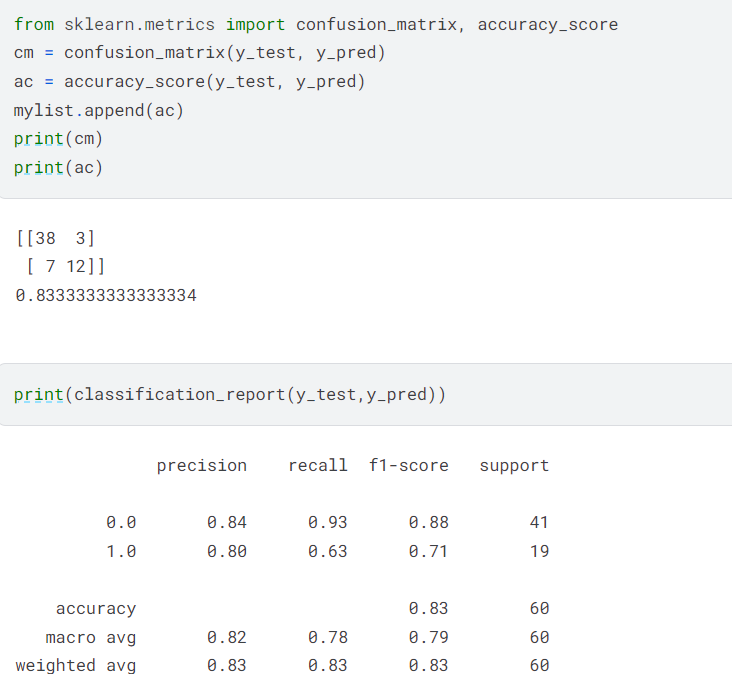
Như mọi người có thể thấy được ở phần kết quả của bài dự đoán này thì dự đoán của mô hình Logistic Regreesion chiếm 0.83

### **3.1.2. Decision Tree Classifier**

Bên cạnh thuật Logistic Regreesion em còn sử dụng thêm cả thuật toán Decision Tree Classifier là các thuật toán học máy có thể thực hiện công việc phân loại và hồi quy, thậm chí có thể thực hiện các công việc yêu cầu xuất nhiều dạng dữ liệu khác nhau. Cây quyết định có tiềm năng lớn, hiệu suất ổn định khi thực hiện học trên các tập dữ liệu phức tạp. Bọn em vẫn sẽ dựa trên thuật toán này để tiếp tục dự đoán về số ca mắc bênh có khả năng qua đời trong thời gian điều trị



**Hình 9: Khởi tạo mô hình Decision Tree Classifier**

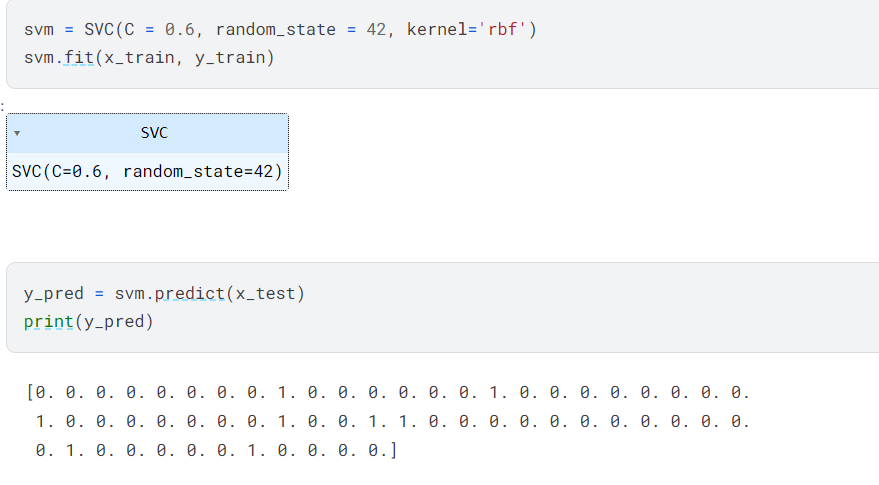
****

**Hình 10: Kết quả mô hình Decision Tree Classifier**

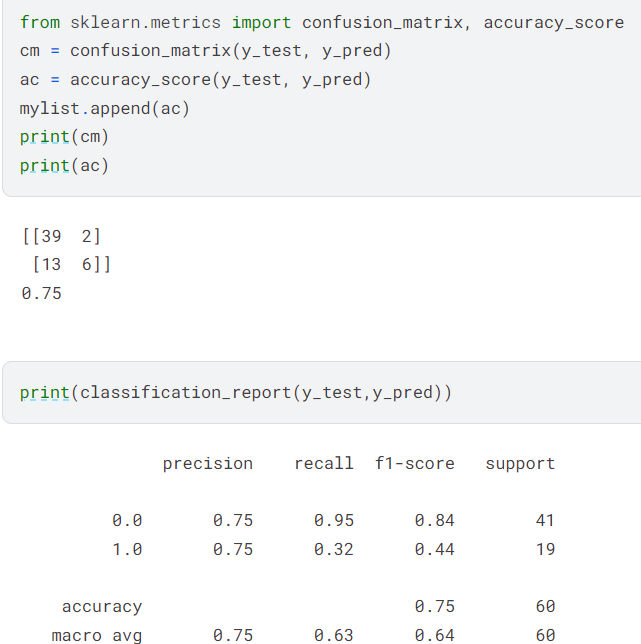
Kết quả của dự đoán này cũng không có sự trên lệch quá nhiều so với dự đoán Logistic Regreesion là mấy cũng chiếm 0.83

### **3.1.3. Support Vector Classifier**

Support Vector Classifier là bài toán đi tìm mặt phân cách sao cho margin tìm được là lớn nhất, đồng nghĩa với việc các điểm dữ liệu an toàn nhất so với mặt phân cách.



**Hình 11: Khởi tạo Support Vector Classifier**

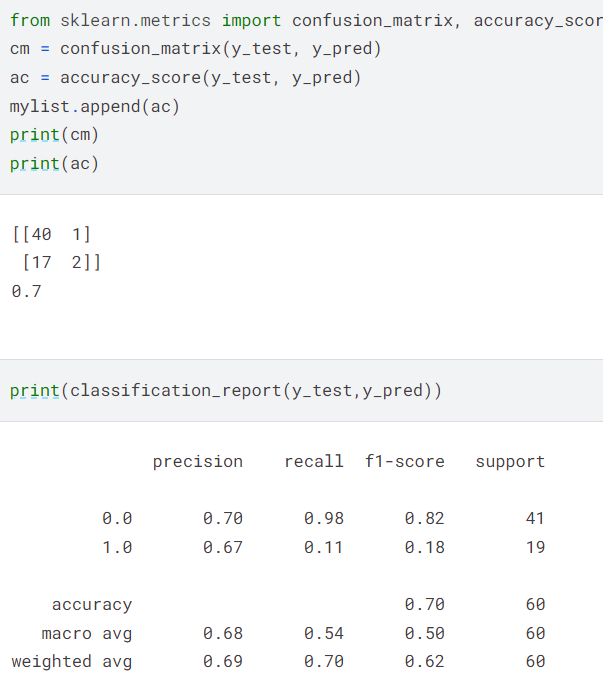


**Hình 12: Kết quả mô hình Support Vector Classifier**

### **3.1.4. K-Nearest Neighbors**



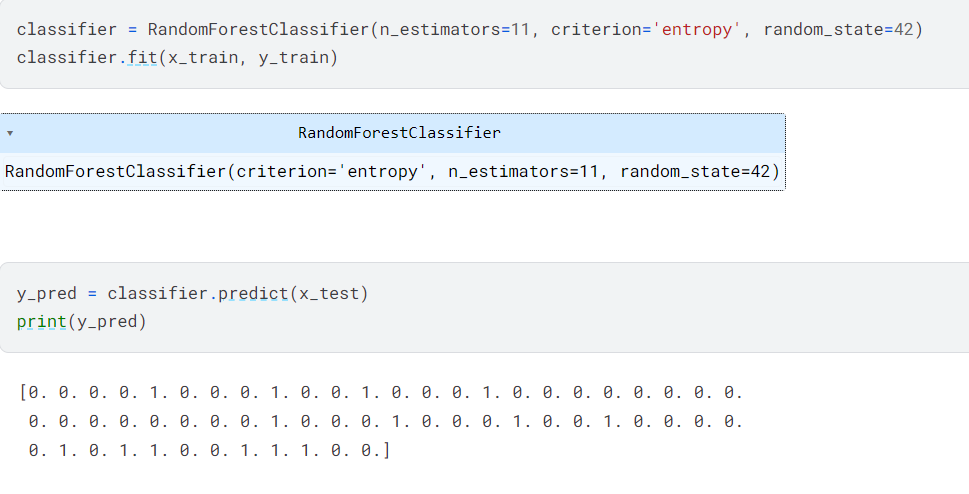
**Hình 13: Khởi tạo K-Nearest Neighbors**

****

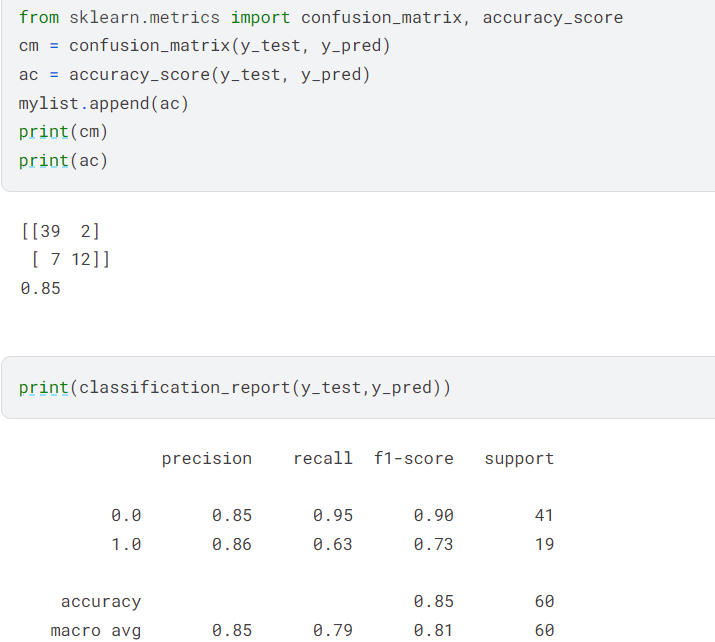
**Hình 14: Kết quả mô hình K-Nearest Neighbors**

Như mọi người có thể thấy thì theo dự đoán K-Nearest Neighbors chiếm 0.7

### **3.1.5. Random Forest**



**Hình 15: Khởi tạo Random Forest**

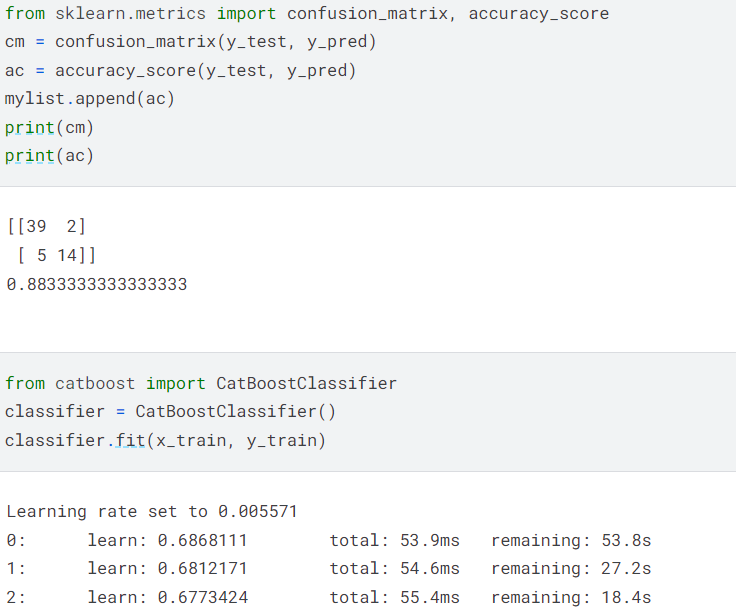


**Hình 16: Kết quả mô hình Random Forest**

### **3.1.6. XGBoost**



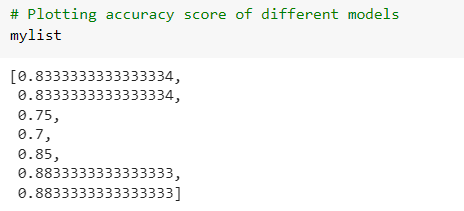
**Hình 17: Khởi tạo XGBoost**

****

**Hình 18: Kết quả mô hình XGBoost**

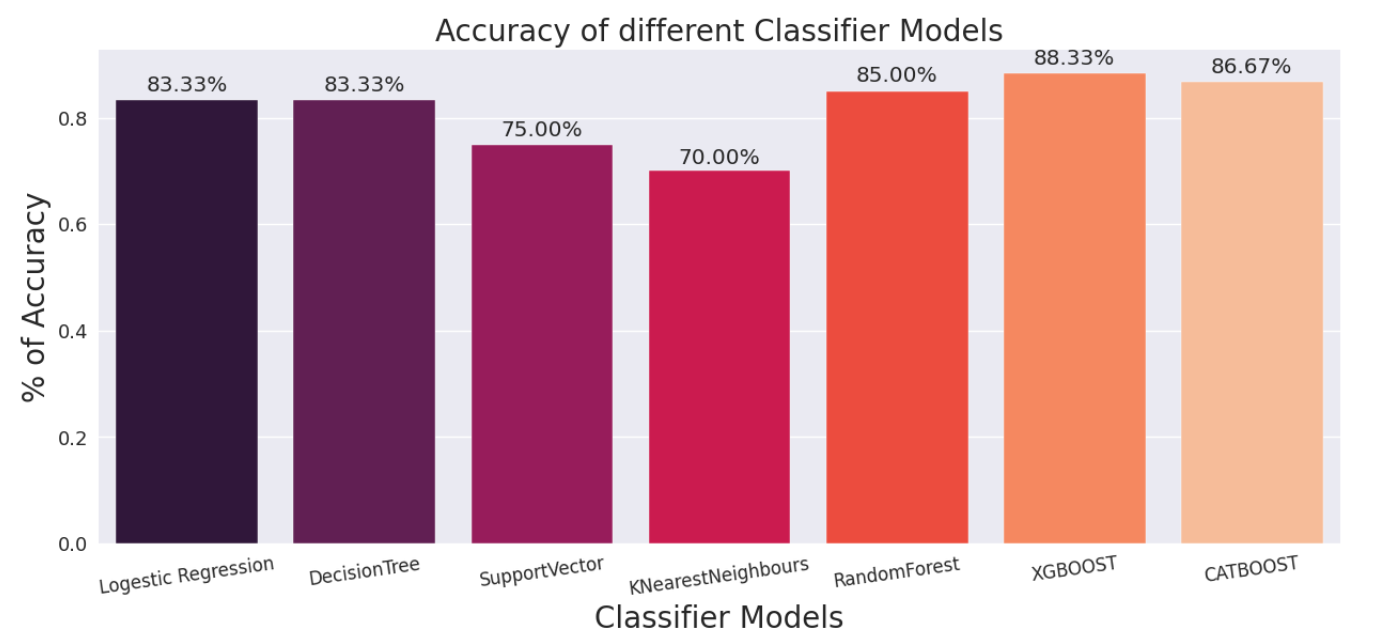
Từ phương pháp XGBoost ta có thể thấy được rằng phương pháp này có độ sai lệch là 0.883%

**3.1.7. So sanh các phương pháp dự đoán**



**Hình 19: Xuất ra phần trăm sai lệch của các dự báo**

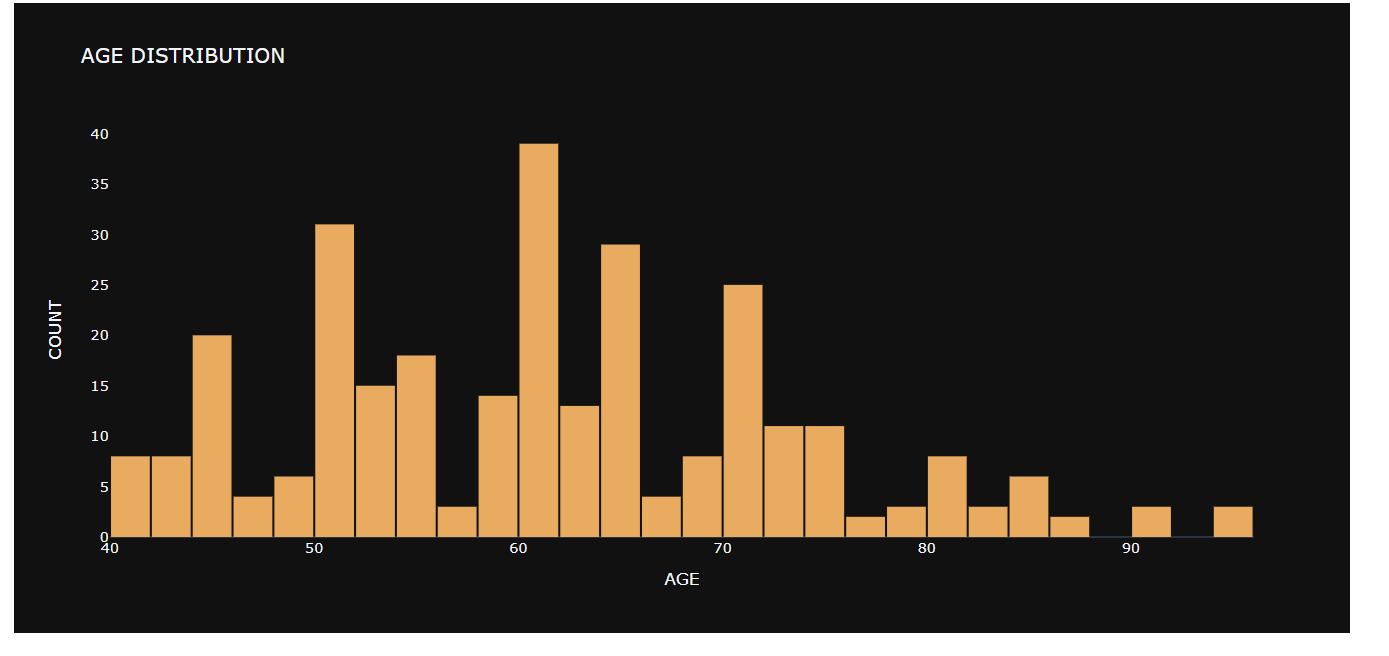
Từ bảy phương pháp dự đoán bao gồm: Logistic Regression, Decision Tree Classifier, Support Vector Classifier, K-Nearest Neighbors, Random Forest, XGBoost. Ta có thể thấy được rằng mô hình K-Nearest Neighbors là mô hình có độ dự báo sai lệch thấp nhất



**Hình 20: Trực quan hóa của phần trăm sai lệch của các dự báo**

**3.2. Một số ảnh trực phân hóa được phân tích trong dự án**

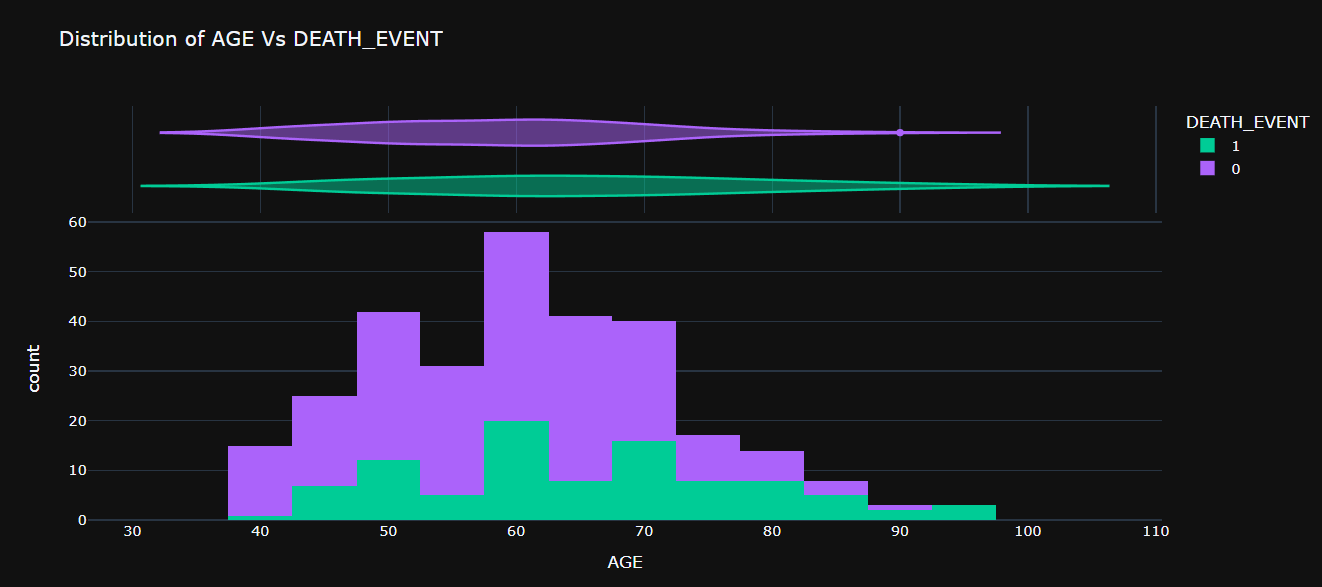
**3.2.1. Trực quan theo độ tuổi**



**Hình 21: Trực quan hóa về tuổi**

Như mọi người đã thấy ở ảnh trực quan hóa ở trên thì ta có thể thấy được rằng trong bộ dữ liệu này thì số bệnh nhân mắc bệnh thuộc ở giai đoạn trung niên của con người phân bố nhiều nhất ở độ tuổi từ 50 tới 70 tuôi trong đó thì dưới 40 tuổi lại có rất ít người mắc bệnh này

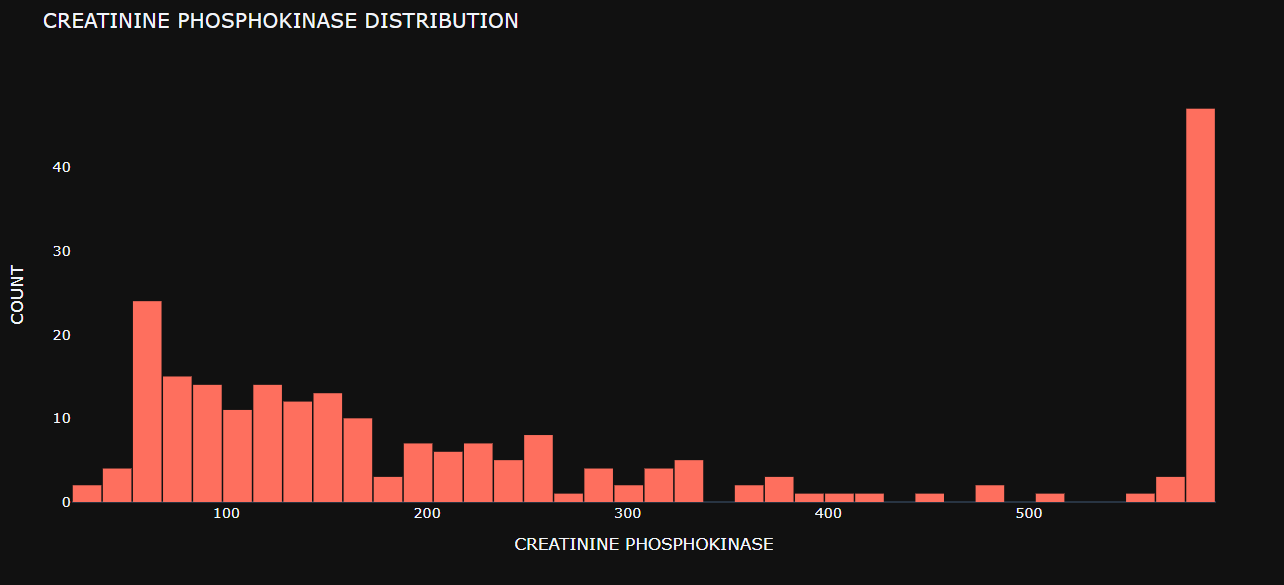
**3.2.2. Trực quan theo khả năng mất ở các độ tuổi**

****

**Hình 22: Trực quan hóa về tuổi ảnh hưởng tới khả năng sống sót**

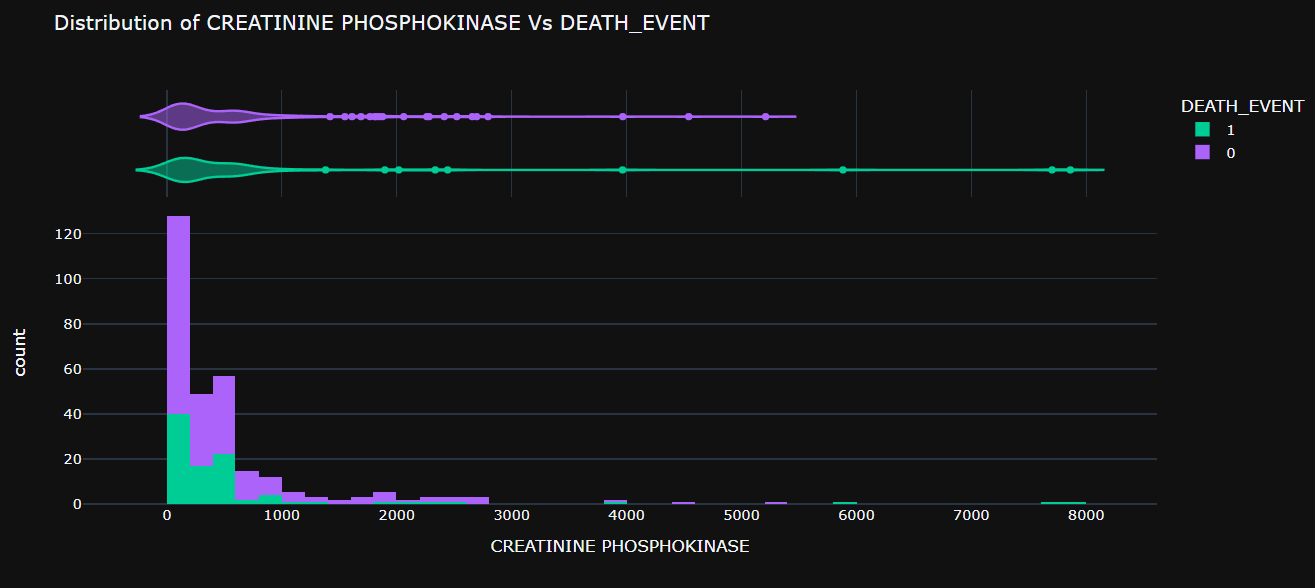
Người xưa từng có câu“Qua tứ tuần đối với người trẻ thì đã già, quan ngũ tuần thì đối với người già thì vẫn còn trẻ, qua lục tuần là bước vào giai đoạn khó khăn nhất của một đời người” có thể thấy được đây hoàn toàn là tinh hoa mà cả người xưa đã tích góp được nó hoàn toàn đúng khi ta nhìn dữ liệu đã được trực quan hóa lại khi tỷ lệ của người mất ở năm 60 tuổi là cao nhất.

**3.2.3. Trực quan nồng độ enzyme trong máu**

****

**Hình 23: Trực quan hóa về nồng độ enzyme trong máu**

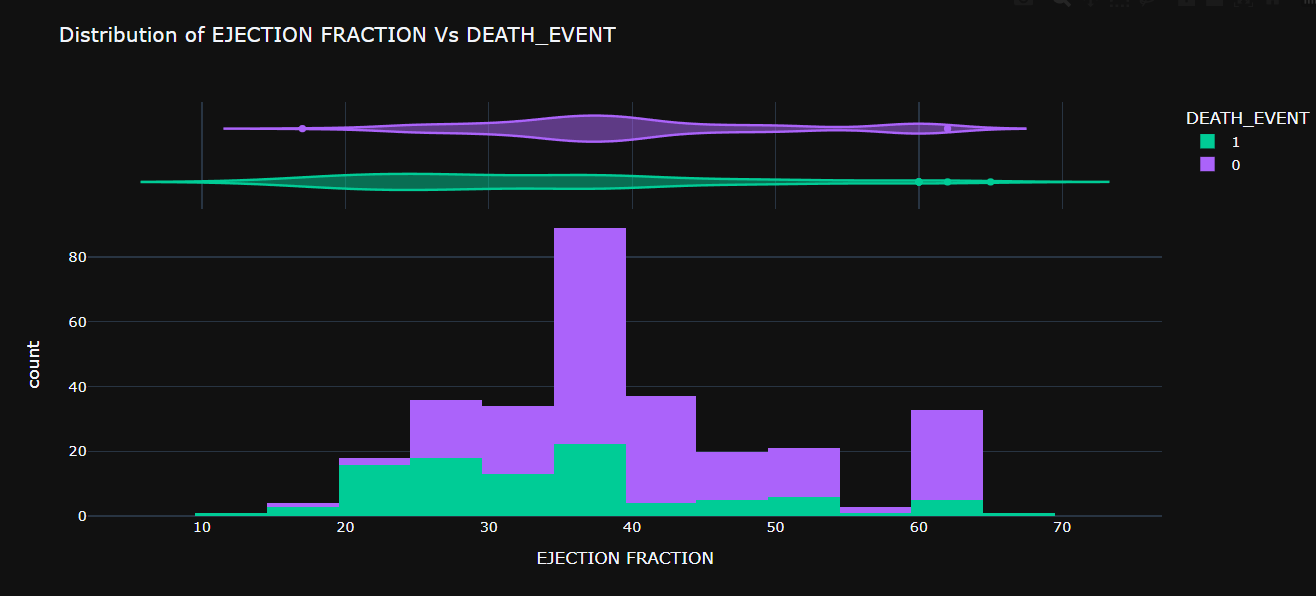
**3.2.4. Trực quan nồng độ enzyme ảnh hưởng tới việc mất**

****

**Hình 24: Trực quan hóa về nồng độ enzyme ảnh hưởng khả năng sống sót**

Theo sơ đồ trên mà mọi người có thể thấy được người có nồng độ enzyme càng thấp thì tỷ lệ họ mất càng cao trong khi đó những người có tỷ lệ enzyme cao thì tỷ lệ sinh của họ rất thấp

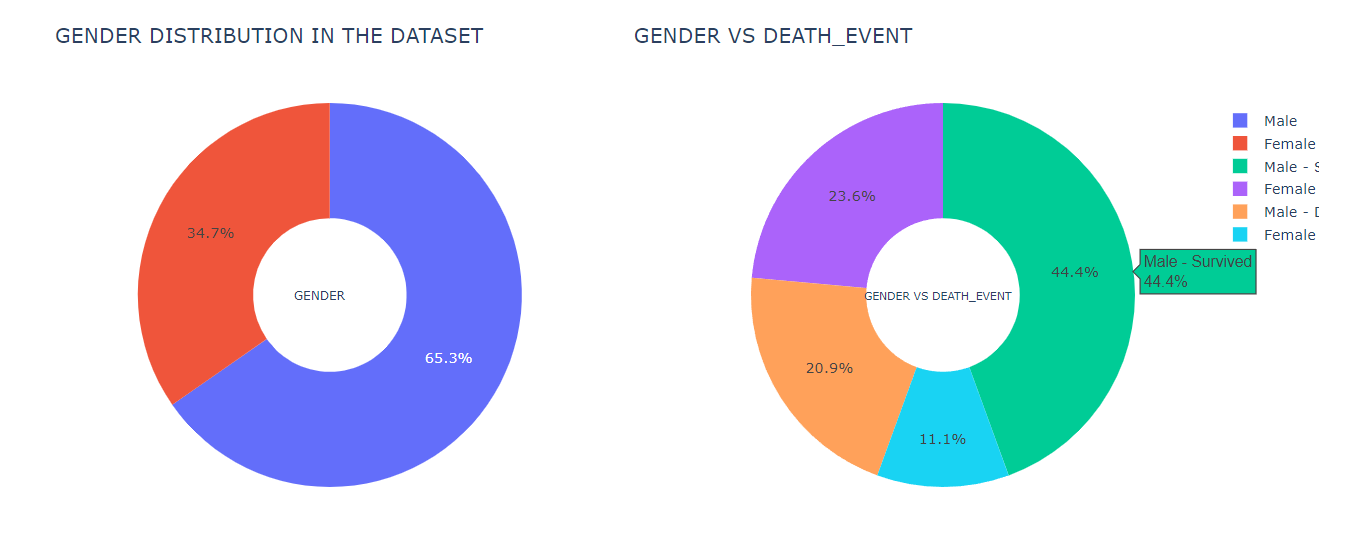
**3.2.5. Trực quan về khả năng co bớp của tim**



**Hình 25: Trực quan hóa về tình trạng co bớp của tim**

Theo biểu đồ trên thì tuổi tác và tình trạng co bớp của tim có mối tương quan rất mật thiếu vì tùy từng độ tuổi mà sẽ có tình trạng co thắng khác nhau từ đó cũng ảnh hưởng tới tình trạng của bệnh nhân đồng thời cũng quyết định khả năng sinh tồn của các bệnh nhân

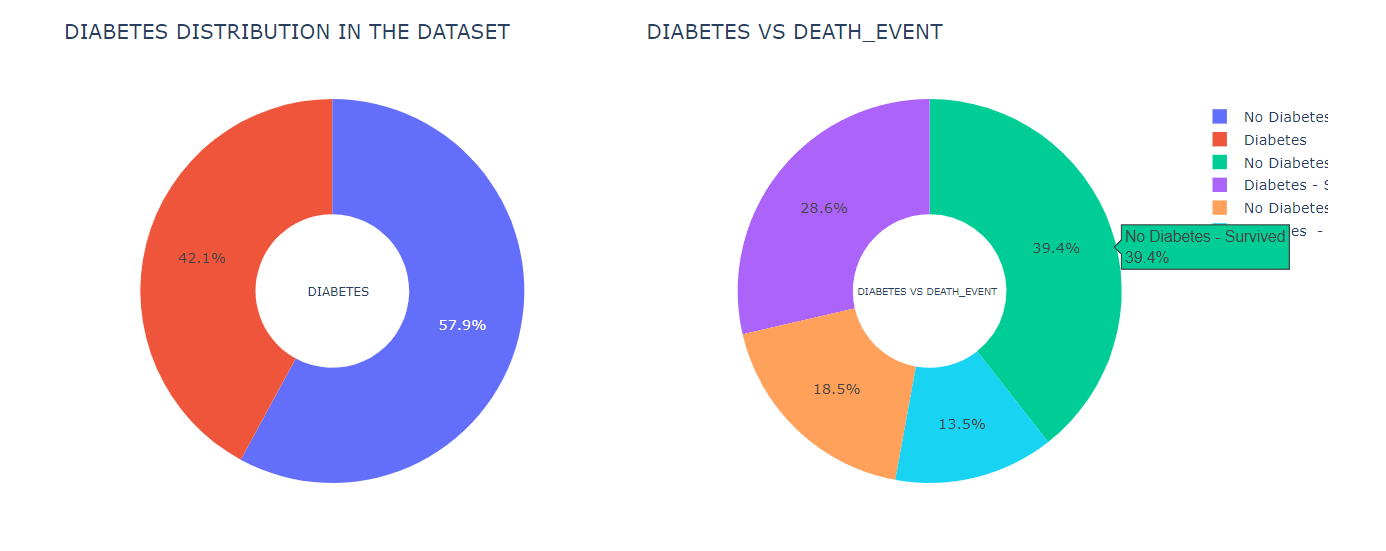
**3.2.6. Trực quan giới tính**



**Hình 26: Trực quan hóa về giới tính ảnh hưởng tới bệnh tim**

Theo như hình trên mọi người có thể quan sát được gằn giới tính cũng là một yếu tố ảnh hướng tới người mắc bệnh tim. Trong đó số người nam mắc bệnh tim là 65.3% và nữ là 34.7% dù rằng nam chiếm % mắc bệnh rất cao nhưng theo dữ liệu cho thấy rằng số bệnh nhân nam vẫn còn sống chiếm 44.4% trong khi đó nữ dù phần trăm mắc bệnh ít hơn nhưng có thể thấy được rằng số người nữ mắc bệnh vẫn còn sống chỉ chiếm 23.6%. Từ đó cho ta thấy dù rằng khả năng nam giới mắc bệnh tim mạch có thể cao hơn nhưng khả năng sinh tồn của phụ nữ lại thấp hơn.

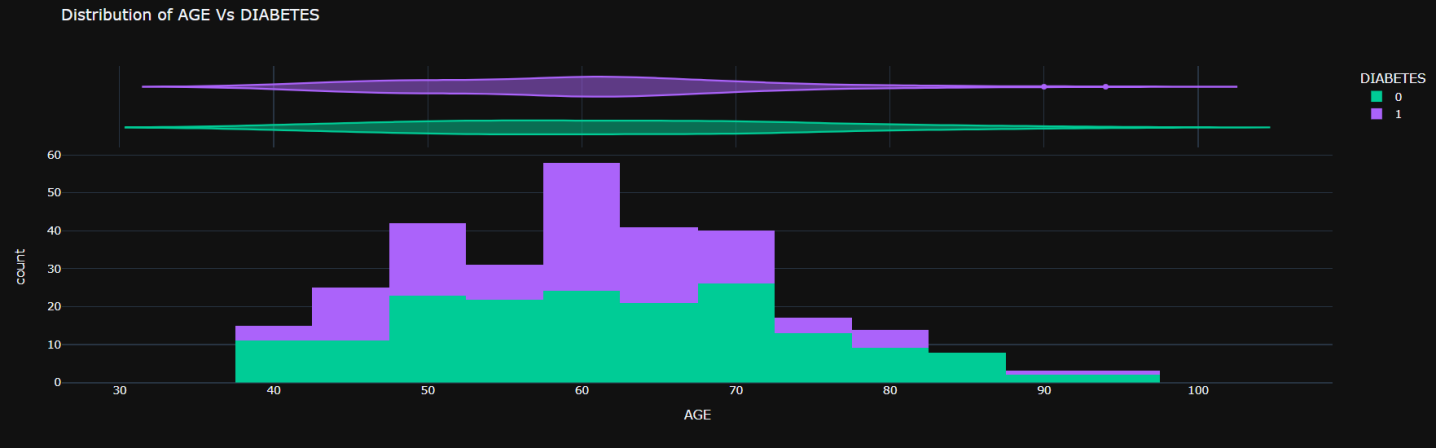
**3.2.7. Trực quan bệnh tiểu đường ảnh hưởng tới bệnh tim**



**Hình 27: Trực quan hóa về bệnh tiểu đường và bệnh tim**

Trong bộ dữ liệu này thì người không mắc bệnh tiểu đường là 57.9% và số người mắc bệnh là 42.1%. Trong đó số người không mắc bệnh tiểu đường mà mắc bệnh tim vẫn còn sống chiếm 39.4% và số người không mắc bệnh tiểu đường mà mắc bệnh tim tình trạng đã qua đời là chiếm 18.5%. Ngoài ra mắc bệnh tiểu đường và bệnh tim đã không qua khỏi là 13.5%. Từ đó có thể cho ta thấy rằng bệnh tiểu đường và bệnh tim mạch có mối quan tương quan rất chắc chẽ người bị bệnh tiểu đường sẽ có nguy cơ khó qua khỏi bệnh tim rất cao.

**3.2.8. Tương quan bệnh tiểu đường với tuổi tác**



**Hình 28: Trực quan hóa về bệnh tiểu đường và tuổi tác**

Có thể thấy rằng độ tuổi bị mắc bệnh tiểu đường tập trung từ 45 cho đến 70 từ đó ta có thể thấy được rằng mối tương quan của tuổi tác ảnh hưởng rất lớn đến các người mắc bệnh. Vì gần như đây là giai đoạn mà con người ta sẽ bắt đầu quan tâm nhiều hơn về sức khỏe của bản thân từ đó cũng có thể thấy được rằng người càng lớn tuổi sẽ dễ mắc các bệnh về tiểu đường hơn người trẻ.

# KẾT LUẬN

## Kết quả đạt được.

Nhờ mô hình dự đoán này mà mà chúng em có thể có được một số kết quả rút ra được từ dự án đó là giúp các cơ sở y tế tăng mức độ chính xác và khả năng dự đoán chó các bệnh nhân có nguy cơ mắc bệnh tim mạch được phát hiện sớm hơn từ đó giúp cho việc điều trị y tế trở nên thuận lợi hơn giúp cho bệnh nhân có thể sớm tiếp cận được việc chữa bệnh do đó sẽ tránh được các biến chứng về sau nhờ vậy nâng cao khả năng sống sót của bệnh nhân. Cũng như giúp phát triển hơn về y học vừa có thể vận dụng được ngành khoa học dữ liệu vào y tế ngoài ra cũng có thể từ những dữ liệu này mà có thể đưa ra cho mọi người thấy được rằng nếu muốn có thể vận dụng các kiến thức của ngành công nghệ thông tin vào y học để phát triển nên các thiết bị thúc đẩy y tế đưa ra các thông số giúp cho bệnh nhân có thể các nhìn nhận hơn về sức khỏe cho bản thân dù rằng có thể đó là các số liệu khô khan nhưng sau khi được trực quan hóa ta có thể giúp bệnh nhân hiểu thêm về tình hình của bản thân cũng như mô hình dự đoán này có thể dự đoán được xác suất mắc bệnh dựa tren thông tin cá nhân như tuổi tác, giới tính, … giúp cho bệnh nhân hiểu thêm về tình hình hiện tại từ đó giúp có các biện pháp can thiệp kiệp thời. Bên cạnh đó cũng giúp cho mọi người hiểu rõ hơn về mức độ nguy hiểm của bệnh tình từ đó đưa ra các biện pháp chuẩn đoán và phòng ngừa hiệu quả hơn.

Hiểu rõ hơn về mức độ ảnh hưởng của các yếu tố nguy cơ đối với nguy cơ mắc bệnh tim mạch. Mô hình có thể giúp xác định đặc trưng quan trọng nhất trong việc dự đoán nguy cơ, từ đó cung cấp thông tin hữu ích cho các nỗ lực can thiệp và phòng ngừa.

Đối với cộng đồng, việc áp dụng mô hình dự đoán có thể giúp cải thiện sức khỏe cộng đồng thông qua việc xác định và can thiệp sớm vào các trường hợp có nguy cơ cao mắc bệnh tim mạch, từ đó giảm thiểu tỷ lệ mắc bệnh và tăng cường chất lượng cuộc sống.

Những kết quả trên sẽ mang lại lợi ích rất lớn trong lĩnh vực y tế, giúp cải thiện dự đoán và quản lý bệnh tim mạch, từ đó cải thiện sức khỏe cá nhân và cộng đồng.

## Hạn chế và hướng phát triển .

**Hạn chế**

Bên cạnh những thuận lợi thì cũng có những hạn chế như: Vì thời gian làm dự án này khá ngắn nên bọn em không thể tự tổng hợp cũng như raw được bộ dữ liệu không ưng ý nên không thể crawl dự liệu như mong muốn dẫn đến bộ dữ liệu không đủ lớn và còn khá thiếu sót dẫn đến tập dữ liệu không đầy đủ và đa dạng từ đó dẫn đến việc mô hình không tổng quát hoá tốt trên các trường hợp mới hoặc không đưa ra dự đoán chính xác đó là những hạn chế về bộ dữ liệu này sau khi bọn em phân tích.

**Hướng phát triển**

Nếu có nhiều thời gian hơn bọn em sẽ mở rộng và cải thiện dữ liệu: Nghiên cứu có thể tìm kiếm và tích hợp thêm dữ liệu từ nhiều nguồn khác nhau để nó phong phú hơn và cải thiện khả năng dự đoán của mô hình.

Tinh chỉnh Mô hình: Cần tiếp tục tinh chỉnh các tham số của mô hình, thử nghiệm với nhiều cấu hình khác nhau để tìm ra mô hình tốt nhất và tránh overfitting hoặc underfitting.

Đó là những hạn chế cũng như hướng phát triển thêm mà nhóm em có thể nghĩ ra được rất mong được quý thầy cô xem xét và cho ý kiến thêm.

Em chân thành cảm ơn!

# TÀI LIỆU THAM KHẢO

[1] DATA MINING: CONCEPTS AND TECHNIQUES 3RD EDITION

[2] Nguyễn Thanh Tuấn(2009), Deeplearning cơ bản The Legrand Orange Book Template by Mathias Legrand is used : <https://nttuan8.com/sach-deep-learning-co-ban/>

[3] Lecturer in deep learning at the Faculty of Information Technology