**TRƯỜNG ĐẠI HỌC GIAO THÔNG VẬN TẢI**

**PHÂN HIỆU TẠI TP. HỒ CHÍ MINH**

**BỘ MÔN CÔNG NGHỆ THÔNG TIN**



**BÁO CÁO BÀI TẬP LỚN**

**HỌC PHẦN KHAI PHÁ DỮ LIỆU**

**ĐỀ TÀI :** **PHÂN TÍCH DỮ LIỆU Y TẾ VÀ**

**PHÁT TRIỂN MÔ HÌNH DỰ ĐOÁN BỆNH TIỂU ĐƯỜNG**

**TÍCH HỢP TRÊN WEBSITE**

Giảng viên hướng dẫn: NGUYỄN THIỆN DƯƠNG

Sinh viên thực hiện: ĐINH NGUYỄN HẢI ĐĂNG - 6351071019

PHẠM VĂN THUẬN - 6351071068

ĐẶNG QUANG VINH - 6351071079

NGUYỄN THANH HÒA – 6351071024

NGUYỄN PHI KHANH - 6351071032

Lớp : Công nghệ thông tin

Khoá :63

Tp. Hồ Chí Minh, năm 2025

**TRƯỜNG ĐẠI HỌC GIAO THÔNG VẬN TẢI**

**PHÂN HIỆU TẠI TP. HỒ CHÍ MINH**

**BỘ MÔN CÔNG NGHỆ THÔNG TIN**



**BÁO CÁO BÀI TẬP LỚN**

**HỌC PHẦN KHAI PHÁ DỮ LIỆU**

**ĐỀ TÀI :** **PHÂN TÍCH DỮ LIỆU Y TẾ VÀ**

**PHÁT TRIỂN MÔ HÌNH DỰ ĐOÁN BỆNH TIỂU ĐƯỜNG**

**TÍCH HỢP TRÊN WEBSITE**

Giảng viên hướng dẫn: NGUYỄN THIỆN DƯƠNG

Sinh viên thực hiện: ĐINH NGUYỄN HẢI ĐĂNG - 6351071019

PHẠM VĂN THUẬN - 6351071068

ĐẶNG QUANG VINH - 6351071079

NGUYỄN THANH HÒA - 6351071024

NGUYỄN PHI KHANH - 6351071032

Lớp : Công nghệ thông tin

Khoá :63

Tp. Hồ Chí Minh, năm 2025

|  |  |
| --- | --- |
| **TRƯỜNG ĐẠI HỌC GIAO THÔNG VẬN TẢI**  **PHÂN HIỆU TẠI THÀNH PHỐ HỒ CHÍ MINH** | **CỘNG HÒA XÃ HỘI CHỦ NGHĨA VIỆT NAM**  **Độc lập – Tự do – Hạnh phúc** |

**NHIỆM VỤ BÁO CÁO BÀI TẬP LỚN**

**BỘ MÔN:** **CÔNG NGHỆ THÔNG TIN**

-------\*\*\*-------

**Sinh viên thực hiện:**

|  |  |
| --- | --- |
| **Họ và Tên** | **Mã số sinh viên** |
| Đinh Nguyễn Hải Đăng | 6351071019 |
| Phạm Văn Thuận | 6351071068 |
| Đặng Quang Vinh | 6351071079 |
| Nguyễn Thanh Hòa | 6351071024 |
| Nguyễn Phi Khanh | 6351071032 |

**1. Tên đề tài**

* Phân tích dữ liệu y tế và phát triển mô hình dự đoán bệnh tiểu đường tích hợp trên website.

1. **Lý do chọn đề tài**

* Tiểu đường là một bệnh mãn tính làm tăng nguy cơ đột quỵ, suy thận, biến chứng thận, bệnh mạch máu ngoại biên, bệnh tim và tử vong. Liên đoàn Tiểu đường Quốc tế ước tính rằng đến năm 2045, với tốc độ tăng trưởng hiện tại, 693 triệu người sẽ mắc bệnh tiểu đường trên toàn thế giới. Theo Trung tâm Kiểm soát và Phòng ngừa Dịch bệnh (CDC), vào năm 2012, 29,1 triệu người ở Hoa Kỳ được chẩn đoán mắc bệnh tiểu đường, khiến đây trở thành nguyên nhân gây tử vong đứng thứ bảy tại quốc gia này. Bệnh tiểu đường gây gánh nặng tài chính lớn cho nền kinh tế Hoa Kỳ. Các nghiên cứu cho thấy tổng chi phí ước tính cho bệnh tiểu đường được chẩn đoán đã tăng lên 327 tỷ đô la vào năm 2017, bao gồm 237 tỷ đô la chi phí y tế trực tiếp và 90 tỷ đô la do giảm năng suất.
* Trong những năm gần đây, bệnh tiểu đường trở thành một trong những vấn đề sức khỏe nghiêm trọng trên toàn cầu. Theo nhiều thống kê y tế, số người mắc bệnh tiểu đường ngày càng tăng nhanh, đặc biệt tại các quốc gia đang phát triển như Việt Nam. Điều đáng lo ngại là phần lớn bệnh nhân chỉ phát hiện bệnh khi đã ở giai đoạn muộn, gây khó khăn cho quá trình điều trị và làm tăng nguy cơ biến chứng nguy hiểm.
* Song song với đó, sự phát triển mạnh mẽ của khoa học dữ liệu (Data Science) và trí tuệ nhân tạo (AI) mở ra nhiều cơ hội mới trong lĩnh vực y tế. Các mô hình học máy (Machine Learning) có khả năng phân tích dữ liệu sức khỏe và dự đoán nguy cơ mắc bệnh với độ chính xác cao, hỗ trợ phát hiện sớm và đưa ra lời khuyên phù hợp. Tuy nhiên, phần lớn các mô hình này chưa được tích hợp vào các nền tảng dễ sử dụng cho người dân và các cơ sở y tế.
* Nhận thấy tính cấp thiết trong việc hỗ trợ sàng lọc bệnh tiểu đường cũng như tiềm năng của các phương pháp học máy, nhóm quyết định chọn đề tài “Phân tích dữ liệu y tế và phát triển mô hình dự đoán bệnh tiểu đường tích hợp trên website”.
* Đề tài vừa mang ý nghĩa thực tiễn, vừa ứng dụng được các kiến thức về xử lý dữ liệu, thuật toán học máy và phát triển web. Việc xây dựng một hệ thống dự đoán trực tuyến giúp:
* Hỗ trợ người dùng tự đánh giá nguy cơ mắc bệnh tiểu đường một cách nhanh chóng.
* Tăng khả năng phát hiện sớm nhằm giảm thiểu biến chứng.
* Đưa công nghệ AI vào ứng dụng thực tế phục vụ cộng đồng.
* Giúp sinh viên rèn luyện kỹ năng phân tích dữ liệu, lập trình mô hình ML và phát triển website.

1. **Nội dung, phạm vi đề tài**

* Đề tài tập trung nghiên cứu và xây dựng một hệ thống dự đoán bệnh tiểu đường tích hợp trí tuệ nhân tạo, hướng đến hỗ trợ người dùng tự đánh giá nguy cơ mắc bệnh và nâng cao nhận thức về sức khỏe. Nội dung nghiên cứu bắt đầu từ việc khảo sát các lý thuyết liên quan đến học máy trong lĩnh vực y tế, bao gồm các thuật toán thường dùng trong phân loại như Logistic Regression, SVM...Đồng thời, đề tài tìm hiểu các kỹ thuật xử lý dữ liệu sức khỏe, phân tích các đặc trưng có ảnh hưởng đến bệnh tiểu đường như tuổi, BMI, huyết áp, sức khỏe tinh thần... nhằm phục vụ cho việc xây dựng mô hình dự đoán phù hợp và chính xác.
* Hệ thống tiến hành thu thập và xử lý bộ dữ liệu tiểu đường, làm sạch, chuẩn hóa và phân chia dữ liệu trước khi huấn luyện mô hình. Sau đó, các mô hình được xây dựng, thử nghiệm và đánh giá dựa trên các chỉ số như Accuracy, Precision, Recall, F1-score và AUC để lựa chọn mô hình có hiệu năng tốt nhất. Trên nền tảng mô hình AI đã tối ưu, nhóm tiếp tục phát triển website với giao diện đơn giản, trực quan, hỗ trợ người dùng nhập các chỉ số sức khỏe, xem kết quả dự đoán nguy cơ tiểu đường.
* Phạm vi đề tài tập trung vào nhóm người dùng phổ thông và giới hạn trong chức năng dự đoán nguy cơ tiểu đường dựa trên dữ liệu cơ bản từ bộ dữ liệu công khai. Hệ thống không đi sâu vào việc chẩn đoán y khoa chuyên sâu, không thu thập dữ liệu từ bệnh viện hay triển khai mô hình học sâu phức tạp. Cuối cùng, đề tài đánh giá hiệu quả mô hình dự đoán, khả năng hoạt động của website và tính tiện dụng của giao diện, từ đó đề xuất hướng phát triển trong tương lai như mở rộng dữ liệu, tăng độ chính xác mô hình và tích hợp thêm các chức năng theo dõi sức khỏe khác.

1. **Công nghệ, công cụ và ngôn ngữ lập trình**

* Trong quá trình phát triển hệ thống, nhóm sử dụng kết hợp nhiều ngôn ngữ lập trình, công nghệ và công cụ hiện đại nhằm đảm bảo tính ổn định, hiệu năng và khả năng mở rộng của ứng dụng. Ở phía backend, nhóm lựa chọn Python và framework FastAPI để xây dựng hệ thống RESTful API. FastAPI mang lại tốc độ xử lý cao, hỗ trợ bất đồng bộ (asynchronous I/O), đồng thời cung cấp khả năng tích hợp linh hoạt với các mô hình học máy, giúp tối ưu quá trình triển khai.
* Trong lĩnh vực học máy, nhóm áp dụng các thuật toán như Decision Tree, Random Forest, XGBoost, Logistic Regression, Support Vector Machine (SVM) và Neural Network để so sánh và lựa chọn mô hình có hiệu suất tốt nhất. Mô hình cuối cùng được sử dụng để dự đoán nguy cơ mắc bệnh tiểu đường dựa trên các chỉ số y tế quan trọng như BMI, sức khỏe tinh thần, sức khỏe thể chất, độ tuổi và các yếu tố liên quan khác.
* Phần frontend của hệ thống được xây dựng bằng ReactJS, kết hợp cơ chế kết xuất phía máy chủ (Server-side Rendering – SSR) nhằm cải thiện tốc độ tải trang và tối ưu SEO. Giao diện người dùng được thiết kế bằng Tailwind CSS, mang lại sự linh hoạt, nhất quán và tính thẩm mỹ hiện đại. Các thao tác và luồng tương tác với hệ thống được tối ưu hóa để đảm bảo trải nghiệm trực quan, dễ sử dụng.
* Tổng thể hệ thống được tổ chức theo kiến trúc microservices với ba thành phần chính: frontend (ReactJS) đảm nhiệm giao diện người dùng, backend (FastAPI) xử lý nghiệp vụ và giao tiếp với cơ sở dữ liệu, và cụm dịch vụ học máy vận hành độc lập phục vụ cho chức năng dự đoán và hỗ trợ tư vấn sức khỏe. Kiến trúc này không chỉ giúp hệ thống dễ dàng mở rộng và bảo trì, mà còn bảo đảm hiệu năng cao và khả năng tích hợp linh hoạt với các công nghệ mới trong tương lai.

1. **Các kết quả chính dự kiến sẽ đạt được:**

* Xây dựng thành công mô hình dự đoán nguy cơ mắc bệnh tiểu đường dựa trên các thuật toán học máy như Logistic Regression, Random Forest, XGBoost, SVM và Neural Network. Mô hình được huấn luyện, tinh chỉnh và đánh giá theo các chỉ số Accuracy, Precision, Recall, F1-score và AUC - ROC, từ đó lựa chọn ra mô hình có hiệu suất tốt nhất.
* Hoàn thiện một hệ thống API xử lý dự đoán được xây dựng bằng FastAPI, đảm bảo khả năng phản hồi nhanh, ổn định và tích hợp hiệu quả với mô hình học máy đã triển khai. API hỗ trợ truyền dữ liệu đầu vào, xử lý dự đoán và trả về kết quả theo chuẩn REST.
* Phát triển giao diện website trực quan, thân thiện, sử dụng ReactJS và Tailwind CSS, cho phép người dùng nhập các thông số sức khỏe, nhận kết quả dự đoán nguy cơ tiểu đường, xem thông tin bệnh và nhận một số khuyến nghị cơ bản về sức khỏe.
* Xây dựng một hệ thống hoàn chỉnh theo kiến trúc microservices, bao gồm frontend, backend và dịch vụ học máy độc lập. Hệ thống đảm bảo tính linh hoạt, dễ mở rộng, dễ bảo trì và có khả năng tích hợp thêm các chức năng khác trong tương lai.
* Tạo ra một công cụ hỗ trợ người dùng trong việc nhận biết sớm nguy cơ mắc bệnh tiểu đường, góp phần nâng cao ý thức chăm sóc sức khỏe cá nhân. Hệ thống có tính ứng dụng thực tiễn và có thể được mở rộng thành các sản phẩm hỗ trợ y tế khác.
* Hoàn thiện báo cáo nghiên cứu đầy đủ, bao gồm khảo sát lý thuyết, phân tích dữ liệu, quy trình xây dựng mô hình, kiến trúc hệ thống, kết quả thử nghiệm và đánh giá hiệu năng, cùng các đề xuất hướng phát triển trong thời gian tới.

1. **Giảng viên và cán bộ hướng dẫn:**

Họ và tên: Nguyễn Thiện Dương

Đơn vị công tác: Bộ môn CNTT

Điện thoại: 0777207700

Email: [ntduong@utc2.edu.vn](mailto:ntduong@utc2.edu.vn)

1. **Đã nhận nhiệm vụ**

Sinh viên: Đinh Nguyễn Hải Đăng

Phạm Văn Thuận

Đặng Quang Vinh

Nguyễn Thanh Hòa

Nguyễn Phi Khanh

Điện thoại:

Email: 6351071019@st.utc2.edu.vn

|  |  |
| --- | --- |
| **Ngày 13 tháng 12 năm 2025**  **Trưởng BM Công nghệ Thông tin**  **ThS. Trần Phong Nhã** | **Đã giao nhiệm vụ TKTN**  **Giảng viên hướng dẫn**  **ThS. Nguyễn Thiện Dương** |

# 

# LỜI CẢM ƠN

Trong quá trình thực hiện và hoàn thành bài báo cáo này, nhóm chúng em xin gửi lời cảm ơn sâu sắc đến thầy **Nguyễn Thiện Dương**. Thầy không chỉ tận tình giảng dạy mà còn luôn sẵn sàng hỗ trợ, giải đáp những thắc mắc của chúng em trong suốt thời gian học tập và nghiên cứu. Những kiến thức, kinh nghiệm và định hướng mà thầy truyền đạt đã trở thành nền tảng quan trọng giúp chúng em hiểu rõ vấn đề, phát triển tư duy và hoàn thành bài báo cáo một cách đầy đủ và khoa học hơn.

Chúng em thật sự trân trọng sự tận tâm của thầy trong từng buổi học, từng góp ý và từng hướng dẫn chi tiết. Chính sự nhiệt huyết và tinh thần trách nhiệm ấy đã tạo động lực để chúng em cố gắng hơn, chủ động học hỏi và không ngừng hoàn thiện bản thân.

Một lần nữa, nhóm chúng em xin chân thành cảm ơn thầy Nguyễn Thiện Dương và kính chúc thầy luôn mạnh khỏe, nhiều niềm vui, tiếp tục thành công trong sự nghiệp giảng dạy và truyền cảm hứng cho các thế hệ sinh viên tiếp theo.

Chúng em xin trân trọng cảm ơn!

# 

# NHẬN XÉT CỦA GIẢNG VIÊN HƯỚNG DẪN

…………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………

…………………………………………………………………………………………

…………………………………………………………………………………………

…………………………………………………………………………………………

…………………………………………………………………………………………

|  |
| --- |
| ***Tp. Hồ Chí Minh, ngày 13 tháng 12 năm 2025***  **Giáo viên hướng dẫn**  **Nguyễn Thiện Dương** |

**MỤC LỤC**

[**NHIỆM VỤ BÁO CÁO BÀI TẬP LỚN i**](#_Toc216474243)

[**LỜI CẢM ƠN vi**](#_Toc216474244)

[**NHẬN XÉT CỦA GIẢNG VIÊN HƯỚNG DẪN vii**](#_Toc216474245)

[**DANH MỤC CHỮ VIẾT TẮT x**](#_Toc216474246)

[**MỤC LỤC HÌNH ẢNH xi**](#_Toc216474247)

[**MỤC LỤC BẢNG BIỂU xiii**](#_Toc216474248)

[**BẢNG PHÂN CÔNG CÔNG VIỆC xiv**](#_Toc216474249)

[**CHƯƠNG 1. GIỚI THIỆU ĐỀ TÀI 1**](#_Toc216474250)

[1.1. Bối cảnh và lý do xây dựng hệ thống 1](#_Toc216474251)

[1.2. Mục tiêu của đề tài 1](#_Toc216474252)

[1.3. Phạm vi đề tài 1](#_Toc216474253)

[1.4. Phương pháp thực hiện 2](#_Toc216474254)

[**CHƯƠNG 2. CƠ SỞ LÝ THUYẾT 3**](#_Toc216474255)

[2.1. Lý thuyết học máy - Machine Learning 3](#_Toc216474256)

[2.1.1. Khái niệm về Học máy 3](#_Toc216474257)

[2.1.2. Các bước xây dựng mô hình 3](#_Toc216474258)

[2.1.3. Các nhóm phương pháp trong học máy 4](#_Toc216474259)

[2.1.4. Các thuật toán phân loại phổ biến 6](#_Toc216474260)

[2.1.5. Lý thuyết về độ đánh giá mô hình 7](#_Toc216474261)

[2.2. Tổng quan về công nghệ sử dụng 9](#_Toc216474262)

[2.2.1. FASP Protocol là gì? 9](#_Toc216474263)

[2.2.2. Các thành phần chính của FASP API 10](#_Toc216474264)

[2.2.3. Bảo mật và Troubleshooting 12](#_Toc216474265)

[**CHƯƠNG 3. DATASET VÀ TIỀN XỬ LÝ DỮ LIỆU 13**](#_Toc216474266)

[3.1. Nguồn dữ liệu 13](#_Toc216474267)

[3.1.1. Dữ liệu công khai 13](#_Toc216474268)

[3.1.2. Phân tích và lựa chọn tập dữ liệu phù hợp 13](#_Toc216474269)

[3.1.3. Hướng dẫn tải dataset 14](#_Toc216474270)

[3.2. Mô tả dữ liệu 16](#_Toc216474271)

[3.2.1. Chi tiết các biến trong bộ dữ liệu 16](#_Toc216474272)

[3.2.2. Làm sạch dữ liệu 18](#_Toc216474273)

[3.2.2.1. Xử lý dữ liệu bị thiếu 18](#_Toc216474274)

[3.2.2.2. Kiểm tra giá trị độc nhất 19](#_Toc216474275)

[3.2.2.3. Kiểm tra và loại bỏ giá trị trùng lặp 20](#_Toc216474276)

[3.2.2.4. Xác định Outliers 21](#_Toc216474277)

[3.2.3. Giải thích sự xuất hiện của các giá trị ngoại lai 22](#_Toc216474278)

[3.2.3.1. Phân tích tổng quan dữ liệu của 3 biến BMI, MentHlth, PhysHlth 22](#_Toc216474279)

[3.2.3.2. Xác định các giá trị hợp lý 23](#_Toc216474280)

[3.2.3.3. Xác định Outliers theo phương pháp IQR 25](#_Toc216474281)

[3.2.3.4. Phân tích dữ liệu khám phá 27](#_Toc216474282)

[3.2.3. Trực quan hóa dữ liệu 28](#_Toc216474283)

[3.2.3.1. Mối quan hệ giữa từng biến độc lập chứa ngoại lai và GenHlth tới biến phụ thuộc Diabetes\_binary 28](#_Toc216474284)

[3.2.3.2. Mối quan hệ giữa nhóm biến độc lập và biến phụ thuộc Diabetes\_binary 31](#_Toc216474285)

[3.3. Xử lý giá trị ngoại lai 37](#_Toc216474286)

[3.4. Chọn các biến quan trọng 39](#_Toc216474287)

[3.5. Chia tập train và test 42](#_Toc216474288)

[**CHƯƠNG 4. XÂY DỰNG MÔ HÌNH 43**](#_Toc216474289)

[4.1. Sensitive Models 43](#_Toc216474290)

[4.2. Robust Models 45](#_Toc216474291)

[4.3. So sánh và lựa chọn mô hình tốt nhất 47](#_Toc216474292)

[**CHƯƠNG 5. KẾT QUẢ VÀ KIẾN NGHỊ 49**](#_Toc216474293)

[5.1. Kết quả đạt được 49](#_Toc216474294)

[5.1.1. Hoàn thiện mô hình dự đoán bệnh tiểu đường 49](#_Toc216474295)

[5.1.2. Xây dựng giao diện người dùng (UX/UI) hoàn chỉnh và trực quan 49](#_Toc216474296)

[5.1.3. Triển khai hệ thống backend ổn định và đảm bảo tính mở rộng 49](#_Toc216474297)

[5.2. Đánh giá chung 49](#_Toc216474298)

[5.3. Kiến nghị phát triển 49](#_Toc216474299)

[5.4. Kết luận 50](#_Toc216474300)

[**TÀI LIỆU THAM KHẢO 51**](#_Toc216474301)

# DANH MỤC CHỮ VIẾT TẮT

|  |  |  |  |
| --- | --- | --- | --- |
| **STT** | **Mô tả** | **Ý nghĩa đầy đủ** | **Ghi chú** |
| 1 | AI | Artificial Intelligence | Trí tuệ nhân tạo |
| 2 | ML | Machine Learning | Học máy |
| 3 | SVM | Support Vector Machine | Thuật toán máy vectơ hỗ trợ |
| 4 | UI/UX | User Interface / User Experience | Giao diện & trải nghiệm người dùng |
| 5 | API | Application Programming Interface | Giao diện lập trình ứng dụng |
| 6 | FASP | Fast, Adaptive, Secure, Protocol | Xuất hiện trong mục 1.4 (FASP API) |
| 7 | MLP | Multi-Layer Perceptron | Mạng nơ-ron truyền thẳng |
| 8 | CNN | Convolutional Neural Network | Mạng nơ-ron tích chập |
| 9 | RNN | Recurrent Neural Network | Mạng nơ-ron hồi quy |

# 

# MỤC LỤC HÌNH ẢNH

[Hình 2.1: Các bước xây dựng mô hình. 4](#_Toc216450371)

[Hình 2.2: Công thức tính accuracy 7](#_Toc216450372)

[Hình 2.3: Công thức tính precision 7](#_Toc216450373)

[Hình 2.4: Công thức tính recall 7](#_Toc216450374)

[Hình 2.5: Công thức tính F1-score 7](#_Toc216450375)

[Hình 2.6: Công thức tính TPR và FPR 8](#_Toc216450376)

[Hình 2.7: Đường cong ROC 8](#_Toc216450377)

[Hình 2.8: Chỉ số đánh giá phân loại ROC – AUC 9](#_Toc216450378)

.

[Hình 3.1: Nhấp vào nút “Code”. 14](#_Toc216450379)

[Hình 3.2: Cửa sổ hiển thị đoạn mã dùng để tải dataset. 15](#_Toc216450380)

[Hình 3.3: Nơi điền tên file. 15](#_Toc216450381)

[Hình 3.4: Kết quả kiểm tra dữ liệu có dữ liệu thiếu hay không 18](#_Toc216450382)

[Hình 3.5: Kết quả kiểm tra giá trị độc nhất 19](#_Toc216450383)

[Hình 3.6: Kết quả kiểm tra giá trị trùng lặp 20](#_Toc216450384)

[Hình 3.7: Thông tin dữ liệu sau khi xóa dữ liệu trùng lặp 20](#_Toc216450385)

[Hình 3.8: Biểu đồ boxplot kiểm tra Outliers 21](#_Toc216450386)

[Hình 3.9: Kết quả phân tích tổng quan dữ liệu của 3 biến MBI, MentHlth, PhysHlth 22](#_Toc216450387)

[Hình 3.10: Kết quả các giá trị hợp lý 23](#_Toc216450388)

[Hình 3.11: Kết quả xác định Outliers 25](#_Toc216450389)

[Hình 3.12: Ma trận tương quan của các biến 27](#_Toc216450390)

[Hình 3.13: Biểu đồ cột thể hiện mối quan hệ của BMI và Diabetes\_binary 28](#_Toc216450391)

[Hình 3.14: Biểu đồ thể hiện mối quan hệ của MentHlth và Diabetes\_binary 28](#_Toc216450392)

[Hình 3.15: Biểu đồ cột chồng thể hiện mối quan hệ của PhysHlth và Diabetes\_binary 29](#_Toc216450393)

[Hình 3.16: Biểu đồ đường thể hiện mối quan hệ của PhysHlth và Diabetes\_binary 29](#_Toc216450394)

[Hình 3.17: Biểu đồ cột chồng thể hiện mối quan hệ của GenHlth và Diabetes\_binary 30](#_Toc216450395)

[Hình 3.18: Biểu đồ cột thể hiện sự phân bố số lượng người mắc bệnh tiểu đường theo mức độ đánh giá sức khỏe tổng quát (GenHlth) 30](#_Toc216450396)

[Hình 3.19: Biểu đồ cột đôi thể hiện tỷ lệ mắc bệnh tiểu đường theo hai yếu tố 31](#_Toc216450397)

[Hình 3.20: Biểu đồ thể hiện mối quan hệ của Smoker và HvyAlcoholConsump với Diabetes\_binary 32](#_Toc216450398)

[Hình 3.21: Biểu đồ thể hiện mối quan hệ của Stroke và HeartDiseaseorAttack với Diabetes\_binary 32](#_Toc216450399)

[Hình 3.22: Biểu đồ cột đôi thể hiện mối quan hệ của PhysActivity, Fruits và Veggies với Diabetes\_binary 33](#_Toc216450400)

[Hình 3.23: Biểu đồ thể hiện mối quan hệ của NoDocbcCost và AnyHealthcare với Diabetes\_binary 34](#_Toc216450401)

[Hình 3.24: Biểu đồ tổng hợp 4 phần thể hiện tỷ lệ mắc bệnh tiểu đường theo các yếu tố nhân khẩu học (Sex, Age, Education và Income) 35](#_Toc216450402)

[Hình 3.25: Biểu đồ cột đôi thể hiện mối quan hệ của DiffWalk và CholCheck với Diabetes\_binary 36](#_Toc216450403)

[Hình 3.26: Thông tin dữ liệu df1 sau khi xử lý giá trị ngoại lai bằng phương pháp 2 39](#_Toc216450404)

[Hình 3.27: Thông tin dữ liệu df2 sau khi xử lý giá trị ngoại lai bằng phương pháp 2 39](#_Toc216450405)

[Hình 3.28: Kết quả chọn lọc thuộc tính của df1 40](#_Toc216450406)

[Hình 3.29: Kết quả chọn lọc thuộc tính của df2 40](#_Toc216450407)

[Hình 3.30: Thông tin dữ liệu của df1 sau khi chọn lọc 15 thuộc tính có Score cao nhất 41](#_Toc216450408)

[Hình 3.31: Thông tin dữ liệu của df2 sau khi chọn lọc 15 thuộc tính có Score cao nhất 41](#_Toc216450409)

[Hình 3.32: Thông tin chia tập train - test theo tỷ lệ 80 – 20 42](#_Toc216450410)

.

[Hình 4.1: Quá trình thực hiện train 3 mô hình Logistic Regression, Support Vector Machine (SVM) và Multi-Layer Perceptron (MLP) 44](#_Toc216450411)

[Hình 4.2: Tóm tắt kết quả đánh giá mô hình trên df2 45](#_Toc216450412)

[Hình 4.3: Quá trình thực hiện train 3 mô hình Decision Tree, Random Forest và XGBoost 46](#_Toc216450413)

[Hình 4.4: Tóm tắt kết quả đánh giá mô hình trên df1 47](#_Toc216450414)

# MỤC LỤC BẢNG BIỂU

[Bảng 3.1: Bảng biến nhị phân 16](#_Toc216450353)

[Bảng 3.2: Bảng biến thứ bậc 17](#_Toc216450354)

[Bảng 3.3: Bảng biến liên tục 17](#_Toc216450355)

[Bảng 3.4: Bảng so sánh kết quả 3 phương pháp xử lý giá trị ngoại lai 38](#_Toc216450356)

[Bảng 3.5: Bảng đánh giá 3 phương pháp xử lý giá trị ngoại lai 38](#_Toc216450357)

.

[Bảng 4.1: Bảng lý do chọn 3 mô hình Logistic Regression, Support Vector Machine (SVM) và Multi-Layer Perceptron (MLP) 43](#_Toc216450361)

[Bảng 4.2: Bảng lý do chọn 3 mô hình Decision Tree, Random Forest và XGBoost 45](#_Toc216450362)

[Bảng 4.3: Bảng đánh giá các mô hình 47](#_Toc216450363)

# BẢNG PHÂN CÔNG CÔNG VIỆC

|  |  |  |
| --- | --- | --- |
| **Tên thành viên** | **Phân công công việc** | **Đánh giá** |
| Đinh Nguyễn Hải Đăng | * Mức độ nâng cao * Xây dựng được Chương trình ứng dụng demo sử dụng kết quả đạt được |  |
| Phạm Văn Thuận | * Phát biểu kết quả * So sánh, đánh giá |  |
| Đặng Quang Vinh | * Kỹ thuật tiền xử lý dữ liệu được lựa chọn * Thuật toán khai thác dữ liệu được lựa chọn * Kết quả đạt được |  |
| Nguyễn Thanh Hoà | * Dữ liệu được chọn * Mô tả bài toán |  |
| Nguyễn Phi Khanh | * Kết luận * Ưu điểm * Hạn chế * Hướng phát triển |  |

# CHƯƠNG 1. GIỚI THIỆU ĐỀ TÀI

## 1.1. Bối cảnh và lý do xây dựng hệ thống

* Bệnh tiểu đường là một trong những bệnh mãn tính phổ biến nhất hiện nay và có xu hướng gia tăng nhanh chóng trên toàn thế giới. Việc phát hiện bệnh muộn thường dẫn đến nhiều biến chứng nguy hiểm như suy thận, mù lòa, đột quỵ và bệnh tim mạch. Do đó, nhu cầu dự đoán sớm nguy cơ mắc bệnh để hỗ trợ người dân theo dõi sức khỏe là vô cùng cần thiết.
* Bên cạnh đó, với sự phát triển mạnh mẽ của công nghệ dữ liệu và trí tuệ nhân tạo, việc áp dụng các mô hình học máy vào lĩnh vực y tế đã và đang mang lại hiệu quả cao trong công tác dự đoán và hỗ trợ chẩn đoán. Vì vậy, nhóm quyết định lựa chọn đề tài “Phân tích dữ liệu y tế và phát triển mô hình dự đoán bệnh tiểu đường tích hợp trên website” nhằm ứng dụng kiến thức đã học, đồng thời xây dựng một hệ thống có tính thực tiễn, dễ tiếp cận và hữu ích cho cộng đồng.

## 1.2. Mục tiêu của đề tài

* Phân tích và làm sạch bộ dữ liệu y tế liên quan đến bệnh tiểu đường.
* Khảo sát và lựa chọn các thuật toán học máy phù hợp để xây dựng mô hình dự đoán.
* Huấn luyện và đánh giá mô hình dựa trên các chỉ số hiệu suất (accuracy, precision, recall, F1-score…).
* Xây dựng website tích hợp mô hình dự đoán, cho phép người dùng nhập thông tin sức khỏe và nhận kết quả đánh giá nguy cơ.
* Tạo ra một hệ thống đơn giản, dễ sử dụng, góp phần giúp người dùng nâng cao nhận thức về sức khỏe.

## 1.3. Phạm vi đề tài

* Phạm vi dữ liệu: sử dụng các bộ dữ liệu y tế công khai có liên quan đến bệnh tiểu đường.
* Phạm vi kỹ thuật: đề tài tập trung vào các thuật toán học máy cơ bản như Logistic Regression, Decision Tree, Random Forest, SVM, Neural Netword, XGBoost.
* Phạm vi ứng dụng: xây dựng website giao diện đơn giản, hỗ trợ nhập dữ liệu sức khỏe và trả về kết quả dự đoán. Đề tài không thay thế vai trò chẩn đoán của bác sĩ mà chỉ mang tính hỗ trợ tham khảo.
* Phạm vi thời gian: thực hiện trong khuôn khổ thời gian của môn học nên mức độ chức năng sẽ ở mức cơ bản, phù hợp năng lực và thời lượng.

## 1.4. Phương pháp thực hiện

* Phương pháp thu thập và khảo sát tài liệu: tìm hiểu các bộ dữ liệu y tế liên quan đến tiểu đường, tài liệu về học máy và các nghiên cứu trước đó.
* Phương pháp phân tích dữ liệu: tiến hành làm sạch, xử lý thiếu dữ liệu, chuẩn hóa và trực quan hóa dữ liệu để hiểu rõ đặc điểm bộ dữ liệu.
* Phương pháp mô hình hóa: lựa chọn và xây dựng các thuật toán học máy, sau đó huấn luyện và tối ưu mô hình.
* Phương pháp đánh giá: sử dụng các chỉ số thống kê để so sánh hiệu quả và lựa chọn mô hình phù hợp nhất.
* Phương pháp triển khai: tích hợp mô hình đã chọn lên website bằng công nghệ phù hợp (Python/Flask hoặc Node.js), tạo giao diện người dùng và chức năng dự đoán.
* Phương pháp thử nghiệm: kiểm tra giao diện, thử nghiệm nhận dữ liệu đầu vào và đánh giá kết quả dự đoán.

# 

# CHƯƠNG 2. CƠ SỞ LÝ THUYẾT

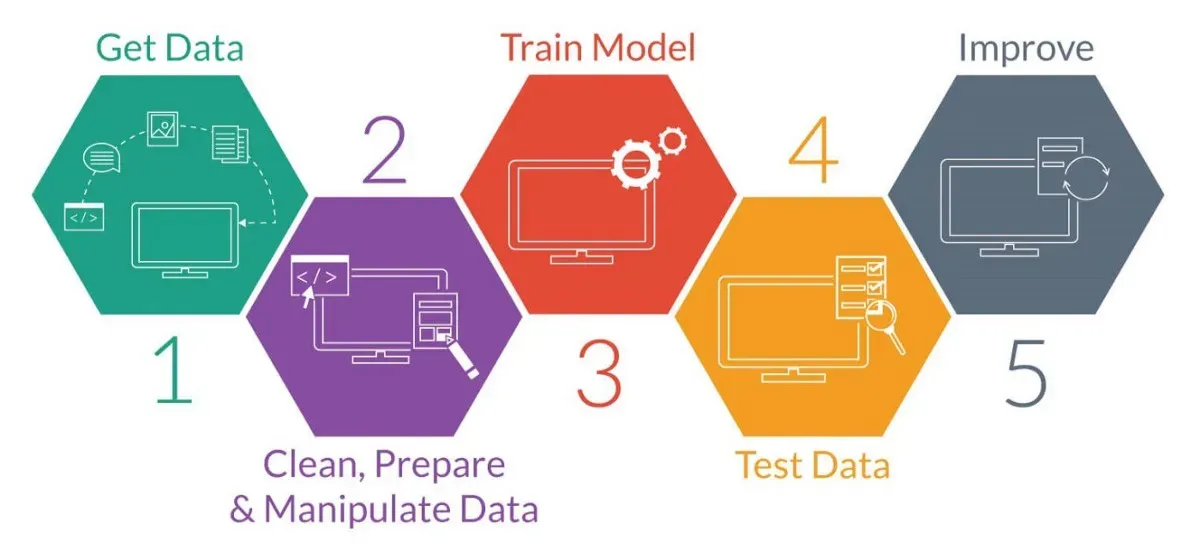
## 2.1. Lý thuyết học máy - Machine Learning

**2.1.1. Khái niệm về Học máy**

* Machine Learning hay còn gọi được là học máy, một lĩnh vực của trí tuệ nhân tạo ([AI](https://vietnix.vn/ai-la-gi/)) và [khoa học máy tính](https://vietnix.vn/khoa-hoc-may-tinh-la-gi/). Machine Learning liên quan đến việc nghiên cứu và xây dựng các kỹ thuật cho phép các hệ thống “học” tự động từ dữ liệu để giải quyết những vấn đề cụ thể. Machine Learning tập trung vào việc sử dụng dữ liệu và thuật toán để bắt chước cách con người học, sau đó dần dần cải thiện độ chính xác.
* IBM có một lịch sử phong phú về Machine Learning. Một trong đó là Arthur Samuel, được cho là đã đặt ra thuật ngữ “Machine Learning” với nghiên cứu xung quanh trò chơi cờ caro. Robert Nealey, tự xưng là bậc thầy cờ caro, đã chơi trò chơi này trên máy tính IBM 7094 vào năm 1962, và ông đã thua chiếc máy tính này.
* Machine Learning hoạt động dựa trên nhiều thuật toán, tuy vậy công nghệ này được chia làm 02 dạng bài toán chính, gồm: phân loại (classification) và dự đoán (prediction). Trong đó, các bài toán dự đoán có thể là giá xe, giá nhà đất…và những bài phân loại gồm: nhận diện đồ vật, [màu](https://vietnix.vn/bang-ma-mau/" \t "_blank" \o "Bảng code màu đầy đủ – Bảng mã màu HTML, CSS, RGB, CMYK chuẩn) sắc, chữ viết…
* Có thể nói, Machine Learning tuy chỉ là một nhánh trong trí tuệ nhân tạo (AI) nhưng dần trở thành một phần không thể thiếu trong sự phát triển của lĩnh vực khoa học nhân tạo. Bên cạnh đó, Machine Learning vận dụng những thuật toán, thống kê và phân tích dữ liệu nhằm cung cấp các thông tin chi tiết và tạo tác động đến chỉ số tăng trưởng chính của doanh nghiệp.

### 2.1.2. Các bước xây dựng mô hình

* Quy trình làm việc đồng nhất của Machine Learning.
* Bước 1: Data Collection hay Gathering Data: Thu thập dữ liệu được xem là bước quan trọng nhất trong cả quy trình. Thông thường, bước này sẽ chiếm khoảng 70% đến 80% thời gian cho toàn bộ quá trình vận hành. Vì tệp dữ liệu này sẽ ảnh hưởng rất lớn đến mô hình máy học của doanh nghiệp. Do đó, bạn cần nghiên cứu dữ liệu một cách cẩn thận, sau đó chắt lọc và chọn lựa những dữ liệu phù hợp nhất.
* Bước 2: Data Processing: Data processing hay còn được gọi là xử lý dữ liệu sơ khởi. Công việc này tương đương với bước thu thập dữ liệu. Trong quá trình vận hành, bước xử lý dữ liệu sẽ thực hiện các công việc loại bỏ thuộc tính dư thừa, mã hóa và thu gọn dữ liệu…
* Bước 3: Training model: Training model được hiểu là huấn luyện mô hình. Ở giai đoạn này, bạn sẽ để cho mô hình học trên các tệp dữ liệu mà bạn đã thu thập và chắt lọc được.
* Bước 4: Evaluating model: Evaluating model có nghĩa là đánh giá mô hình. Công việc này sẽ được thực hiện sau khi mô hình đã học xong tệp dữ liệu và bắt đầu cho kết quả. Dựa vào kết quả mà Machine Learning Workflow cung cấp, doanh nghiệp sẽ đánh giá được chất lượng cả tệp dữ liệu và hiệu suất hoạt động của mô hình này. Trên thực tế, một mô hình hoàn thiện sẽ cho kết quả đạt từ 80% trở lên.
* Bước 5: Improve: Cái thiện mô hình sẽ là giai đoạn cuối cùng cần thực hiện. Vì sau khi chọn được các mô hình phù hợp, số còn lại sẽ thuộc nhóm chưa phù hợp và cần được huấn luyện lại sao cho đạt được kỳ vọng mà doanh nghiệp đã đề ra. Theo nghiên cứu, bước 3,4 và 5 sẽ chiếm khoảng 30% thời gian của toàn quy trình.



Hình 2.1: Các bước xây dựng mô hình.

**2.1.3. Các nhóm phương pháp trong học máy**

* Trong lĩnh vực học máy (Machine Learning), các thuật toán được phân chia thành ba nhóm chính dựa trên cách thức mô hình học hỏi và rút trích thông tin từ dữ liệu: học có giám sát (Supervised Learning), học không giám sát (Unsupervised Learning) và học tăng cường (Reinforcement Learning). Việc phân loại này giúp người nghiên cứu lựa chọn phương pháp phù hợp với đặc điểm dữ liệu cũng như mục tiêu bài toán.

1. **Học có giám sát (Supervised Learning)**

* Học có giám sát là nhóm thuật toán được sử dụng phổ biến nhất và cũng là phương pháp chính được áp dụng trong hệ thống gợi ý phim của các nền tảng web xem phim trực tuyến. Trong học có giám sát, mô hình được huấn luyện trên tập dữ liệu đã được gán nhãn (labels). Mỗi mẫu dữ liệu gồm hai phần: tập đặc trưng đầu vào (features) như thể loại, diễn viên, thời lượng, lượt xem, và nhãn đầu ra (output) như mức độ yêu thích hoặc xếp hạng của người dùng. Mục tiêu của mô hình là học mối quan hệ giữa hai phần này để có thể dự đoán danh sách phim phù hợp cho người dùng mới.
* Supervised Learning được chia thành hai dạng nhiệm vụ chính:
* Phân loại (Classification): đầu ra là giá trị rời rạc, biểu thị một nhóm hoặc một lớp của đối tượng. Ví dụ: phân loại bộ phim vào các nhóm “phim có khả năng người dùng thích”, “phim không phù hợp”, hoặc dự đoán thể loại mà người dùng quan tâm dựa trên lịch sử xem.
* Hồi quy (Regression): đầu ra là giá trị liên tục như dự đoán điểm đánh giá phim của người dùng, thời gian xem dự kiến hoặc xác suất người dùng sẽ nhấn xem một bộ phim.
* Trong đề tài này, mô hình Random Forest đóng vai trò là thuật toán chính, sử dụng các đặc trưng của phim và hành vi người dùng để dự đoán mức độ phù hợp của mỗi bộ phim, từ đó hỗ trợ hệ thống gợi ý phim hoạt động hiệu quả và chính xác hơn.

1. **Học không giám sát (Unsupervised Learning)**

* Khác với học có giám sát, dữ liệu trong học không giám sát không có nhãn. Mục tiêu của mô hình là tự tìm kiếm cấu trúc ẩn, phân nhóm tự nhiên hoặc các mẫu đặc trưng trong dữ liệu. Đây là phương pháp thường được sử dụng khi người nghiên cứu không biết trước dữ liệu được chia thành bao nhiêu nhóm hoặc không có thông tin đầu ra cụ thể.
* Một số ứng dụng tiêu biểu của học không giám sát gồm:
  + Phân cụm (Clustering): nhóm các đối tượng có đặc điểm tương đồng lại với nhau. Các thuật toán phổ biến gồm K-means, Hierarchical Clustering và DBSCAN.
  + Giảm chiều dữ liệu (Dimensionality Reduction): nén số lượng đặc trưng để dễ quan sát, dễ mô hình hóa hoặc giảm nhiễu, ví dụ PCA (Principal Component Analysis).
  + Phát hiện bất thường (Anomaly Detection): tìm kiếm các điểm dữ liệu bất thường so với phần lớn tập dữ liệu.
* Mặc dù không được sử dụng trực tiếp trong đề tài, nhưng học không giám sát thường hỗ trợ việc chuẩn hóa dữ liệu hoặc phân tích đặc trưng, do đó việc trình bày nhóm thuật toán này giúp hoàn thiện cơ sở lý thuyết.

1. **Học tăng cường (Reinforcement Learning – RL)**

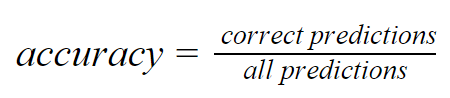
* Học tăng cường là một nhánh đặc biệt trong học máy, trong đó một "tác nhân" (agent) học cách đưa ra hành động tối ưu thông qua quá trình tương tác liên tục với môi trường. Mỗi hành động được đánh giá bằng một tín hiệu phần thưởng (reward), và mục tiêu của agent là tối đa hóa phần thưởng tích lũy theo thời gian.
* Reinforcement Learning thường được ứng dụng trong robot tự hành, điều khiển hệ thống tự động, tối ưu hóa trong công nghiệp, hay các trò chơi nổi tiếng như AlphaGo. Một số thuật toán tiêu biểu bao gồm Q-learning, SARSA và Deep Q-Network (DQN).

**2.1.4. Các thuật toán phân loại phổ biến**

* Trong lĩnh vực học máy có giám sát (supervised learning), bài toán phân loại được sử dụng khi biến mục tiêu (target/output) là rời rạc,… Các thuật toán phân loại đã được phát triển rất phong phú, từ các mô hình tuyến tính đơn giản cho tới các mô hình phức tạp có khả năng học sâu hoặc tạo tập hợp mô hình (ensemble). Một số thuật toán phổ biến thường được áp dụng trong các nghiên cứu và ứng dụng thực tiễn bao gồm:
* Logistic Regression: Thuật toán tuyến tính dùng hàm sigmoid để dự đoán xác suất thuộc lớp. Đơn giản, dễ triển khai và phù hợp với dữ liệu tuyến tính.
* k-Nearest Neighbors (k-NN): Phân loại dựa trên k hàng xóm gần nhất. Không cần huấn luyện nhưng tốn thời gian khi dự đoán.
* Decision Tree: Xây dựng cây quyết định bằng cách phân tách dữ liệu theo thuộc tính. Dễ hiểu, nhưng dễ overfitting nếu không kiểm soát.
* Random Forest: Tập hợp nhiều cây quyết định để tăng độ chính xác và giảm nhiễu. Ổn định và hiệu quả trên nhiều loại dữ liệu.
* Support Vector Machine (SVM): Tìm siêu phẳng tối ưu để tách lớp. Hoạt động tốt với dữ liệu có biên tách rõ ràng, hỗ trợ kernel cho bài toán phi tuyến.
* Naive Bayes: Dựa trên định lý Bayes và giả định độc lập giữa các thuộc tính. Hiệu quả rất tốt cho dữ liệu văn bản và phân loại spam.
* Gradient Boosting / XGBoost / LightGBM: Nhóm thuật toán boosting mạnh mẽ, xây dựng nhiều mô hình nhỏ và kết hợp lại. Cho kết quả chính xác cao, phổ biến trong thực tế.
* Deep Learning (MLP, CNN, RNN, Transformer): Các mô hình học sâu có khả năng trích xuất đặc trưng mạnh, phù hợp với dữ liệu lớn như hình ảnh, âm thanh, văn bản.

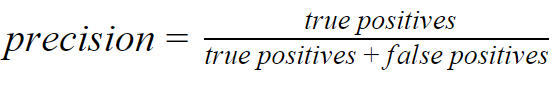
**2.1.5. Lý thuyết về độ đánh giá mô hình**

* Accuracy được định nghĩa là tỷ lệ phần trăm dự đoán đúng cho dữ liệu thử nghiệm. Nó có thể được tính toán dễ dàng bằng cách chia số lần dự đoán đúng cho tổng số lần dự đoán.



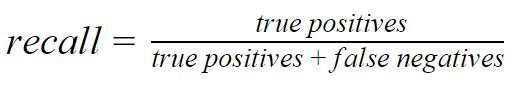
Hình 2.2: Công thức tính accuracy

* Precision được định nghĩa là phần nhỏ của các ví dụ có liên quan (tích cực thực sự) trong số tất cả các ví dụ được dự đoán là thuộc một lớp nhất định.



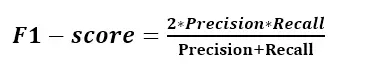
Hình 2.3: Công thức tính precision

* Recall được định nghĩa là phần nhỏ của các ví dụ được dự đoán thuộc về một lớp so với tất cả các ví dụ thực sự thuộc về lớp đó.



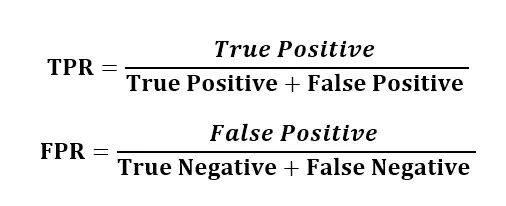
*Hình 2.4: Công thức tính recall*

* F1-score là trung bình điều hòa giữa *độ chuẩn xác* và *độ phủ*. Do đó nó là chỉ số đại diện trong việc đánh giá tỷ lệ dự báo đúng của các trường hợp mẫu *dương tính.*



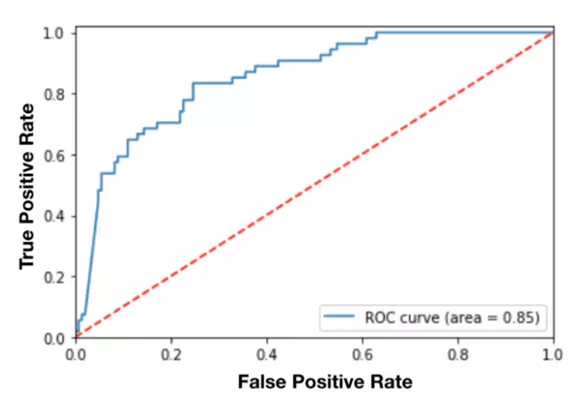
Hình 2.5: Công thức tính F1-score

* AUC (Area Under the Curve) là một phép đo tổng hợp về hiệu suất của phân loại nhị phân trên tất cả các giá trị ngưỡng có thể có. Để hiểu rõ hơn về metric này, chúng ta sẽ tìm hiểu về một khai niệm cơ sở trước, đó là ROC Curve
* ROC Curve (The receiver operating characteristic curve) là một đường cong biểu diễn hiệu suất phân loại của một mô hình phân loại tại các ngưỡng threshold. Về cơ bản, nó hiển thị True Positive Rate (TPR) so với False Positive Rate (FPR) đối với các giá trị ngưỡng khác nhau. Các giá trị TPR, FPR được tính như sau:



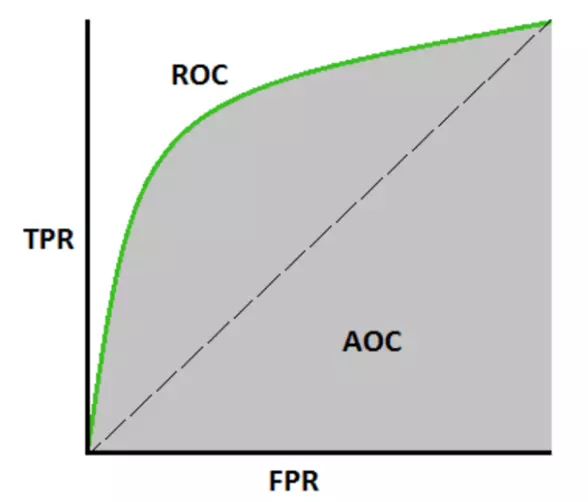
Hình 2.6: Công thức tính TPR và FPR

* ROC tìm ra TPR và FPR ứng với các giá tị ngưỡng khác nhau và vẽ biểu đồ để dễ dàng quan sát TPR so với FPR. Ví dụ dưới đây là một đường cong ROC



Hình 2.7: Đường cong ROC

* AUC là chỉ số được tính toán dựa trên đường cong ROC nhằm đánh giá khả năng phân loại của mô hình tốt như thê nào. Phần diện tích nằm dưới đường cong ROC và trên trục hoành chính là AUC, có giá trị nằm trong khoảng [0, 1].



Hình 2.8: Chỉ số đánh giá phân loại ROC – AUC

**2.2. Tổng quan về công nghệ sử dụng**

**2.2.1. FASP Protocol là gì?**

* FASP (Fast, Adaptive, Secure Protocol) là giao thức truyền dữ liệu độc quyền do IBM Aspera phát triển, được tối ưu hóa cho các mạng IP thông thường (best-effort). Giao thức này kết hợp linh hoạt giữa hai cơ chế truyền dữ liệu: sử dụng UDP cho kênh truyền dữ liệu chính nhằm loại bỏ yêu cầu xác nhận từng gói tin, đồng thời dùng TCP cho kênh điều khiển để đảm bảo tính tin cậy trong quá trình truyền. Nhờ đó, FASP đạt được hiệu suất cao hơn đáng kể so với các giao thức truyền thống vốn dễ bị nghẽn khi băng thông lớn hoặc độ trễ cao.
* Về mặt hiệu năng, FASP có khả năng tận dụng tối đa băng thông mạng và bù đắp độ trễ lên tới 10–20% RTT (Round-Trip Time) ngay cả trong điều kiện mất gói tin từ 1–2%, giúp tốc độ truyền có thể đạt từ 1 đến 10 Gbps trên các mạng diện rộng (WAN). Giao thức này cũng đảm bảo tính an toàn thông qua cơ chế mã hóa AES 128/192/256 bit, tuân thủ tiêu chuẩn bảo mật FIPS 140-2, đồng thời sử dụng xác thực dựa trên token thay cho SSH thông thường để tăng cường khả năng kiểm soát truy cập. Về độ tin cậy, FASP có thể tự động phát hiện và truyền lại các gói tin bị mất, đồng thời hỗ trợ cơ chế dự phòng (fallback) sang HTTP trong trường hợp UDP bị chặn, giúp duy trì luồng truyền ổn định trong nhiều môi trường mạng khác nhau.
* FASP được ứng dụng rộng rãi trong các lĩnh vực yêu cầu truyền tải dữ liệu dung lượng lớn và tốc độ cao như truyền tệp video độ phân giải 4K, đồng bộ dữ liệu khoa học giữa các máy chủ (theo cả hướng đơn, hai chiều hoặc đa hướng), và đặc biệt là trong các hệ thống streaming thời gian thực. Tuy nhiên, hạn chế của giao thức này là mang tính độc quyền, đòi hỏi giấy phép bản quyền của IBM và phải sử dụng phần mềm Aspera chuyên dụng, do đó không hoàn toàn là mã nguồn mở. FASP được phát triển lần đầu bởi công ty Aspera vào năm 2007, sau đó được IBM mua lại vào năm 2014, và công nghệ này đã được trao giải Emmy Award nhờ đóng góp nổi bật trong lĩnh vực truyền dữ liệu hiệu năng cao.

### 2.2.2. Các thành phần chính của FASP API

* FASP API được cung cấp thông qua Aspera SDK (Software Development Kit), là bộ công cụ hỗ trợ đa ngôn ngữ và nhiều giao diện khác nhau, giúp các lập trình viên dễ dàng tích hợp giao thức FASP vào các ứng dụng. Một trong những thành phần chính của SDK là FASP Manager SDK, đây là một thư viện cấp cao cho phép khởi tạo, quản lý và giám sát các tiến trình truyền dữ liệu bằng giao thức FASP, bao gồm cả quá trình tải lên và tải xuống. Thư viện này hỗ trợ các cơ chế như callback hoặc polling để theo dõi tiến trình truyền, đồng thời cho phép người dùng tạm dừng, tiếp tục hoặc điều chỉnh tốc độ truyền dữ liệu tùy theo yêu cầu. FASP Manager SDK tương thích với nhiều ngôn ngữ lập trình phổ biến như Python, C++, Java, C#, .NET, Ruby, Node.js và Go, giúp việc phát triển ứng dụng trở nên linh hoạt và phù hợp với nhiều nền tảng khác nhau.
* Ngoài ra, FASP Stream SDK là phiên bản mở rộng của FASP Manager, được thiết kế để truyền dữ liệu ở dạng byte-stream mà không cần thông qua hệ thống tệp (file system). Điều này đặc biệt phù hợp với các ứng dụng truyền tải dữ liệu thời gian thực hoặc live feed, chẳng hạn như streaming video hoặc các ứng dụng phân tích dữ liệu liên tục. FASP Stream SDK hỗ trợ nhiều mô hình truyền dữ liệu như File-to-Stream, Stream-to-File và Stream-to-Stream, giúp dữ liệu được truyền trực tiếp từ ứng dụng đến máy chủ Aspera mà không cần lưu trữ tạm, từ đó tăng hiệu suất và giảm độ trễ trong quá trình xử lý.
* Một thành phần quan trọng khác của hệ thống là Node API, được thiết kế dưới dạng RESTful API dành cho máy chủ Aspera Node (server). Node API cho phép tạo token xác thực, thiết lập các phiên truyền dữ liệu và quản lý toàn bộ quá trình tải lên hoặc tải xuống. Các điểm cuối (endpoint) chính bao gồm /files/upload\_setup để khởi tạo token tải lên và /files/download\_setup để tạo token tải xuống, trong đó dữ liệu được truyền đi thông qua cấu trúc JSON payload có định dạng {"transfer\_requests": [{"paths": [{"source": "/path"}]}]}. Node API hỗ trợ nhiều cơ chế xác thực khác nhau như Basic Authentication với tên người dùng và mật khẩu hoặc JWT Token, trong đó token có thể được truyền thông qua biến môi trường ASPERA\_SCP\_TOKEN hoặc tham số -W trong công cụ dòng lệnh ascp. Trong trường hợp mạng bị chặn UDP, hệ thống có thể tự động chuyển sang giao thức HTTP (fallback) để đảm bảo quá trình truyền không bị gián đoạn.
* Bên cạnh Node API, Aspera còn cung cấp Faspex API, một giao diện REST API dựa trên JSON được sử dụng trong ứng dụng web chia sẻ file Aspera Faspex. Faspex API cho phép người dùng tạo, gửi và nhận các gói dữ liệu (package) thông qua giao thức FASP, đồng thời hỗ trợ cơ chế xác thực tự động bằng JWT. Việc xác thực trong Faspex API được thực hiện thông qua Client ID kết hợp với RSA key pair, trong đó khóa công khai được lưu trong hệ thống Faspex và khóa riêng dùng để ký JWT assertion trước khi gửi yêu cầu tới endpoint /auth nhằm lấy bearer token phục vụ cho các lệnh tiếp theo. Sau khi xác thực, người dùng có thể gửi yêu cầu đến các endpoint như /send, /received hoặc /issue-token để gửi gói dữ liệu, lấy danh sách nhận hoặc khởi tạo token truyền tải, giúp tự động hóa toàn bộ quy trình trao đổi dữ liệu một cách an toàn và hiệu quả.
* Bên cạnh các API hiện đại này, Aspera Web Services là một giao diện cũ hơn, được xây dựng dựa trên công nghệ SOAP/XML với định nghĩa WSDL, cho phép khởi tạo và giám sát các phiên truyền từ các ứng dụng ở xa. Tuy nhiên, do hạn chế về khả năng mở rộng và tốc độ xử lý, giao diện SOAP này hiện nay ít được sử dụng và phần lớn đã được thay thế bởi các API RESTful như Node API. Ngoài ra, Aspera CLI (Command-Line Interface) cũng là một công cụ quan trọng, cho phép người dùng tương tác trực tiếp với REST API thông qua các lệnh dòng, ví dụ như ascli node upload --src /local --dest user@remote. Công cụ này giúp tự động hóa quá trình truyền dữ liệu và dễ dàng tích hợp với các hệ thống quản lý file hoặc pipeline DevOps.

### 2.2.3. Bảo mật và Troubleshooting

* Trong quá trình vận hành và truyền dữ liệu bằng giao thức FASP, vấn đề bảo mật được đặt lên hàng đầu nhằm đảm bảo an toàn tuyệt đối cho dữ liệu trong suốt quá trình truyền tải. Cơ chế bảo mật của FASP dựa trên xác thực bằng token (Token-based Authentication) sử dụng chuẩn JWT (JSON Web Token) với thuật toán RS256, giúp xác minh danh tính người dùng một cách an toàn và tránh việc lộ thông tin nhạy cảm. Ngoài ra, toàn bộ luồng dữ liệu được mã hóa end-to-end, kết hợp với hỗ trợ cơ chế TLS fallback trong trường hợp giao thức UDP bị chặn, giúp bảo đảm tính toàn vẹn và bảo mật của dữ liệu trên môi trường mạng công cộng. Đặc biệt, FASP không sử dụng mật khẩu SSH truyền thống, nhờ đó giảm thiểu rủi ro rò rỉ thông tin đăng nhập trong quá trình xác thực.
* Trong quá trình sử dụng FASP API, một số lỗi phổ biến có thể xảy ra và cần được xử lý kịp thời. Lỗi “403 Forbidden” thường xuất hiện khi token xác thực đã hết hạn hoặc bị thiếu, khiến hệ thống từ chối truy cập. Ngoài ra, khi UDP bị chặn bởi tường lửa hoặc cấu hình mạng, hệ thống sẽ tự động chuyển sang chế độ HTTP fallback để tiếp tục truyền dữ liệu, mặc dù tốc độ sẽ chậm hơn so với giao thức UDP. Một lỗi khác thường gặp là giới hạn tốc độ truyền (rate limit), có thể được khắc phục bằng cách điều chỉnh tham số target\_rate trong API để tối ưu hóa băng thông theo yêu cầu hệ thống.
* Để hỗ trợ quá trình kiểm tra và khắc phục sự cố, FASP cung cấp các công cụ debug mạnh mẽ. Người vận hành có thể kích hoạt chế độ gỡ lỗi bằng cách sử dụng tùy chọn --log-level=debug trong lệnh ascp hoặc cài đặt callback listener trong SDK để theo dõi chi tiết các sự kiện trong quá trình truyền. Những công cụ này giúp phát hiện sớm các vấn đề về xác thực, tốc độ hoặc kết nối mạng, từ đó đảm bảo hệ thống FASP hoạt động ổn định, an toàn và đạt hiệu năng cao nhất trong mọi điều kiện truyền tải.

# CHƯƠNG 3. DATASET VÀ TIỀN XỬ LÝ DỮ LIỆU

## 3.1. Nguồn dữ liệu

* Dữ liệu đóng vai trò then chốt trong việc quyết định độ chính xác của mô hình dự đoán. Trong đề tài này, tập dữ liệu được xây dựng bằng cách kết hợp bộ dữ liệu công khai về các chỉ số sức khỏe (ví dụ: BRFSS Diabetes) cùng với dữ liệu thu thập bổ sung. Sự kết hợp này giúp tạo ra một dataset đa dạng, bao phủ nhiều nhóm đối tượng và tình trạng sức khỏe, từ đó tăng độ tin cậy và khả năng khái quát của mô hình dự đoán bệnh tiểu đường.

### 3.1.1. Dữ liệu công khai

* Đề tài sử dụng bộ dữ liệu Diabetes Health Indicators Dataset được giới thiệu tại:
* Kaggle - BRFSS 2015
* Link:

https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset

* Bộ dữ liệu này do Hệ thống Giám sát Yếu tố Nguy cơ Hành vi, hay còn gọi là BRFSS (Behavioral Risk Factor Surveillance), năm 2015 cung cấp gồm 3 tập dữ liệu:
* diabetes\_012\_health\_indicators\_BRFSS2015.csv
* diabetes\_binary\_5050split\_health\_indicators\_BRFSS2015.csv
* diabetes\_binary\_health\_indicators\_BRFSS2015.csv
* Lý do chọn bộ dữ liệu này:
* Nguồn dữ liệu lớn, có chất lượng cao
* Được thu thập bởi Trung tâm Kiểm soát và Phòng ngưa Dịch bệnh Hoa Kỳ (CDC)

### 3.1.2. Phân tích và lựa chọn tập dữ liệu phù hợp

* diabetes\_012\_health\_indicators\_BRFSS2015.csv
* Số lượng phản hồi: 253.680 phản hồi
* Biến phụ thuộc:
* 0: Không tiểu đường hoặc chỉ bị tiểu đường trong thời kỳ mang thai
* 1: Tiền tiểu đường
* 2: Tiểu đường
* Đặc điểm: Có sự mất cân bằng giữa các nhóm (phần lớn mục tiêu lớp 0 – không tiểu đường)

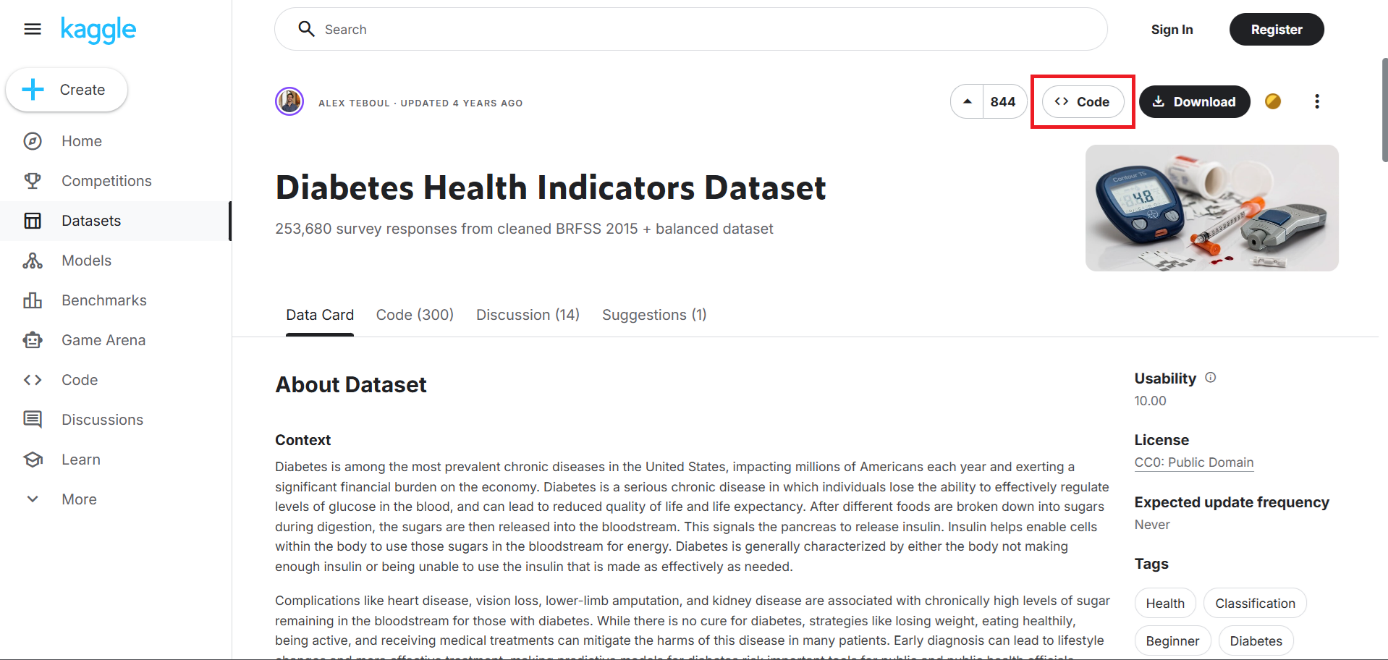
🡪 Có thể gây khó khăn trong việc xây dựng mô hình dự đoán

* diabetes\_binary\_5050split\_health\_indicators\_BRFSS2015.csv
* Số lượng phản hồi: 70.692 phản hồi
* Biến phụ thuộc:
* 0: Không tiểu đường
* 1: Tiền tiểu đường hoặc tiểu đường
* Đặc điểm:
* Cân bằng 50 - 50 giữa 2 lớp, giúp giảm thiểu vấn đề mất cân bằng dữ liệu
* Quy mô nhỏ hơn so với tập dữ liệu đầu tiên
* diabetes\_binary\_health\_indicators\_BRFSS2015.csv
* Số lượng phản hồi: 253.680 phản hồi
* Biến phụ thuộc:
* 0: Không tiểu đường
* 1: Tiền tiểu đường hoặc tiểu đường
* Đặc điểm: Có sự mất cân bằng giữa các lớp nhưng nhờ quy mô lớn nên nó vẫn cung cấp đủ thông tin để xây dựng mô hình

🡪 Chọn tập dữ liệu thứ 3 với biến mục tiêu được chia làm 2 lớp đơn giản 0 và 1, dễ phân loại, quy mô dữ liệu lớn, đảm bảo độ chính xác và tính tổng quát của mô hình.

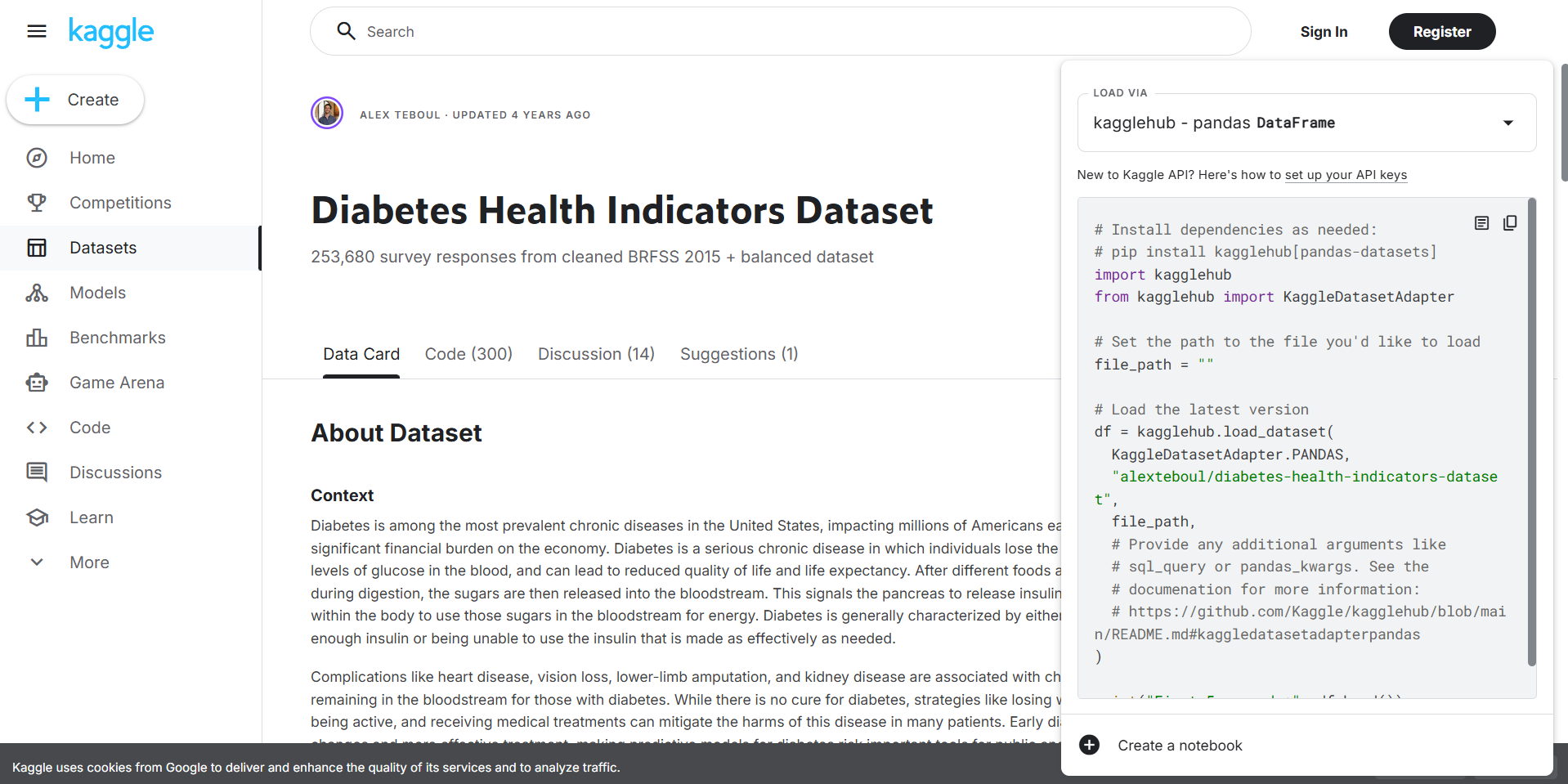
### 3.1.3. Hướng dẫn tải dataset

* Bước 1: Nhấp vào nút “Code” như hình minh họa.

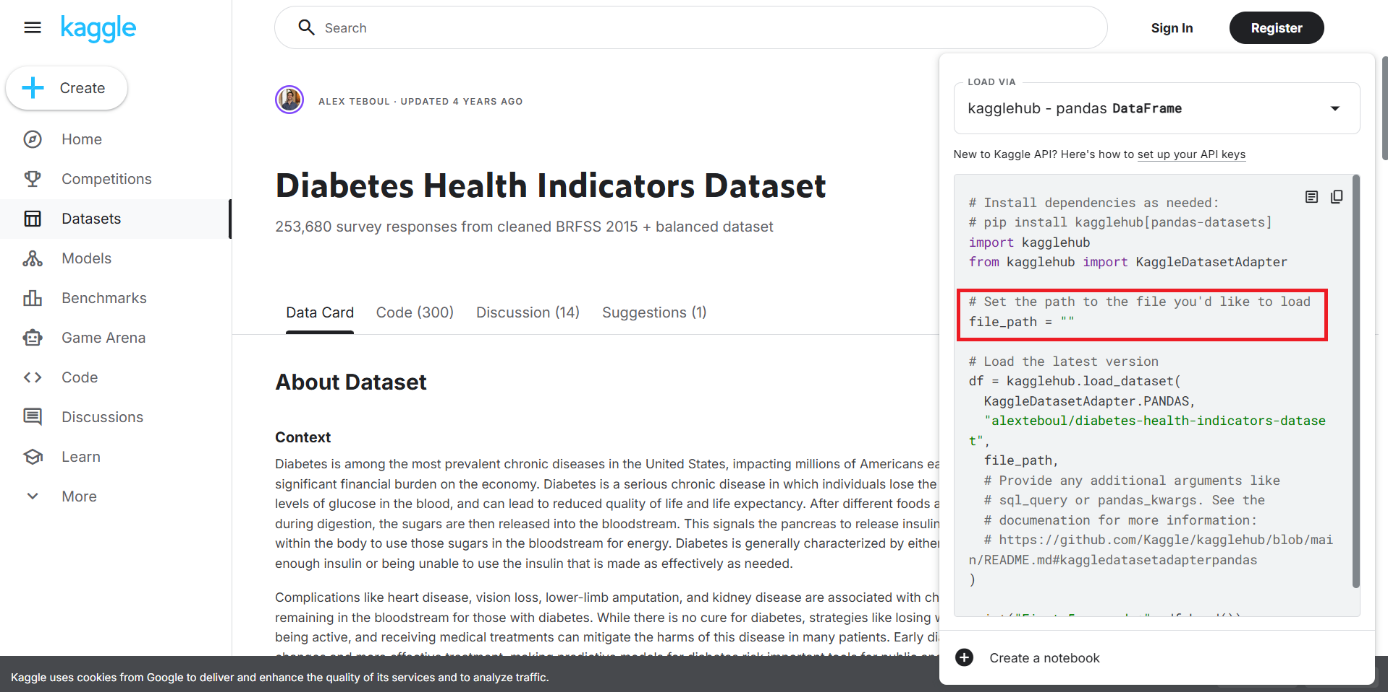


Hình 3.1: Nhấp vào nút “Code”.

* Bước 2: Kaggle sẽ tự động mở một cửa sổ hiển thị đoạn mã dùng để tải dataset.



Hình 3.2: Cửa sổ hiển thị đoạn mã dùng để tải dataset.

* Bước 3: Tại ô hiển thị tên file, bạn chỉ cần thay bằng tên file CSV mà bạn muốn sử dụng. Ở đây, chúng ta dùng file diabetes\_binary\_health\_indicators\_BRFSS2015.csv.

Hình 3.3: Nơi điền tên file.

* Bước 4: Sao chép đoạn code được tạo vào file **.ipynb** hoặc **.py** của bạn, sau đó chạy chương trình. Khi đó dataset sẽ được import thành công vào môi trường làm việc.

## 3.2. Mô tả dữ liệu

### 3.2.1. Chi tiết các biến trong bộ dữ liệu

Gồm 21 biến và được chia làm 3 nhóm:

* Biến nhị phân

|  |  |  |
| --- | --- | --- |
|  | Tên biến | Mô tả |
| Biến nhị phân | Diabetes\_binary | Biến mục tiêu, có giá trị 0 - không tiểu đường và 1 - tiền tiểu đường hoặc tiểu đường |
| HighBP | Cho biết người tham gia có bị huyết áp cao (1) hay không (0)? |
| HighChol | Có bị cholesterol cao (1) hay không (0)? |
| CholCheck | Đã kiểm tra cholesterol trong 5 năm qua (1) hay chưa (1)? |
| Smoker | Có hút thuốc lá (1) hay không (0)? |
| Stroke | Đã từng bị đột quỵ (1) hay chưa (0)? |
| HeartDiseaseorAttack | Có từng bị đau tim hoặc bệnh tim (1) hay không (0)? |
| PhysActivity | Có tham gia hoạt động thể chất trong 30 ngày qua (1) hay không (0)? |
| Fruits | Có tiêu thụ trái cây ít nhất 1 lần trong ngày (1) hay không (0)? |
| Veggies | Có tiêu thụ rau củ ít nhất 1 lần trong ngày (1) hay không (0)? |
| HvyAlcoholConsump | Có tiêu thụ rượu bia nặng (1) hay không (0)? |
| AnyHealthcare | Có tiếp cận dịch vụ chăm sóc sức khỏe (1) hay không (0)? |
| NoDocbcCost | Có từng không đi bác sĩ vì chi phí (1) hay không (0)? |
| DiffWalk | Có gặp khó khăn trong việc đi bộ (1) hay không (0)? |
| Sex | Giới tính, với 0 là nữ 1 là nam |

Bảng 3.1: Bảng biến nhị phân

* Biến thứ bậc

|  |  |  |
| --- | --- | --- |
|  | Tên biến | Mô tả |
| Biến thứ bậc | GenHlth | Đánh giá sức khỏe tổng quát từ 1 (tốt nhất) đến 5 (kém nhất) |
| Age | Độ tuổi, được chia làm 13 nhóm từ 1 (18 - 24 tuổi) đến 13 (80 tuổi trở lên) |
| Education | Trình độ học vấn, từ 1 (không học) đến 6 (đại học trở lên) |
| Income | Mức thu nhập, từ 1 (thấp nhất) đến 8 (cao nhất) |

Bảng 3.2: Bảng biến thứ bậc

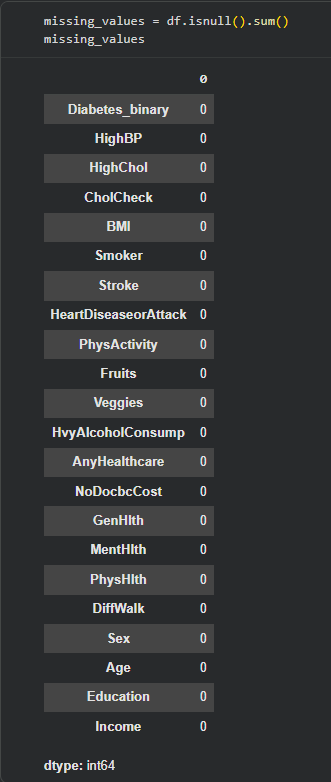
* Biến liên tục

|  |  |  |
| --- | --- | --- |
|  | Tên biến | Mô tả |
| Biến liên tục | BMI | Chỉ số khối cơ thể (Body Mass Index), thể hiện tình trạng cân nặng của người tham gia |
| MenHlth | Số ngày gặp vấn đề về sức khỏe tinh thần trong 30 ngày qua |
| PhysHlth | Số ngày gặp vấn đề về sức khỏe thể chất trong 30 ngày qua |

Bảng 3.3: Bảng biến liên tục

### 3.2.2. Làm sạch dữ liệu

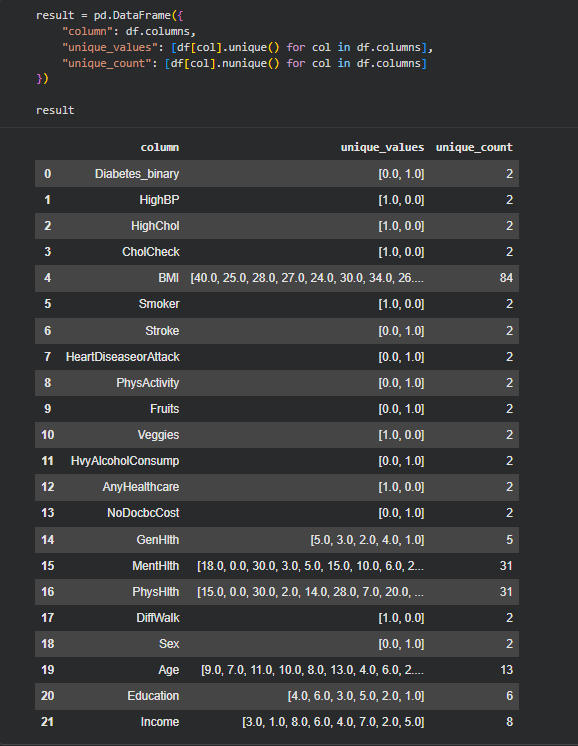
#### 3.2.2.1. Xử lý dữ liệu bị thiếu

**

Hình 3.4: Kết quả kiểm tra dữ liệu có dữ liệu thiếu hay không

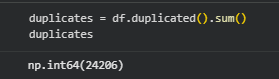
🡪 Không có giá trị bị thiếu

#### 3.2.2.2. Kiểm tra giá trị độc nhất

**

Hình 3.5: Kết quả kiểm tra giá trị độc nhất

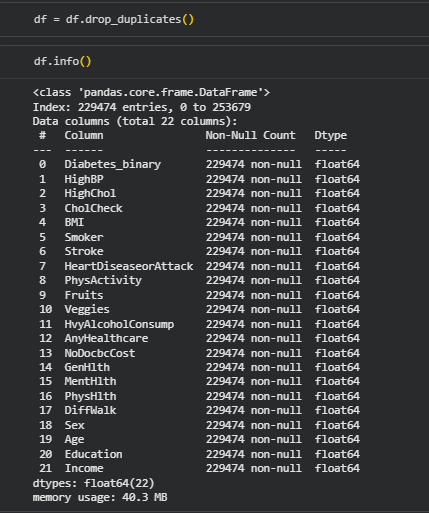
#### 3.2.2.3. Kiểm tra và loại bỏ giá trị trùng lặp



Hình 3.6: Kết quả kiểm tra giá trị trùng lặp

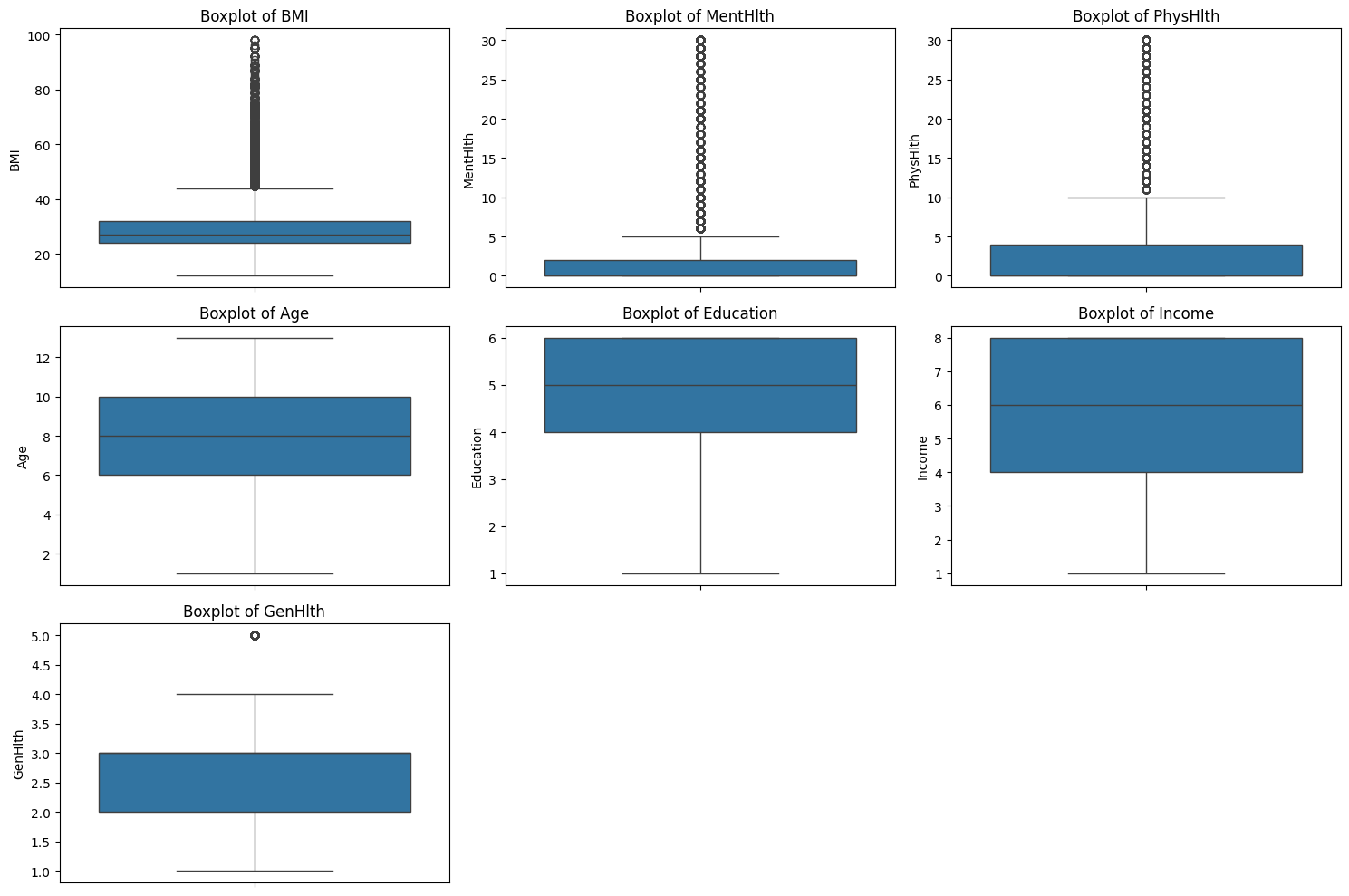
🡪 Có 24.206 dữ liệu trùng lặp

🡪 Tiến hành xóa dữ liệu trùng lặp

**

Hình 3.7: Thông tin dữ liệu sau khi xóa dữ liệu trùng lặp

#### 3.2.2.4. Xác định Outliers

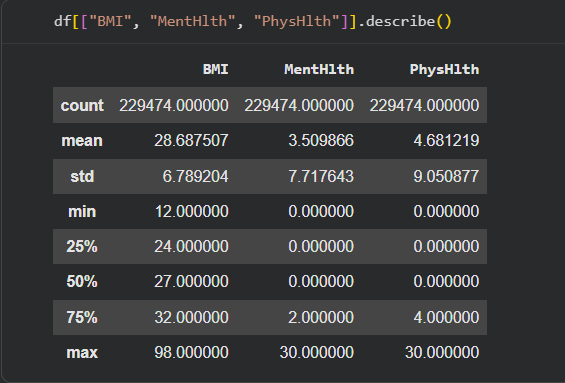


Hình 3.8: Biểu đồ boxplot kiểm tra Outliers

* BMI:
* Có rất nhiều outliers, chủ yếu là BMI > 40.
* BMI bình thường dao động ~18–35 → phần lớn giá trị >40 là bất thường (béo phì / có vấn đề liên quan đến cân nặng).
* Outliers trải dài lên đến ~100, cho thấy dữ liệu bị lệch phải mạnh (right-skewed).
* MentHlth:
* Nhiều outliers > 5.
* Đây là số ngày sức khỏe tâm lý kém trong 30 ngày → giá trị cao là hợp lệ.
* Tuy nhiên phân bố lệch phải, phần lớn là 0–2 ngày, còn 20–30 ngày xuất hiện như outliers.
* PhysHlth:
* Tương tự MentHlth: phần lớn là 0–2, nhưng có nhiều giá trị 20–30.
* Boxplot tạo ra nhiều điểm outlier do phân phối lệch.
* Age, Education, Income, GenHlth:
* Tương đối ổn định, không có quá nhiều outliers đáng chú ý.

### 3.2.3. Giải thích sự xuất hiện của các giá trị ngoại lai

#### 3.2.3.1. Phân tích tổng quan dữ liệu của 3 biến BMI, MentHlth, PhysHlth



Hình 3.9: Kết quả phân tích tổng quan dữ liệu của 3 biến MBI, MentHlth, PhysHlth

* BMI:
* Trung bình: 28,69 -> nằm trong phạm vi thừa cân quốc tế.
* Q3 = 32 nhưng Max = 98 vượt xa trung bình và các phân vị chính.

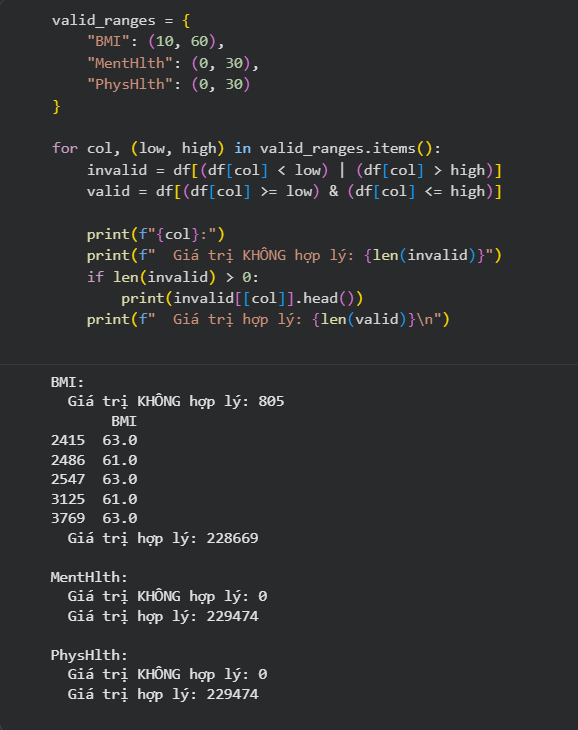
🡪 Có sự xuất hiện của các giá trị bất thường.

* MentHlth:
* Phân bố không đồng đều.
* Q3 = 2 ngày nghĩa là 75% dữ liệu người tham gia chỉ ghi nhận 2 ngày gặp vấn .đề sức khỏe tinh thần trong tháng, tuy nhiên giá trị tối đa là 30 ngày.

🡪 Có một số trường hợp đặc biệt ghi nhận vấn đề sức khỏe tinh thần cả tháng.

* PhysHlth: tương tự như MentHlth.

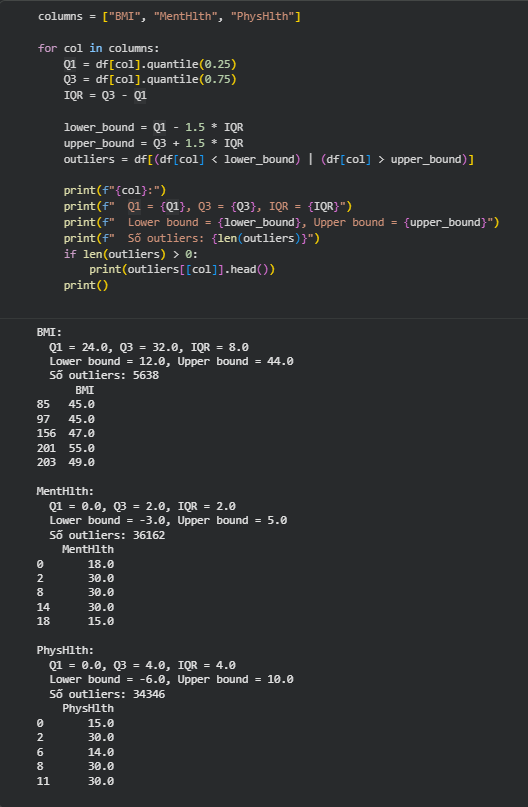
#### 3.2.3.2. Xác định các giá trị hợp lý



Hình 3.10: Kết quả các giá trị hợp lý

* BMI:
* Ngưỡng hợp lý: 10 - 60
* Tổng số giá trị hợp lý: 228669
* Tổng số giá trị không hợp lý: 805 chiếm 0.35%
* MentHlth:
* Ngưỡng hợp lý: 0 - 30
* Tổng số giá trị hợp lý: 229474
* Tổng số giá trị không hợp lý: 0
* PhysHlth:
* Ngưỡng hợp lý: 0 - 30
* Tổng số giá trị hợp lý: 229474
* Tổng số giá trị không hợp lý: 0

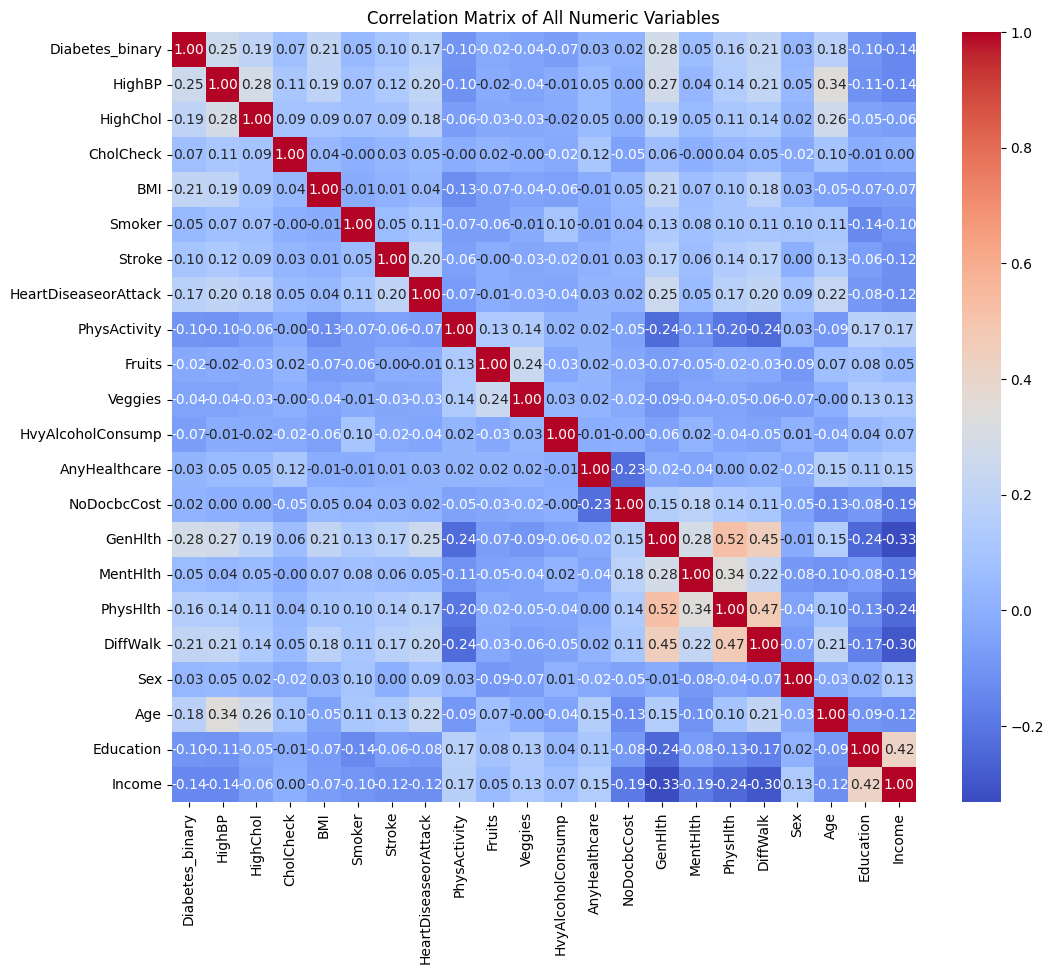
#### 3.2.3.3. Xác định Outliers theo phương pháp IQR



Hình 3.11: Kết quả xác định Outliers

* BMI:
* Ngưỡng dưới: 12
* Ngưỡng trên: 44
* Số lượng outliers: 5638 chiếm 2.46%
* MentHlth:
* Ngưỡng dưới: -3
* Ngưỡng trên: 5
* Số lượng outliers: 36162 chiếm 15.76%
* PhysHlth:
* Ngưỡng dưới: -6
* Ngưỡng trên: 10
* Số lượng outliers: 34346 chiếm 14.97%

#### 3.2.3.4. Phân tích dữ liệu khám phá



Hình 3.12: Ma trận tương quan của các biến

🡪 Không có chỉ số tương quan nào ≥ 0.8

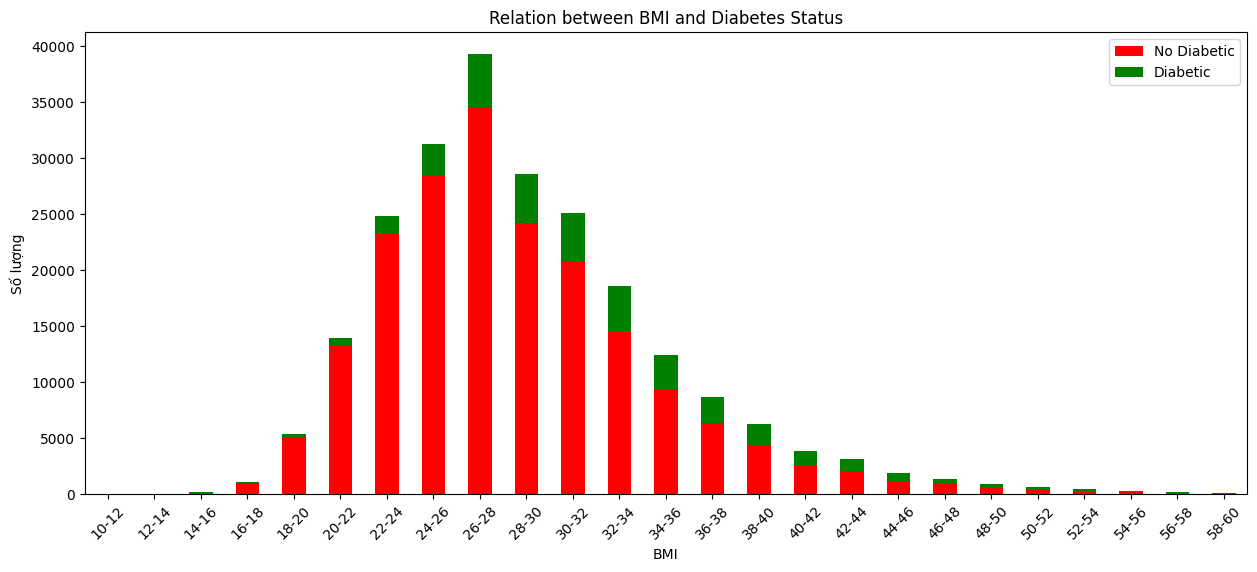
🡪 Các feature không tác động lẫn nhau quá nhiều mà chỉ có ảnh hưởng lên biến mục tiêu.

🡪 Chia làm 2 nhóm để phân tích:

* Mối quan hệ giữa từng biến độc lập chứa ngoại lai và GenHlth tới biến phụ thuộc Diabetes\_binary
* Mối quan hệ giữa nhóm biến độc lập và biến phụ thuộc Diabetes\_binary

### 3.2.3. Trực quan hóa dữ liệu

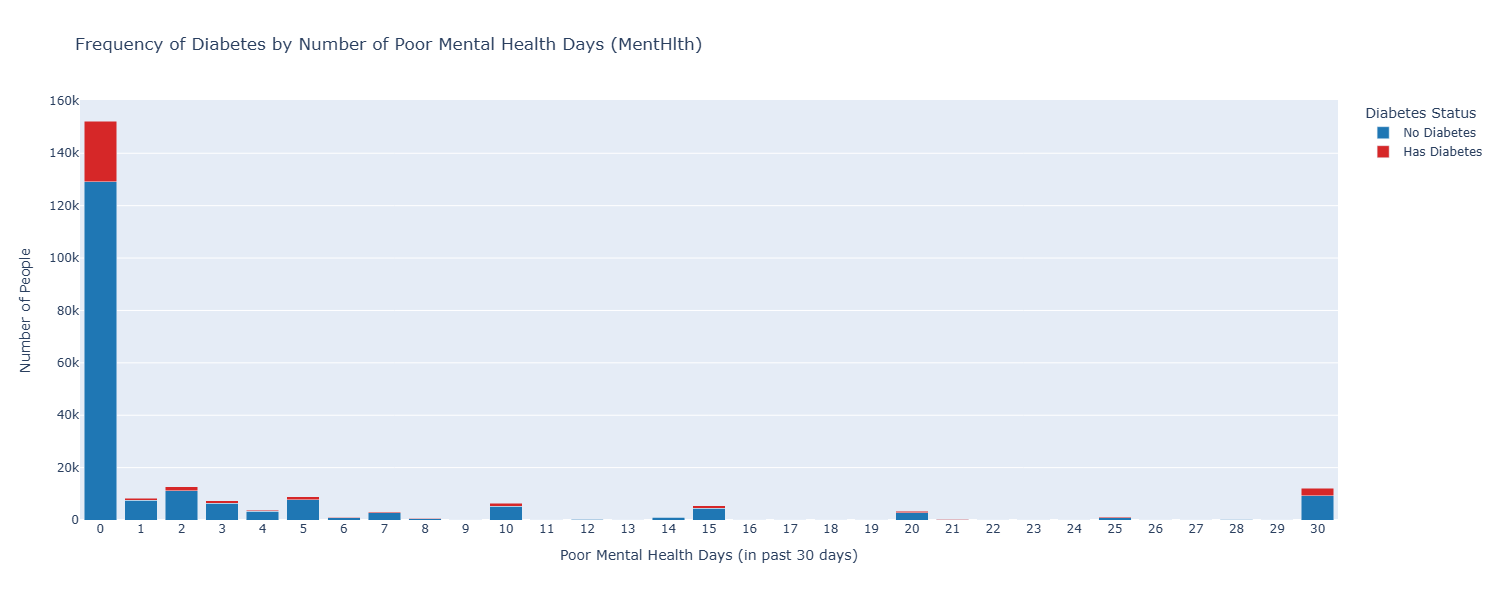
#### 3.2.3.1. Mối quan hệ giữa từng biến độc lập chứa ngoại lai và GenHlth tới biến phụ thuộc Diabetes\_binary



Hình 3.13: Biểu đồ cột thể hiện mối quan hệ của BMI và Diabetes\_binary

* Số lượng người không mắc tiểu đường chiếm ưu thế tuyệt đối ở hầu hết các nhóm BMI.
* Số người mắc tiểu đường rất ít ở nhóm BMI thấp, tăng dần khi BMI tăng và đạt đỉnh ở nhóm BMI cao, sau đó giảm mạnh ở BMI rất cao.
* Thừa cân và béo phì (BMI >= 25, đặc biệt 26–34) có liên quan rõ rệt với tăng nguy cơ tiểu đường.

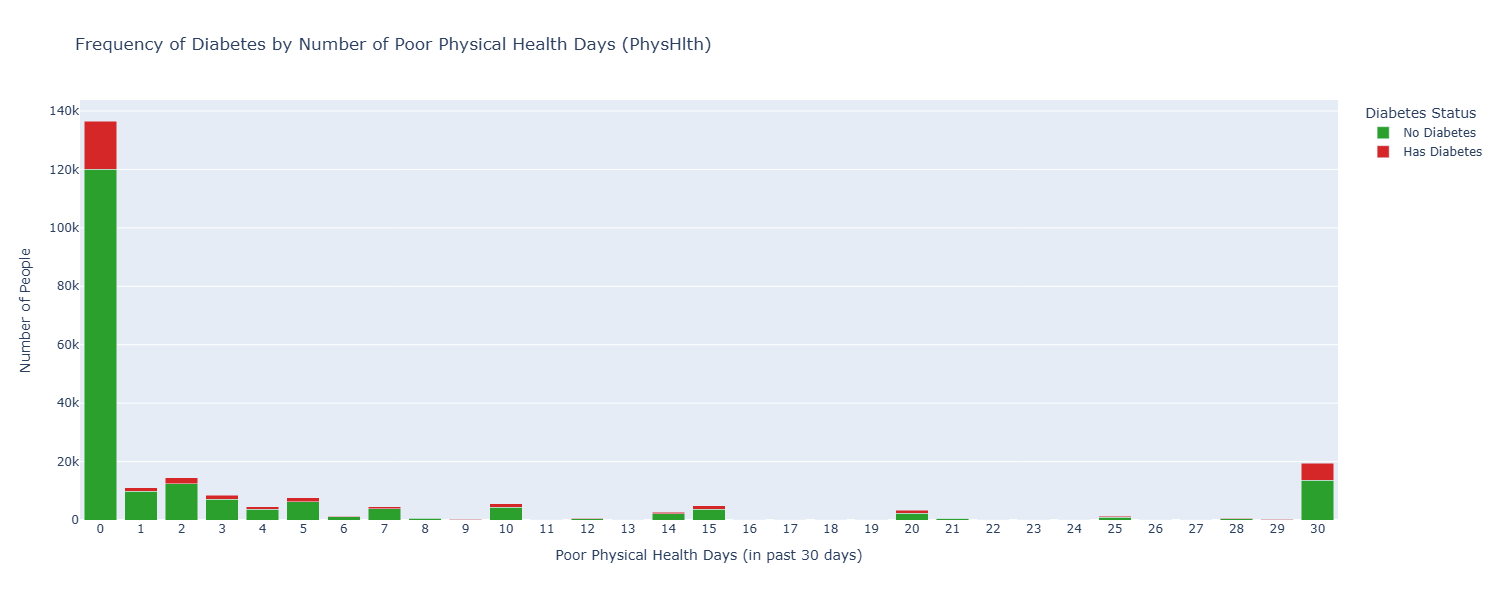
🡪 Người có chỉ số BMI cao từ 25 trở lên có nguy cơ bị tiểu đường.



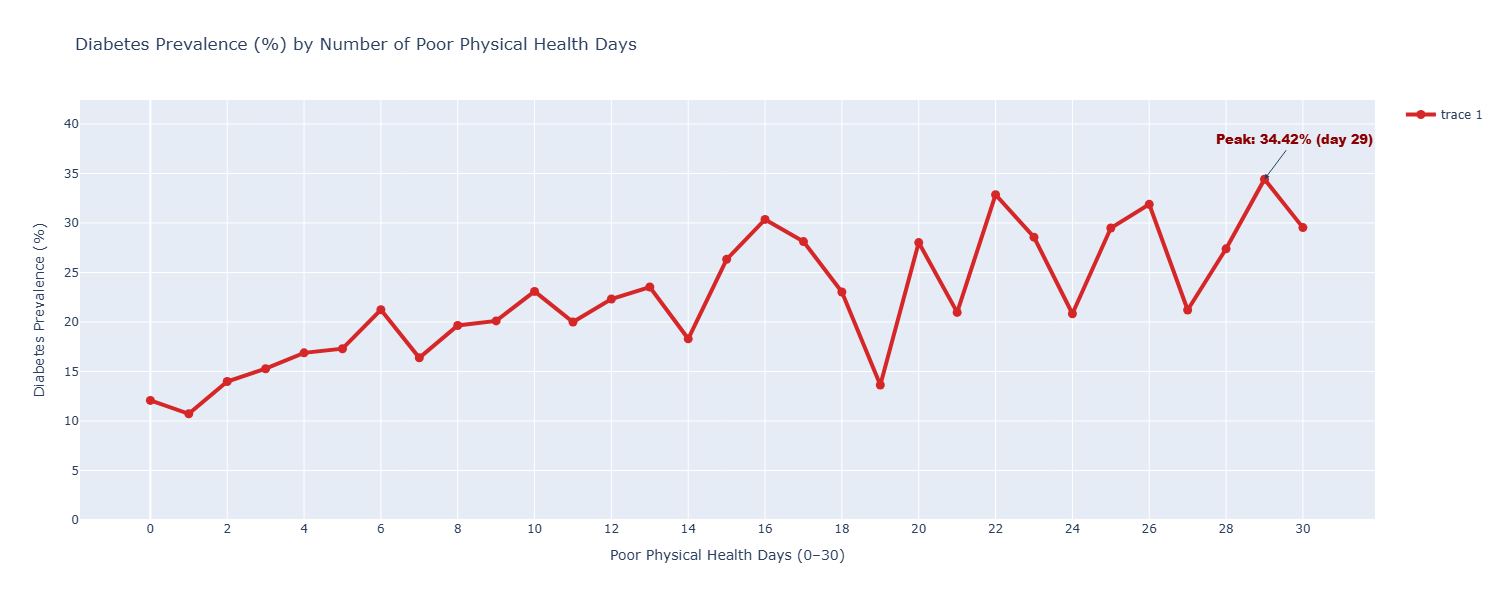
Hình 3.14: Biểu đồ thể hiện mối quan hệ của MentHlth và Diabetes\_binary

* Người có ≥15 ngày sức khỏe tâm thần kém/tháng có nguy cơ mắc tiểu đường cao gấp đôi người khỏe mạnh về tâm lý.
* 0 ngày sức khỏe tinh thần kém: 15,2%.
* 30 ngày sức khỏe tinh thần kém: 22,92%.
* Nguy cơ cao nhất: 25% ở ngày 19.
* Tỷ lệ phần trăm mắc bệnh tiểu đường có xu hướng tăng nhẹ khi sức khỏe tâm lý kém tang.

🡪 Không phải tất cả những người có sức khỏe tâm lý kém đều có khả nặng mắc bệnh tiểu đường nhưng họ có nguy cơ cao hơn.

**

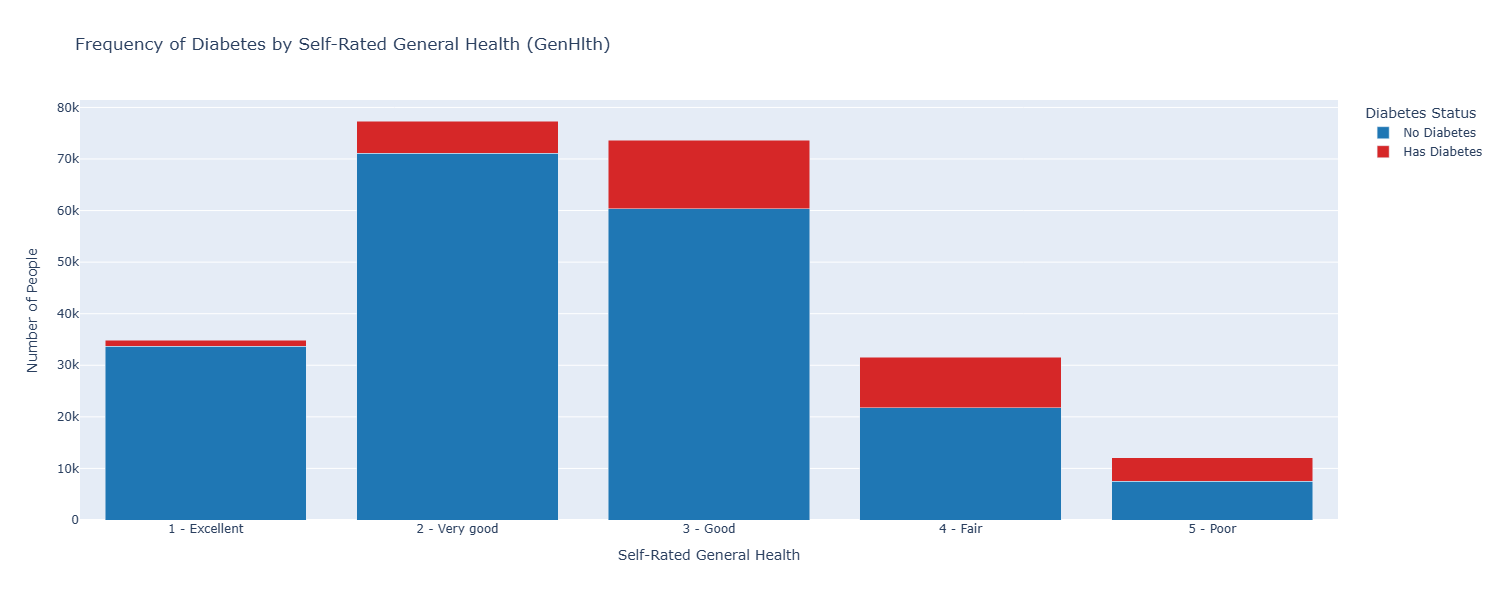
Hình 3.15: Biểu đồ cột chồng thể hiện mối quan hệ của PhysHlth và Diabetes\_binary



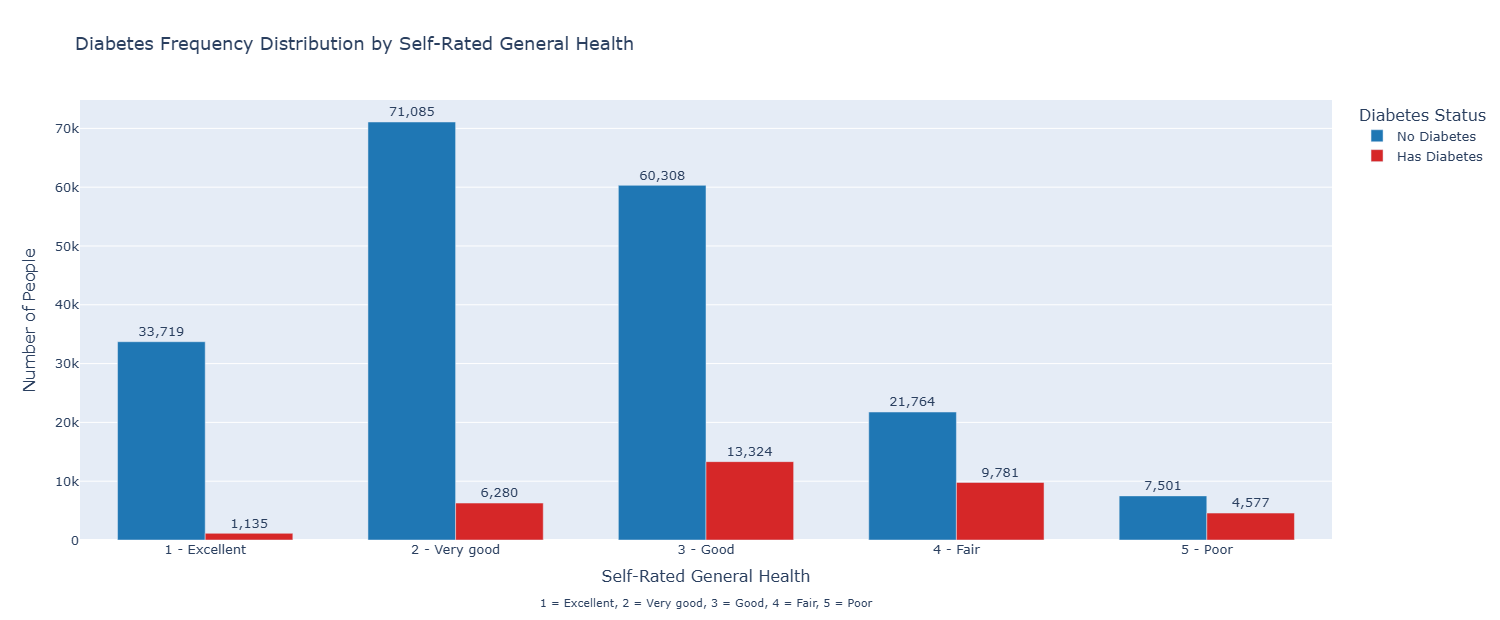
Hình 3.16: Biểu đồ đường thể hiện mối quan hệ của PhysHlth và Diabetes\_binary

* Có sự phân bố không đồng đều trong tần suất mắc bệnh.
* 0 ngày sức khỏe thể chất kém 🡪 tỷ lệ tiểu đường: 12,08%.
* 30 ngày sức khỏe thể chất kém 🡪 tỷ lệ tiểu đường: 29,53%.
* Nguy cơ cao nhất: 34,42% vào ngày thứ 29.
* Nguy cơ tăng gần như tuyến tính với số ngày sức khỏe thể chất kém.

🡪 Sức khỏe thể chất là một yếu tố dự báo tiểu đường CỰC KỲ mạnh – thậm chí còn mạnh hơn cả Sức khỏe tinh thần (trong tập dữ liệu này).



Hình 3.17: Biểu đồ cột chồng thể hiện mối quan hệ của GenHlth và Diabetes\_binary

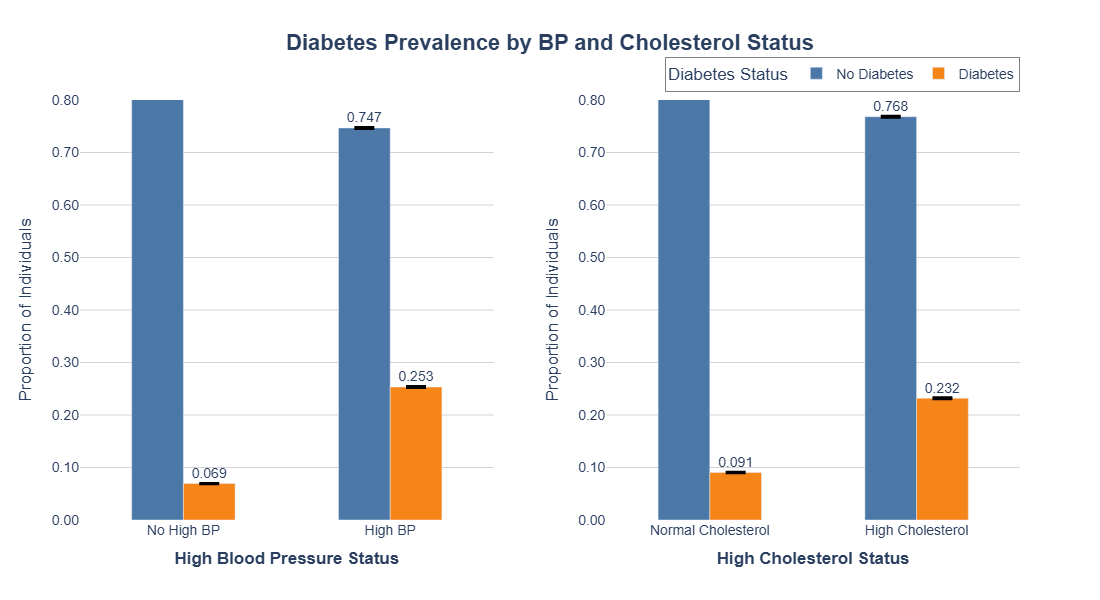


Hình 3.18: Biểu đồ cột thể hiện sự phân bố số lượng người mắc bệnh tiểu đường theo mức độ đánh giá sức khỏe tổng quát (GenHlth)

* Càng tự đánh giá sức khỏe tổng quát kém → càng nhiều người mắc đái tháo đường.
* Tỷ lệ và số lượng tuyệt đối đều tăng dần đều.

🡪 Có mối quan hệ mạnh mẽ giữa sức khỏe tổng quát và nguy cơ mắc bệnh tiểu đường.

#### 3.2.3.2. Mối quan hệ giữa nhóm biến độc lập và biến phụ thuộc Diabetes\_binary



Hình 3.19: Biểu đồ cột đôi thể hiện tỷ lệ mắc bệnh tiểu đường theo hai yếu tố

Huyết áp (High BP) và Cholesterol (HighChol)

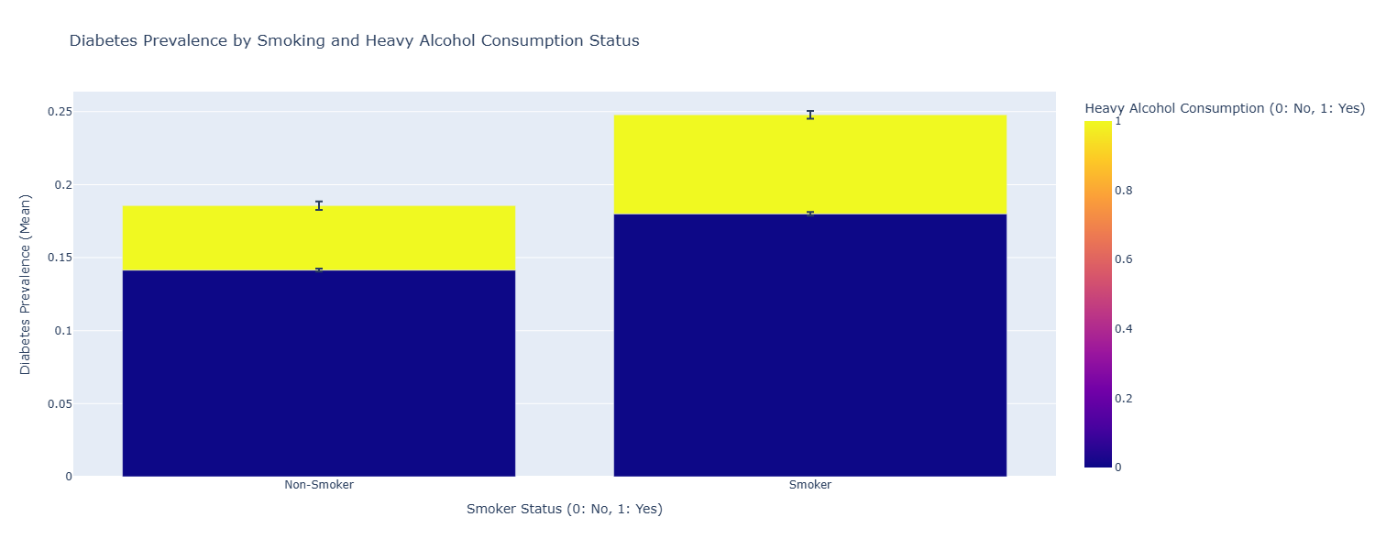
* Ở nhóm Huyết áp Cao (High BP): tỷ lệ cá nhân mắc bệnh tiểu đường là 25.3%.
* Ở nhóm Huyết áp Không Cao (No High BP), tỷ lệ mắc tiểu đường chỉ là 6.9%.

🡪 Những người có huyết áp cao có nguy cơ mắc bệnh tiểu đường cao hơn đáng kể so với những người có huyết áp bình thường.

* Ở nhóm Cholesterol Cao (High Cholesterol), tỷ lệ cá nhân mắc bệnh tiểu đường là 23.2%.
* Ở nhóm Cholesterol Bình thường (Normal Cholesterol), tỷ lệ mắc tiểu đường chỉ là 9.1%.

🡪 Tương tự như huyết áp, những người có mức cholesterol cao có nguy cơ mắc bệnh tiểu đường cao hơn.

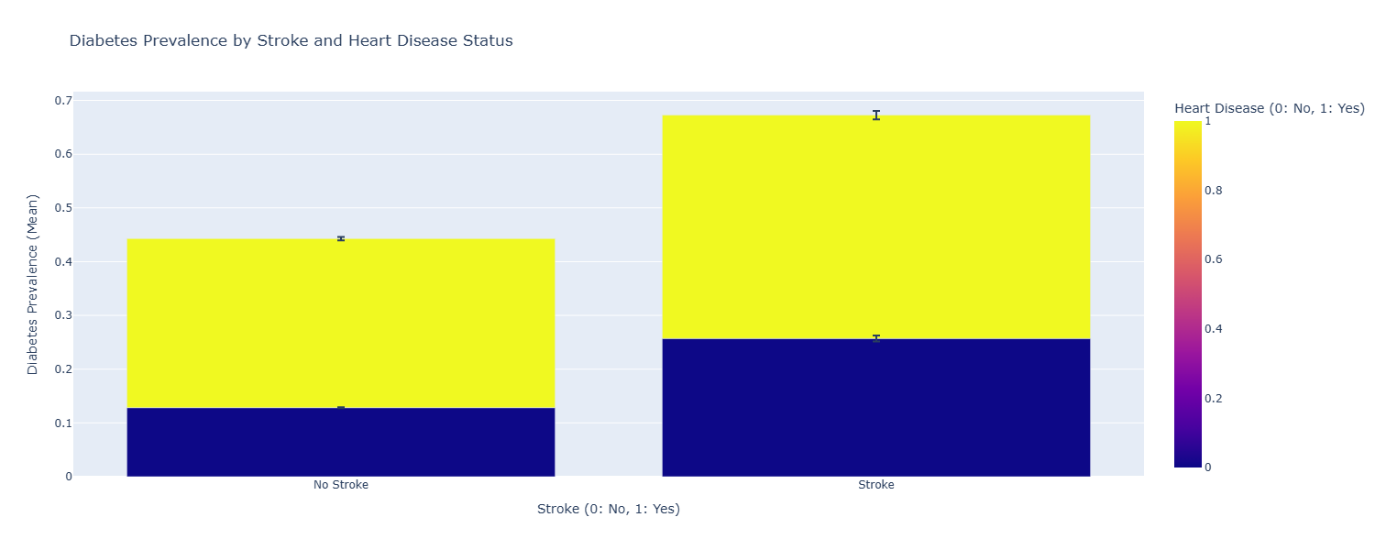
🡺 HighBP và HighChol đểu làm tăng nguy cơ xuất hiện bệnh tiểu đường, đặc biệt là khi cả 2 yếu tố này cùng xuất hiện.



Hình 3.20: Biểu đồ thể hiện mối quan hệ của Smoker và HvyAlcoholConsump với Diabetes\_binary

* Trong số những người không hút thuốc, những người không uống rượu nhiều có tỷ lệ mắc bệnh tiểu đường khoảng 14,15% (SEM ~ 0,001016), trong khi những người uống rượu nhiều có tỷ lệ mắc bệnh thấp hơn đáng kể, khoảng 4,41% (SEM ~ 0,002958).
* Tương tự, trong số những người hút thuốc, những người không uống rượu nhiều có tỷ lệ mắc bệnh tiểu đường khoảng 18,01% (SEM ~ 0,001229), trong khi những người uống rượu nhiều có tỷ lệ mắc bệnh thấp hơn, khoảng 6,78% (SEM ~ 0,002633).

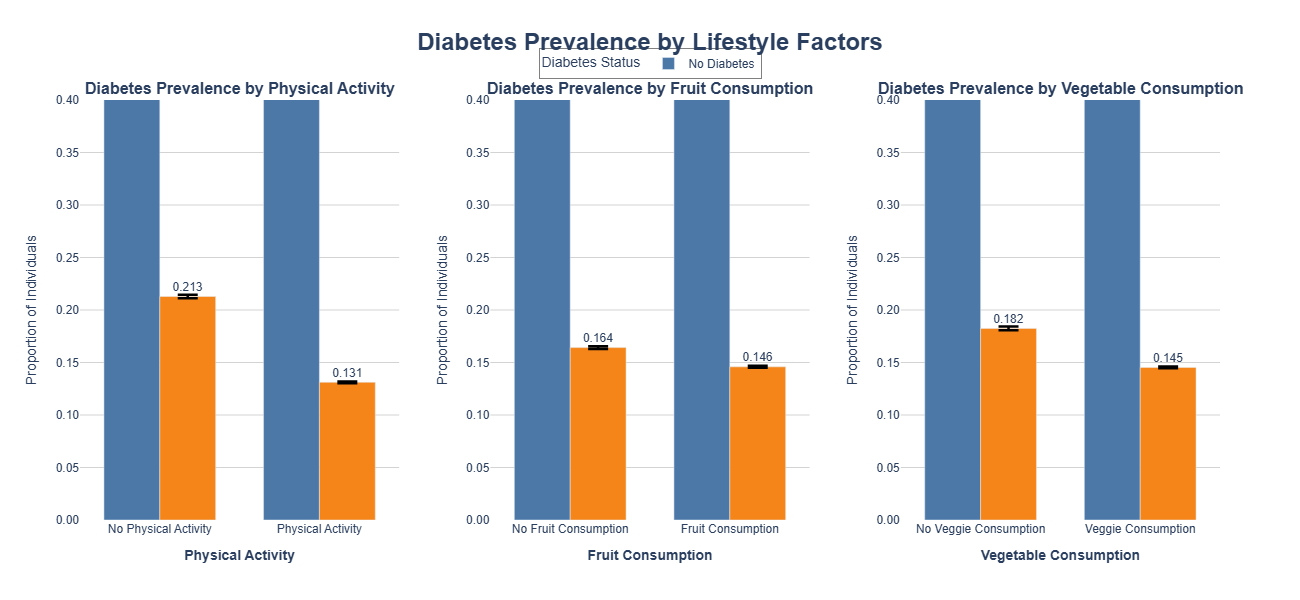
🡪 Thói quen hút thuốc và uống rượu có sự liên hệ nhất định đến bệnh tiểu đường, tuy nhiên mối tương quan chưa mạnh.



Hình 3.21: Biểu đồ thể hiện mối quan hệ của Stroke và HeartDiseaseorAttack với Diabetes\_binary

* Những người không bị đột quỵ và không mắc bệnh tim có tỷ lệ mắc bệnh khoảng 0,1284.
* Những người không bị đột quỵ nhưng mắc bệnh tim có tỷ lệ mắc bệnh khoảng 0,2872.
* Những người bị đột quỵ nhưng không mắc bệnh tim có tỷ lệ mắc bệnh khoảng 0,2312.
* Những người vừa bị đột quỵ vừa mắc bệnh tim có tỷ lệ mắc bệnh cao nhất, khoảng 0,4157.

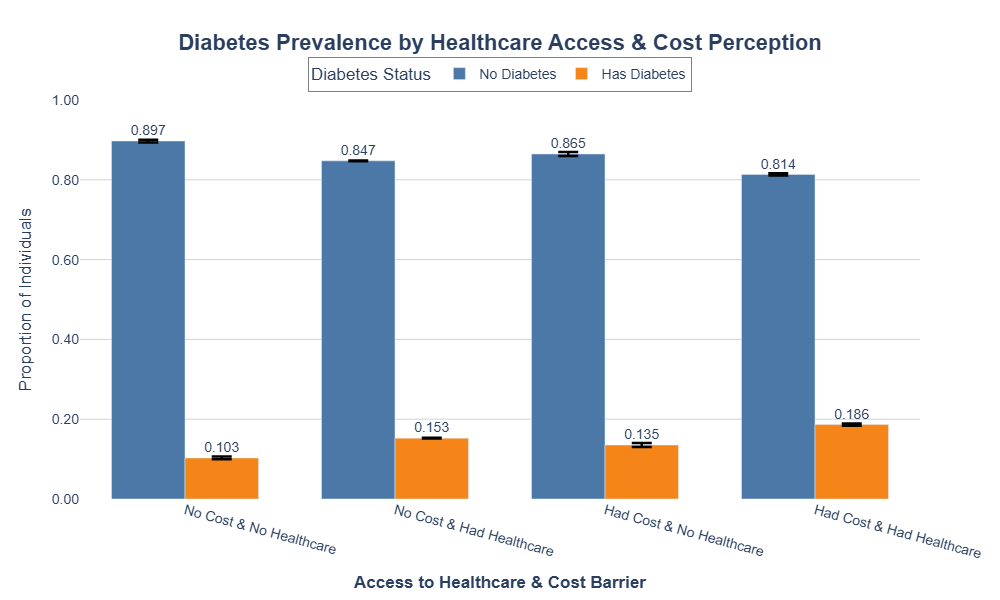
🡪 Đột quỵ và bệnh tim (cơn đau tim) đều làm tăng đáng kể nguy cơ mắc bệnh tiểu đường, đặc biệt khi cả 2 yếu tố này cùng đồng thời.



Hình 3.22: Biểu đồ cột đôi thể hiện mối quan hệ của PhysActivity, Fruits và Veggies với Diabetes\_binary

* Những người không hoạt động thể chất có tỷ lệ mắc bệnh tiểu đường cao hơn (khoảng 21,3%) so với những người hoạt động thể chất (khoảng 13,1%).
* Việc tiêu thụ trái cây có liên quan đến tỷ lệ mắc bệnh tiểu đường thấp hơn: 14,6% đối với những người ăn trái cây so với 16,4% đối với những người không ăn trái cây.
* Tương tự, việc tiêu thụ rau có tương quan với tỷ lệ mắc bệnh tiểu đường thấp hơn: 14,5% đối với những người ăn rau so với 18,2% đối với những người không ăn.

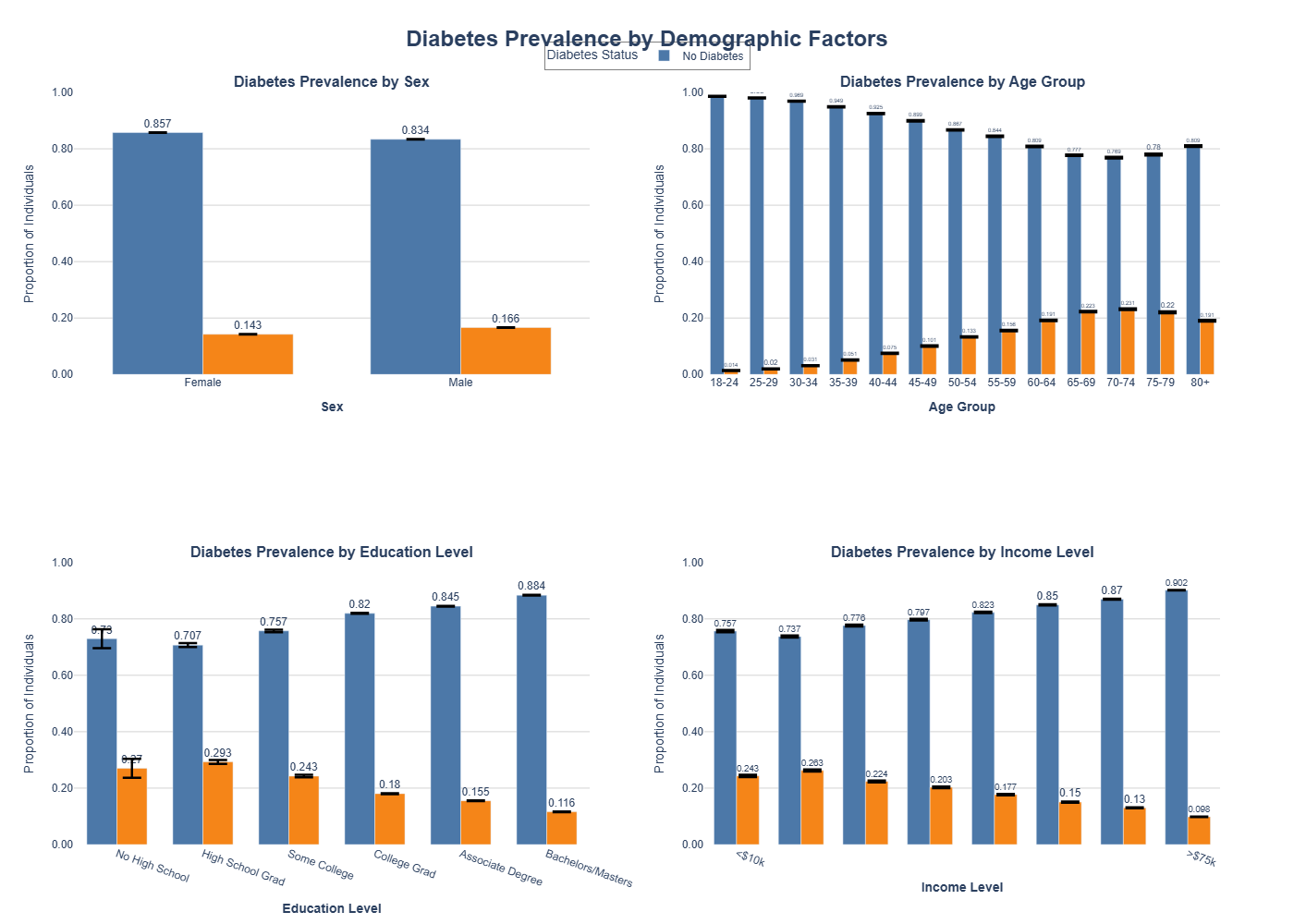
🡪 Việc duy trì hoạt động thể chất và tiêu thụ đủ trái cây và rau củ có thể là những yếu tố quan trọng trong việc giảm nguy cơ mắc bệnh tiểu đường.



Hình 3.23: Biểu đồ thể hiện mối quan hệ của NoDocbcCost và AnyHealthcare với Diabetes\_binary

* Tỷ lệ mắc bệnh tiểu đường thấp nhất (khoảng 10,3%) được ghi nhận ở những người báo cáo không gặp rào cản về chi phí khi đi khám bác sĩ và không được tiếp cận bất kỳ dịch vụ bảo hiểm chăm sóc sức khỏe nào.
* Tỷ lệ mắc bệnh tiểu đường cao nhất (khoảng 18,6%) được ghi nhận ở những người báo cáo gặp rào cản về chi phí khi đi khám bác sĩ nhưng có một số hình thức  bảo hiểm chăm sóc sức khỏe.

🡪 Việc duy trì hoạt động sử dụng dịch vụ bảo hiểm chăm sóc sức khỏe và không gặp rào cản về chi phí là những yếu tố quan trọng trong giúp giảm nguy cơ tiểu đường.



Hình 3.24: Biểu đồ tổng hợp 4 phần thể hiện tỷ lệ mắc bệnh tiểu đường theo các yếu tố nhân khẩu học (Sex, Age, Education và Income)

* Theo giới tính (Sex):
* Không có sự khác biệt lắm, cả nam và nữ đều có tỷ lệ mắc bệnh tương tự.
* Theo nhóm tuổi (Age Group):
* Tỷ lệ mắc tiểu đường tăng rất rõ theo tuổi tác.
* Nhóm tuổi trẻ (18 – 44 tuổi): rất thấp (chỉ 2 – 5%).
* Từ 45 tuổi trở lên bắt đầu tăng mạnh, đạt đỉnh ở nhóm 55–59 tuổi (khoảng 22%).
* Sau 60 tuổi vẫn duy trì ở mức cao (18 – 22%), nhóm 80+ giảm nhẹ xuống còn ~18%.

🡪 Tuổi tác là yếu tố nguy cơ mạnh nhất của bệnh tiểu đường loại 2.

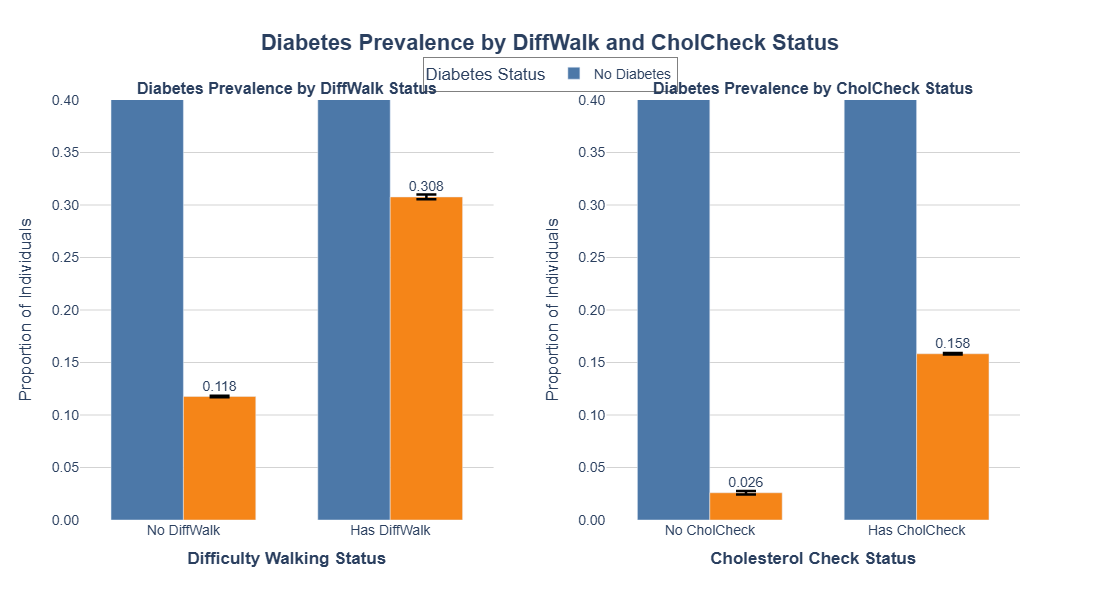
* Theo trình độ học vấn (Education Level):
* Người có trình độ học vấn càng thấp thì tỷ lệ mắc tiểu đường càng cao.
* Không tốt nghiệp phổ thông: ~27%
* Tốt nghiệp phổ thông: ~24,3%
* Một phần đại học: ~24,3%
* Tốt nghiệp đại học trở lên: chỉ còn 11,6 – 15,5%

🡪 Người ít học có nguy cơ mắc tiểu đường cao gấp đôi so với người có bằng đại học/cử nhân.

* Theo mức thu nhập (Income Level):
* Thu nhập càng thấp thì tỷ lệ mắc tiểu đường càng cao.
* Nhóm thu nhập <10k/năm: 24,3%
* Nhóm 10 – 25k: ~22–23%
* Nhóm thu nhập cao nhất (>75k/năm): chỉ còn 8,8%

🡪 Có mối liên hệ ngược rất rõ ràng giữa thu nhập và tỷ lệ mắc tiểu đường: nghèo 🡪 nguy cơ cao gấp gần 3 lần so với người giàu.

* Bệnh tiểu đường phổ biến nhất ở những người: lớn tuổi, nam giới, trình độ học vấn thấp, và đặc biệt là thu nhập thấp.
* Yếu tố kinh tế - xã hội (thu nhập và học vấn) có ảnh hưởng cực kỳ mạnh đến tỷ lệ mắc bệnh tiểu đường, thậm chí còn rõ rệt hơn cả yếu tố giới tính.
* Tuổi tác vẫn là yếu tố nguy cơ quan trọng nhất, nhưng trong cùng độ tuổi, người nghèo và ít học vẫn có nguy cơ cao hơn rất nhiều.



Hình 3.25: Biểu đồ cột đôi thể hiện mối quan hệ của DiffWalk và CholCheck với Diabetes\_binary

* Khó khăn khi đi bộ (DiffWalk = 1) có liên quan mạnh với tỷ lệ mắc bệnh tiểu đường (30,8% so với 11,8%), phản ánh tình trạng suy giảm chức năng vận động thường gặp ở bệnh nhân tiểu đường đã có biến chứng hoặc các yếu tố nguy cơ nặng.
* Tỷ lệ tiểu đường ở nhóm đã kiểm tra cholesterol cao hơn rất nhiều (15,8% so với 2,6%) chủ yếu do nhóm này chứa phần lớn người lớn tuổi và có nguy cơ tim mạch cao, chứ không chứng tỏ việc kiểm tra cholesterol làm tăng nguy cơ tiểu đường. Đây là hiện tượng nhiễu do chỉ định khám bệnh (confounding by indication).

## 3.3. Xử lý giá trị ngoại lai

* Ba biến liên tục BMI, PhysHlth, MentHlth có phân phối lệch phải mạnh (nhiều giá trị cực đại hợp lý về mặt y khoa như số ngày sức khỏe tâm thần/khỏe thể chất xấu = 30).
* Việc loại bỏ outlier quá mạnh bằng IQR thuần túy sẽ làm mất quá nhiều thông tin quan trọng.
* Do đó cần áp dụng các phương pháp giữ lại tối đa thông tin mà vẫn kiểm soát được giá trị ngoại lai cực đoan.
* Ở đây, sẽ sử dụng 3 phương pháp để thử nghiệm và chọn ra cách tối ưu:
* Xử lý bằng IQR:
* IQR = Q3 - Q1, trong đó Q1 là phân vị thứ nhất (25%), và Q3 là phân vị thứ ba (75%).
* Dữ liệu nằm ngoài khoảng [Q1 - 1.5 \* IQR, Q3 + 1.5 \* IQR] được coi là ngoại lai và sẽ bị loại bỏ.
* Xử lý bằng Winsorization:
* Thay thế các giá trị ngoại lai bằng giá trị gần nhất trong khoảng phân vị cụ thể.
* Do các biến tồn tại ngoại lai trên đều phân phối lệch về bên trái, các giá trị ngoại lai tập trung nhiều ở các giá trị lớn (bên phải) của biến. Việc chọn thay thế về mốc dữ liệu 95% sẽ giúp xử lý được các giá trị ngoại lai mà không loại bỏ điểm dữ liệu, giữ được những thông tin ảnh hưởng cần thiết của biến.
* Xử lý bằng Winsorization + IQR: kết hợp cả 2 phương pháp trên.
* Kết quả của 3 phương pháp trên:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Phương pháp** | **Mô tả** | **Số dòng bị xóa** | **% dữ liệu bị xóa** | **Số giá trị bị thay đổi** | **Kích thước cuối** |
| 1. Chỉ IQR | Xóa toàn bộ dòng nếu bất kỳ biến nào vượt ngưỡng IQR × 1.5 | 59.728 | 26.03% | – | 169.746 |
| 2.Winsorization 5% mỗi bên | Giới hạn 5% dưới và 5% trên của mỗi biến | 0 | 0.00% | 18.129 giá trị | 229.474 |
| 3.Winsorization 5% + IQR | Winsorize trước 🡪 sau đó áp dụng IQR để xóa các outlier còn lại | 56.613 | 24.67% | ~18.129 (trước khi xóa) | 172.861 |

Bảng 3.4: Bảng so sánh kết quả 3 phương pháp xử lý giá trị ngoại lai

* Đánh giá và chọn phương pháp:

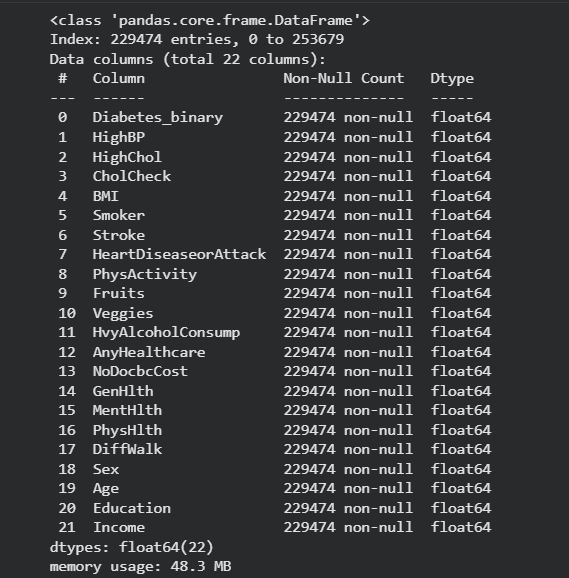
|  |  |  |  |
| --- | --- | --- | --- |
| **Tiêu chí** | **Phương pháp 1** | **Phương pháp 2** | **Phương pháp 3** |
| Giữ lại tối đa số lượng mẫu | Rất kém | Tốt nhất | Tốt |
| Kiểm soát giá trị cực đoan | Rất mạnh | Trung bình – mạnh | Rất mạnh |
| Không làm thay đổi phân phối quá nhiều | Có (xóa mất vùng đuôi) | Không (chỉ cắt đuôi) | Không |
| Phù hợp với mô hình học máy | Không khuyến khích | Rất phù hợp | Phù hợp |
| Dễ giải thích với đội ngũ y khoa | Khó (mất 1/4 bệnh nhân) | Dễ nhất | Dễ |

Bảng 3.5: Bảng đánh giá 3 phương pháp xử lý giá trị ngoại lai

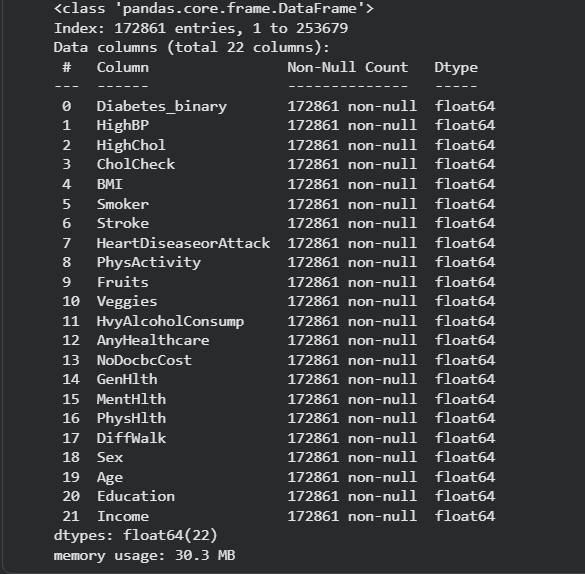
🡪 Chọn phương pháp 2 (Winsorization 5%) và phương pháp 3 (Winsorization 5% + IQR).

## 3.4. Chọn các biến quan trọng

* Tạo 2 dataframe df1 và df2 từ df ban đầu tương ứng với 2 phương pháp xử lý giá trị ngoại lai 2 và 3:

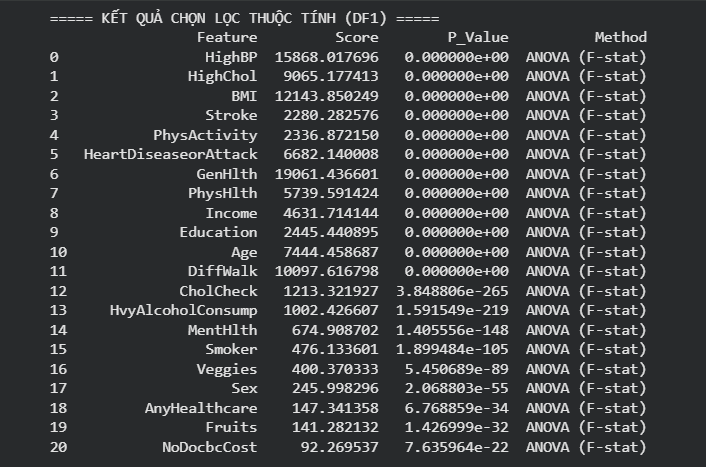


Hình 3.26: Thông tin dữ liệu df1 sau khi xử lý giá trị ngoại lai bằng phương pháp 2

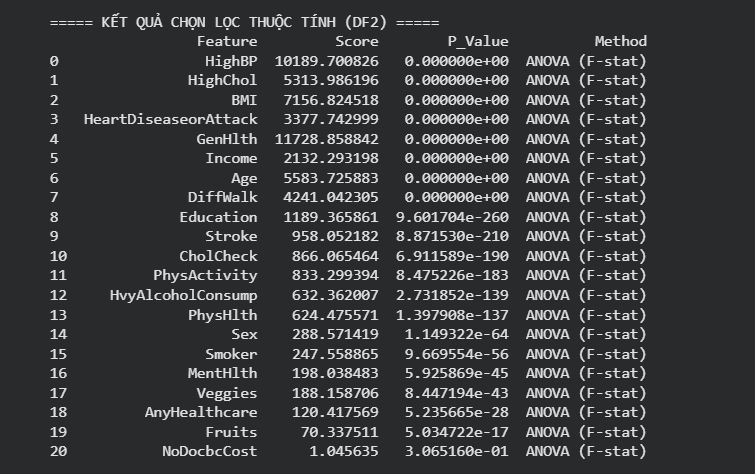


Hình 3.27: Thông tin dữ liệu df2 sau khi xử lý giá trị ngoại lai bằng phương pháp 2

* Tiến hành chọn lọc thuộc tính bằng Chi-squared và ANOVA F-Statistic cho df1 và df2:



Hình 3.28: Kết quả chọn lọc thuộc tính của df1

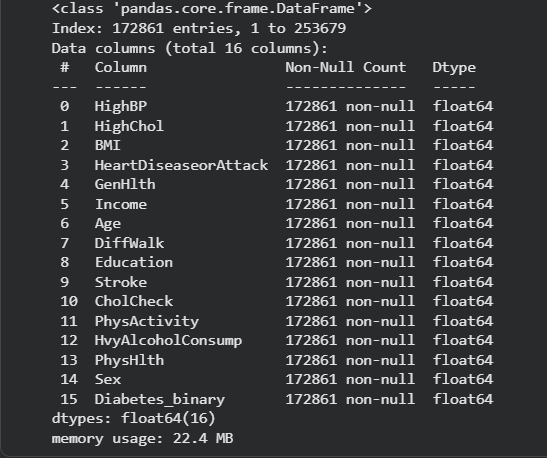


Hình 3.29: Kết quả chọn lọc thuộc tính của df2

🡪 Tương ứng với mỗi dataframe, sẽ chọn 15 thuộc tính có Score cao nhất.



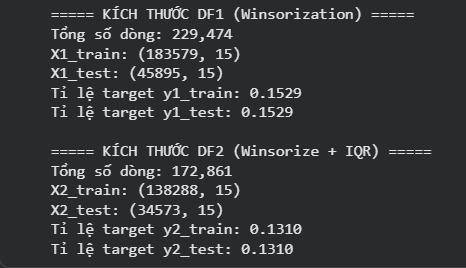
Hình 3.30: Thông tin dữ liệu của df1 sau khi chọn lọc 15 thuộc tính có Score cao nhất

**

Hình 3.31: Thông tin dữ liệu của df2 sau khi chọn lọc 15 thuộc tính có Score cao nhất

## 3.5. Chia tập train và test

* Target\_col: ‘Diabetes\_binary’.
* X: 15 biến.
* Tỉ lệ train/test: 80 – 20.



Hình 3.32: Thông tin chia tập train - test theo tỷ lệ 80 – 20

# CHƯƠNG 4. XÂY DỰNG MÔ HÌNH

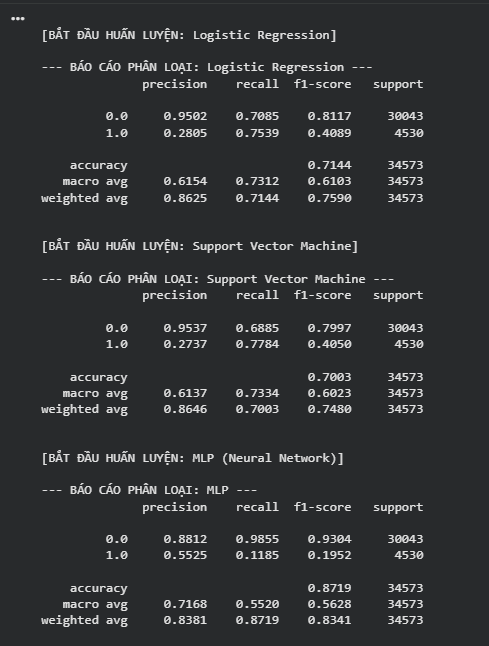
## 4.1. Sensitive Models

* Gồm các mô hình: Logistic Regression, Suport Vector Machine (SVM), Multi-Layer Perceoption (MLP).
* Dataframe được train: df2.
* Lý do chọn các mô hình trên:
* Các mô hình này dựa trên khoảng cách, tối ưu hóa hàm lỗi bằng phương pháp độ dốc, hoặc giả định phân phối dữ liệu.
* Outlier có thể kéo đường biên quyết định hoặc làm sai lệch các tham số.

|  |  |
| --- | --- |
| **Mô hình** | **Lý do** |
| Logistic Regression | Winsorization giúp ổn định các hệ số hồi quy (weights) bằng cách giới hạn ảnh hưởng của các giá trị cực đoan. |
| Support Vector Machine (SVM) | Nhạy cảm vì chúng dựa trên khoảng cách (margins). Outlier có thể ảnh hưởng đến việc tìm ra siêu phẳng phân cách tối ưu. |
| Multi-Layer Perceptron (MLP) | Các mô hình dựa trên gradient descent cũng nhạy cảm với outlier, vì chúng có thể làm tăng gradient quá mức và làm chậm hoặc làm sai lệch quá trình học. |

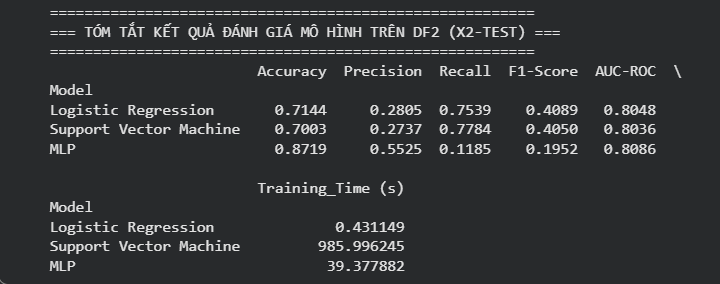
Bảng 4.1: Bảng lý do chọn 3 mô hình Logistic Regression, Support Vector Machine (SVM) và Multi-Layer Perceptron (MLP)

* Quá trình thực hiện:



Hình 4.1: Quá trình thực hiện train 3 mô hình Logistic Regression, Support Vector Machine (SVM) và Multi-Layer Perceptron (MLP)

* Kết quả đạt được:

**

Hình 4.2: Tóm tắt kết quả đánh giá mô hình trên df2

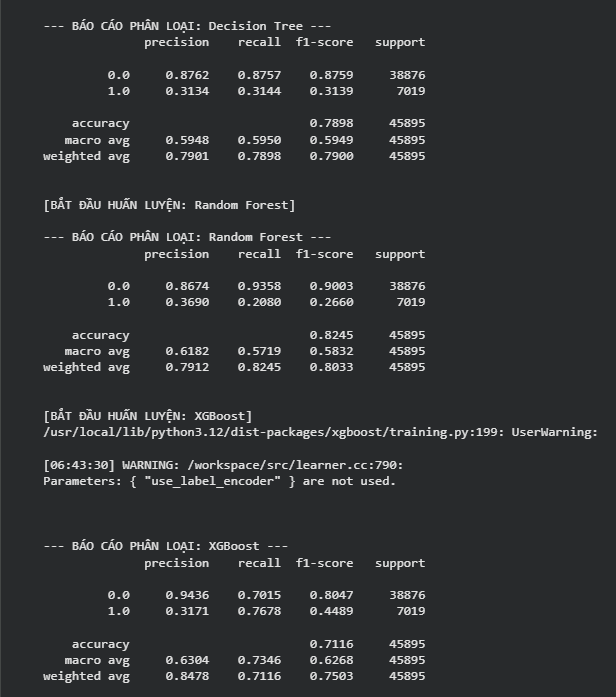
## 4.2. Robust Models

* Gồm các mô hình: Decision Tree, Random Forest, XGBoost.
* Dataframe được train: df1.
* Lý do chọn các mô hình trên:
* Các mô hình này hoạt động dựa trên việc phân chia dữ liệu theo ngưỡng (splits) chứ không phải khoảng cách.
* Outlier thường chỉ ảnh hưởng đến việc phân chia ở nút cuối cùng.

|  |  |
| --- | --- |
| **Mô hình** | **Lý do** |
| Decision Tree | Ít nhạy cảm nhất. Outlier không ảnh hưởng nhiều đến cấu trúc cây. Winsorization vẫn hữu ích để giữ lại toàn bộ dữ liệu. |
| Random Forest | Là tập hợp của các Decision Tree. Rất mạnh mẽ trước outlier. (B) là tốt nhất vì nó bảo toàn kích thước mẫu, giúp mỗi cây có nhiều dữ liệu hơn để học. |
| XGBoost | Mô hình tăng cường (Boosting) cũng là mô hình cây. (B) là lựa chọn tối ưu vì giữ lại toàn bộ dữ liệu giúp tăng cường khả năng học của mô hình. |

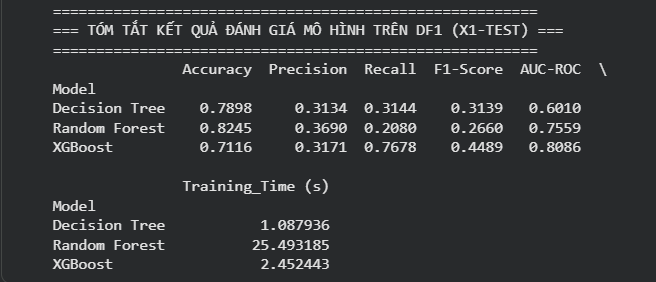
Bảng 4.2: Bảng lý do chọn 3 mô hình Decision Tree, Random Forest và XGBoost

* Quá trình thực hiện:



Hình 4.3: Quá trình thực hiện train 3 mô hình Decision Tree, Random Forest và XGBoost

* Kết quả đạt được:



Hình 4.4: Tóm tắt kết quả đánh giá mô hình trên df1

## 4.3. So sánh và lựa chọn mô hình tốt nhất

|  |  |
| --- | --- |
| **Tiêu chí** | **Mô hình tốt nhất** |
| 1. AUC-ROC (khả năng phân biệt tổng thể) | XGBoost & MLP = 0.8086 |
| 2. Recall (bắt được bao nhiêu bệnh nhân thực sự có nguy cơ) | SVM 0.778 🡪 XGBoost 0.768 🡪 Logistic 0.754 |
| 3. F1-Score (cân bằng Precision–Recall) | XGBoost 0.4489 |
| 4. Không được có Recall quá thấp | MLP (0.1185) 🡪 loại ngay |
| 5. Training time hợp lý | XGBoost 2.5 giây |
| 6. Dễ tuning và ổn định khi dữ liệu mới | XGBoost |

Bảng 4.3: Bảng đánh giá các mô hình

🡪 Chọn XGBoost vì:

* AUC cao nhất hoặc ngang bằng tất cả mô hình.
* Không bỏ lỡ quá nhiều bệnh nhân nguy cơ cao.
* MLP chỉ 0.195, RF chỉ 0.266 🡪 XGBoost hơn hẳn.
* MLP dù Accuracy cao nhưng không có ý nghĩa trong y tế.
* SVM 1000 giây, MLP 39 giây, RF 25 giây 🡪 XGBoost nhanh nhất trong nhóm mạnh.
* XGBoost cực kỳ ổn định sau khi đã tune.

# CHƯƠNG 5. KẾT QUẢ VÀ KIẾN NGHỊ

**5.1. Kết quả đạt được**

### ****5.1.1. Hoàn thiện mô hình dự đoán bệnh tiểu đường****

* Nhóm đã xây dựng và tối ưu hóa mô hình dự đoán nguy cơ mắc bệnh tiểu đường dựa trên bộ dữ liệu BRFSS 2015. Quy trình bao gồm tiền xử lý dữ liệu, lựa chọn đặc trưng, thử nghiệm nhiều thuật toán và đánh giá bằng các chỉ số Accuracy, Precision, Recall, F1-score và AUC - ROC. Kết quả cuối cùng cho thấy mô hình đạt độ chính xác cao và ổn định, đáp ứng được yêu cầu đặt ra ban đầu.

### 5.1.2. Xây dựng giao diện người dùng (UX/UI) hoàn chỉnh và trực quan

* Hệ thống được thiết kế với giao diện thân thiện, dễ sử dụng và có khả năng hiển thị kết quả dự đoán rõ ràng. Người dùng có thể nhập các thông tin sức khỏe cần thiết, theo dõi kết quả mô hình cũng như nhận được những gợi ý về tình trạng sức khỏe. UX/UI được tối ưu theo hướng đơn giản – trực quan giúp tăng trải nghiệm người dùng.

### 5.1.3. Triển khai hệ thống backend ổn định và đảm bảo tính mở rộng

* Backend được xây dựng bằng FastAPI, đảm bảo tốc độ phản hồi nhanh, có khả năng mở rộng khi số lượng người dùng tăng. Hệ thống API vận hành ổn định, tổ chức rõ ràng theo mô hình RESTful và hỗ trợ triển khai mô hình dự đoán trong môi trường thực tế.

## ****5.2. Đánh giá chung****

* Qua quá trình triển khai, hệ thống đã đạt được các mục tiêu ban đầu đề ra: xây dựng được mô hình dự đoán bệnh tiểu đường hiệu quả, phát triển giao diện trực quan và hoàn thiện quy trình vận hành backend - frontend. Mặc dù mô hình hoạt động tốt trên dữ liệu thử nghiệm, nhưng hiệu quả thực tế vẫn có thể thay đổi khi triển khai trên dữ liệu đa dạng và cập nhật hơn. Nhóm đánh giá sản phẩm hiện tại có tính thực tiễn cao và có thể mở rộng để sử dụng trong các hệ thống hỗ trợ chăm sóc sức khỏe cộng đồng.

## ****5.3. Kiến nghị phát triển****

* Để nâng cao chất lượng hệ thống trong tương lai, nhóm đề xuất:
* Mở rộng bộ dữ liệu bằng cách cập nhật thêm dữ liệu từ các năm hoặc từ các nguồn y tế khác nhau để tăng độ bao phủ của mô hình.
* Phát triển thêm tính năng theo dõi sức khỏe dài hạn, cho phép người dùng lưu lịch sử và xem xu hướng thay đổi nguy cơ theo thời gian.
* Tối ưu hiệu năng giao diện và API, hướng đến khả năng triển khai trên môi trường cloud thực tế.
* Tăng cường bảo mật dữ liệu, đảm bảo thông tin sức khỏe của người dùng được bảo vệ tuyệt đối.

## ****5.4. Kết luận****

* Đề tài đã hoàn thành đầy đủ các mục tiêu nghiên cứu ban đầu: xây dựng một hệ thống dự đoán nguy cơ mắc bệnh tiểu đường dựa trên dữ liệu sức khỏe; triển khai mô hình lên giao diện thân thiện với người dùng; và xây dựng backend vững chắc, có khả năng mở rộng trong tương lai. Sản phẩm cuối cùng không chỉ mang tính ứng dụng thực tiễn cao mà còn mở ra nhiều hướng phát triển mới cho các hệ thống hỗ trợ chăm sóc sức khỏe thông minh. Kết quả đạt được là nền tảng quan trọng để tiếp tục nghiên cứu và hoàn thiện trong các giai đoạn tiếp theo.

# TÀI LIỆU THAM KHẢO

|  |  |
| --- | --- |
| [1] | “Building Risk Prediction Models for Type 2 Diabetes Using Machine Learning Techniques,” 19 September 2019. [Trực tuyến]. Available: https://www.cdc.gov/pcd/issues/2019/19\_0109.htm. [Đã truy cập 1 October 2025]. |
| [2] | Y. Krisna, “Diabetes Risk Prediction Using BRFSS 2015 Data,” 6 July 2024. [Trực tuyến]. Available: https://rpubs.com/KiriYuuki/Diabetes-Risk-Prediction-Using-BRFSS-2015-Data. [Đã truy cập 15 October 2025]. |
| [3] | T. D. School, “Top 5 thuật toán phân loại (Classification) trong Machine learning,” 27 May 2024. [Trực tuyến]. Available: https://blog.tomorrowmarketers.org/thuat-toan-phan-loai-classification-machine-learning/. [Đã truy cập 5 November 2025]. |
| [4] | T. Đức, “Đánh giá các mô hình học máy,” 28 September 2021. [Trực tuyến]. Available: https://viblo.asia/p/danh-gia-cac-mo-hinh-hoc-may-RnB5pp4D5PG. [Đã truy cập 5 November 2025]. |
| [5] | P. D. Khanh, “Thước đo mô hình phân loại,” 2021. [Trực tuyến]. Available: https://phamdinhkhanh.github.io/deepai-book/ch\_ml/modelMetric.html. [Đã truy cập 5 November 2025]. |
| [6] | T. t. y. t. K. v. C. Lớn, “Bệnh đái tháo đường ở Việt Nam ngày càng trẻ hoá, nhiều cháu 14 tuổi đã mắc bệnh,” 17 November 2024. [Trực tuyến]. Available: https://bvquan5.medinet.gov.vn/benh-dai-thao-duong/benh-dai-thao-duong-o-viet-nam-ngay-cang-tre-hoa-nhieu-chau-14-tuoi-da-mac-benh-cmobile17005-196576.aspx. [Đã truy cập 7 October 2025]. |
| [7] | B. Y. Tế, “Việt Nam hiện tỷ lệ người mắc bệnh đái tháo đường đang gia tăng nhanh,” 12 November 2023. [Trực tuyến]. Available: https://moh.gov.vn/tin-noi-bat/-/asset\_publisher/3Yst7YhbkA5j/content/viet-nam-hien-ty-le-nguoi-mac-benh-ai-thao-uong-ang-gia-tang-nhanh. [Đã truy cập 8 October 2025]. |