   
BỘ GIÁO DỤC VÀ ĐÀO TẠO

**TRƯỜNG ĐẠI HỌC SƯ PHẠM KỸ THUẬT TP. HỒ CHÍ MINH**

**KHOA CÔNG NGHỆ THÔNG TIN**

****

**MÔN NHẬP MÔN KHOA HỌC DỮ LIỆU**

**HỌC KÌ II/2023-2024**

**BÁO CÁO CUỐI KỲ**

**ĐỀ TÀI: XÂY DỰNG MÔ HÌNH DỰ ĐOÁN BỆNH TIỂU ĐƯỜNG**

**GVHD: ThS. Lê Minh Tân**

**Lớp HP: INDS331085**

**Học kỳ: 2**

**Năm học: 2023 – 2024**

**Thành viên nhóm:**

1. **Nguyễn Hoàng Việt Pháp 21110587**
2. **Cao Thị Thu Thủy 21110670**
3. **Lê Trạc Tiến 21110677**

*Thành phố Thủ Đức, tháng 5 năm 2024*

**MỤC LỤC**

[I. PHẦN GIỮA KỲ 1](#_Toc165327499)

[1. Tên dự án: Xây dựng mô hình dự đoán bệnh tiểu đường 1](#_Toc165327500)

[2. Mô tả ngắn dự án (10-30 sentences) 1](#_Toc165327501)

[2.1. Dự án về vấn đề gì? 1](#_Toc165327502)

[2.2. Phạm vi dự án 1](#_Toc165327503)

[2.3. Thu thập dữ liệu 1](#_Toc165327504)

[2.3.1. Tập dữ liệu bệnh tiểu đường tại US 1](#_Toc165327505)

[2.3.2. Tập dữ liệu bệnh tiểu đường của phụ nữ Ấn Độ 2](#_Toc165327506)

[2.3.3. Tập dữ liệu bệnh tiểu đường tại IRAQ 2](#_Toc165327507)

[3. Chosen datasets 3](#_Toc165327508)

[3.1. Tập dữ liệu bệnh tiểu đường tại US 3](#_Toc165327509)

[3.2. Tập dữ liệu bệnh tiểu đường của phụ nữ gốc Ấn Độ Pima 3](#_Toc165327510)

[3.3. Tập dữ liệu bệnh tiểu đường tại IRAQ 3](#_Toc165327511)

[4. Giới thiệu về datasets 3](#_Toc165327512)

[4.1. Tập dữ liệu bệnh tiểu đường tại US 3](#_Toc165327513)

[4.1.1. Tập mẫu 3](#_Toc165327514)

[4.1.2. Thuộc tính 3](#_Toc165327515)

[4.1.3. Thống kê 7](#_Toc165327516)

[4.1.4. Đánh giá chất lượng 10](#_Toc165327517)

[4.2. Tập dữ liệu bệnh tiểu đường của phụ nữ gốc Ấn Độ Pima 10](#_Toc165327518)

[4.2.1. Tập mẫu 10](#_Toc165327519)

[4.2.2. Thuộc tính 10](#_Toc165327520)

[4.2.3. Thống kê 12](#_Toc165327521)

[4.2.4. Đánh giá chất lượng 14](#_Toc165327522)

[4.3. Tập dữ liệu bệnh tiểu đường tại IRAQ 15](#_Toc165327523)

[4.3.1. Tập mẫu 15](#_Toc165327524)

[4.3.2. Thuộc tính 15](#_Toc165327525)

[4.3.3. Thống kê 16](#_Toc165327526)

[4.3.4. Đánh giá chất lượng 19](#_Toc165327527)

[5. So sánh 19](#_Toc165327528)

[6. Kết luận 21](#_Toc165327529)

[II. PHẦN CUỐI KỲ 22](#_Toc165327530)

[1. Thông tin chi tiết về dữ liệu 22](#_Toc165327531)

[1.1. Tập dữ liệu bệnh tiểu đường tại US 22](#_Toc165327532)

[1.2. Tập mẫu 22](#_Toc165327533)

[1.3. Thuộc tính 22](#_Toc165327534)

[2. So sánh và kết luận sâu 26](#_Toc165327535)

[2.1. Bệnh tiểu đường là gì? Nguyên nhân. Tỉ lệ người trưởng thành mắc bệnh trong tập dữ liệu thể hiện như thế nào? 26](#_Toc165327536)

[2.2. Các thuộc tính có mối tương quan với bệnh tiểu đường 27](#_Toc165327537)

[2.3. Hoạt động về thể chất, thói quen ăn uống, lối sống không lành mạnh có ảnh hưởng thế nào đến nguy cơ mắc bệnh tiểu đường 31](#_Toc165327538)

[2.4. Độ tuổi càng cao thì càng dễ mắc bệnh tiểu đường đúng không? So sánh với thực trạng tại Việt Nam hiện nay 34](#_Toc165327539)

[2.5. Bệnh tiểu đường có thể là nguồn gốc cho những căn bệnh khác hay không? Dữ liệu có thể hiện được điều đó hay không? 35](#_Toc165327540)

[2.5.1. Tăng huyết áp 35](#_Toc165327541)

[2.5.2. Bệnh đau tim 36](#_Toc165327542)

[2.5.3. Đột quỵ 38](#_Toc165327543)

[2.6. Việc xây dựng mô hình chẩn đoán bệnh tiểu đường thông qua tập dữ liệu trên có cho ra kết quả tốt với độ chính xác > 90% theo phương thức đánh giá độ chính xác là F1-score hoặc công thức (Số dự đoán đúng / Tổng số \* 100%) hay không? Với mô hình có độ chính xác tốt nhất có nên ứng dụng vào các phòng khám, bệnh viện hay không? 41](#_Toc165327544)

[2.6.1. Xây dựng mô hình dự đoán khi dữ liệu chưa cân bằng 42](#_Toc165327545)

[2.6.2. Xây dựng mô hình với dữ liệu được cân bằng thông qua xử lý thêm dữ liệu 48](#_Toc165327546)

[2.6.3. Xây dựng mô hình với dữ liệu được cân bằng bằng cách lựa chọn ra từ tập dữ liệu ban đầu 51](#_Toc165327547)

[2.7. Những chính sách cần đề ra để phòng tránh bệnh tiểu đường. 54](#_Toc165327548)

[3. Kết luận 55](#_Toc165327549)

[TÀI LIỆU THAM KHẢO 57](#_Toc165327550)

# PHẦN GIỮA KỲ

## Tên dự án: Xây dựng mô hình dự đoán bệnh tiểu đường

## Mô tả ngắn dự án (10-30 sentences)

### Dự án về vấn đề gì?

Dự án giúp dự đoán nguy cơ mắc bệnh tiểu đường từ việc kết hợp giữa dữ liệu lâm sàng, thông tin về lối sống, và các chỉ số sinh học sẽ được phân tích bằng các thuật toán học máy để tạo ra mô hình dự đoán chính xác.

### Phạm vi dự án

* Dự án được áp dụng tại các bệnh viện có chẩn đoán và điều trị bệnh tiểu đường.
* Dự án hỗ trợ các bác sĩ trong quá trình chẩn đoán bệnh tiểu đường, đánh giá các yếu tố nguy cơ và cung cấp dự đoán về khả năng mắc bệnh dựa trên thông tin dự đoán về bệnh tiểu đường.

### Thu thập dữ liệu

#### Tập dữ liệu bệnh tiểu đường tại US

P: Dữ liệu bệnh tiểu đường của những người cư trú tại Hoa Kỳ, của những người từ 18 tuổi trở lên.

A: Dữ liệu bệnh tiểu đường của người cư trú tại các tiểu bang, của những người từ 18 tuổi trở lên, thực hiện thông qua điện thoại.

S: Dữ liệu bệnh tiểu đường của người cư trú tại các tiểu bang, của những người từ 18 tuổi trở lên, có phản hồi từ 236.378 cá nhân.

Các yếu tố ảnh hưởng: sai số của các dụng cụ thực hiện xét nghiệm cho bệnh nhân được khảo sát đó, những người thực hiện khảo sát trả lời không trung thực, quá trình nhập liệu không chính xác.

Instrument: Sử dụng phương tiện điện thoại để thực hiện các cuộc khảo sát nhằm thu thập dữ liệu.

Protocol: Người nghe điện thoại sẽ trả các câu hỏi về thông tin về: thông tin về tình trạng tiểu đường, chỉ số áp huyết cao, chỉ số cholesterol cao, thông tin về việc kiểm tra cholesterol, chỉ số khối cơ thể, thông tin về việc hút thuốc lá, thông tin về cơn đột quỵ, thông tin về bệnh tim hoặc cơn đau tim, mức độ hoạt động thể chất, số lượng trái cây tiêu thụ hàng ngày, số lượng rau tiêu thụ hàng ngày, tiêu thụ rượu mạnh, có sử dụng dịch vụ chăm sóc sức khỏe nào không, không phải trả chi phí cho bác sĩ hoặc dịch vụ y tế, tình trạng sức khỏe tổng quát, sức khỏe tâm thần, sức khỏe về mặt thể chất, gặp khó khăn khi đi bộ, giới tính, tuổi, trình độ học vấn, thu nhập.

#### Tập dữ liệu bệnh tiểu đường của phụ nữ Ấn Độ

P: Dữ liệu sức khỏe của những người trưởng thành gốc Ấn Độ Pima.

A: Dữ liệu sức khỏe của phụ nữ gốc Ấn Độ Pima.

S: Dữ liệu sức khỏe của phụ nữ gốc Ấn Độ Pima, ít nhất 21 tuổi, có phản hồi từ 768 cá nhân.

Các yếu tố ảnh hưởng: sai số của các dụng cụ thực hiện xét nghiệm cho bệnh nhân được khảo sát đó, những người thực hiện khảo sát trả lời không trung thực, quá trình nhập liệu không chính xác.

Instrument: Sử dụng phương tiện điện thoại để thực hiện các cuộc khảo sát nhằm thu thập dữ liệu.

Protocol: Người nghe điện thoại sẽ trả các câu hỏi về số lần người phụ nữ đã mang thai, nồng độ glucose trong huyết tương, huyết áp, độ dày của nếp gấp da ở cơ tam đầu, nồng độ insulin trong máu, chỉ số BMI, một chức năng cho thấy khả năng mắc bệnh tiểu đường dựa trên tiền sử gia đình, tuổi của người phụ nữ.

#### Tập dữ liệu bệnh tiểu đường tại IRAQ

P: Dữ liệu bệnh tiểu đường của những người kiểm tra y tế về bệnh tiểu đường tại IRAQ.

A: Dữ liệu bệnh tiểu đường được thu thập tại Bệnh viện Thành Phố lớn.

S: Dữ liệu bệnh tiểu đường tại Bệnh viện Y Tế Thành phố và Trung tâm Chuyên khoa Nội tiết và Bệnh viện Giảng dạy Tiểu đường-Al-Kindy, từ 1000 cá nhân.

Các yếu tố ảnh hưởng: sai số của các dụng cụ thực hiện xét nghiệm cho bệnh nhân được khảo sát đó, những người thực hiện khảo sát trả lời không trung thực, quá trình nhập liệu không chính xác.

Instrument: Sử dụng phương tiện điện thoại để thực hiện các cuộc khảo sát nhằm thu thập dữ liệu.

Protocol: Người nghe điện thoại sẽ trả các câu hỏi về lượng đường trong máu, tuổi, giới tính, tỷ lệ Creatinin (Cr), chỉ số khối cơ thể (BMI), chỉ số mỡ máu, ure, cholesterol (Chol), hồ sơ lipid lúc đói, HBA1C.

## Chosen datasets

### Tập dữ liệu bệnh tiểu đường tại US

Name: Diabetes Health Indicators Dataset

File name: diabetes\_binary\_health\_indicators\_BRFSS2021.csv

Author(s): JULLIEN NAZREEN

Downloadable

Links: [Diabetes Health Indicators Dataset (kaggle.com)](https://www.kaggle.com/datasets/julnazz/diabetes-health-indicators-dataset/data?select=diabetes_binary_health_indicators_BRFSS2021.csv)

### Tập dữ liệu bệnh tiểu đường của phụ nữ gốc Ấn Độ Pima

Name: Diabetes Prediction Dataset

File name: diabetes.csv

Author(s): PENTA KRISHNA KISHORE

Downloadable links:

[Diabetes Prediction Dataset (kaggle.com)](https://www.kaggle.com/datasets/pentakrishnakishore/diabetes-csv/data?select=diabetes.csv)

### Tập dữ liệu bệnh tiểu đường tại IRAQ

Name: Diabetes Dataset

File name: Dataset of Diabetes.csv

Author(s): ARAVINDPCODER

Downloadable links: [Diabetes Dataset (kaggle.com)](https://www.kaggle.com/datasets/aravindpcoder/diabetes-dataset?select=Dataset+of+Diabetes+.csv)

## Giới thiệu về datasets

### Tập dữ liệu bệnh tiểu đường tại US

#### Tập mẫu

* Bộ dữ liệu "diabetes\_binary\_health\_indicators\_BRFSS2021.csv là một tập dữ liệu rõ ràng gồm 236.378 câu trả lời khảo sát. Biến mục tiêu Diabetes\_binary có 2 lớp. 0 là không mắc bệnh tiểu đường và 1 là tiền tiểu đường hoặc tiểu đường. Tập dữ liệu này có 21 biến đặc trưng và không cân bằng.

#### Thuộc tính

* Diabetes\_binary (Tiểu đường), nhận giá trị 0 hoặc 1:

0: Không mắc bệnh tiểu đường

1: Tiền tiểu đường hoặc đang mắc bệnh tiểu đường

* HighBP (Huyết áp cao), nhận giá trị 0 hoặc 1:

0: Không có huyết áp cao

1: Có huyết áp cao

* HighChol (Cholesterol cao), nhận giá trị 0 hoặc 1:

0: Không có cholesterol cao

1: Có cholesterol cao

* CholCheck (Kiểm tra cholesterol), nhận giá trị 0 hoặc 1:

0: Không kiểm tra cholesterol trong vòng 5 năm

1: Đã kiểm tra cholesterol trong vòng 5 năm

* BMI (Chỉ số khối cơ thể), nhận các giá trị lớn hơn 0, chỉ số này đo lường sự tương quan giữa cân nặng và chiều cao của một người.
* Smoker (Hút thuốc), nhận giá trị 0 hoặc 1:

0: Không hút thuốc

1: Đã hút ít nhất 100 điếu thuốc trong cả đời

* Stroke (Tai biến), nhận giá trị 0 hoặc 1:

0: Không bị tai biến

1: Đã bị tai biến

* HeartDiseaseorAttack (Bệnh tim mạch hoặc đau tim), nhận giá trị 0 hoặc 1:

0: Không mắc bệnh tim mạch hoặc đau tim

1: Đã mắc bệnh tim mạch hoặc đã trải qua đau tim

* PhysActivity (Hoạt động vận động), nhận giá trị 0 hoặc 1:

0: Không tham gia hoạt động vận động trong 30 ngày qua (không tính công việc)

1: Tham gia hoạt động vận động trong 30 ngày qua (không tính công việc)

* Fruits (Tiêu thụ trái cây), nhận giá trị 0 hoặc 1:

0: Không tiêu thụ ít nhất 1 loại trái cây mỗi ngày

1: Tiêu thụ ít nhất 1 loại trái cây mỗi ngày

* Veggies (Tiêu thụ rau củ):

0: Không tiêu thụ ít nhất 1 loại rau củ mỗi ngày

1: Tiêu thụ ít nhất 1 loại rau củ mỗi ngày

* HvyAlcoholConsump (Tiêu thụ rượu nặng):

0: Không tiêu thụ rượu nặng (nam giới uống hơn 14 đơn vị mỗi tuần, nữ giới uống hơn 7 đơn vị mỗi tuần)

1: Tiêu thụ rượu nặng

* AnyHealthcare (Bảo hiểm y tế):

0: Không có bảo hiểm y tế hoặc bất kỳ hình thức bảo hiểm nào khác

1: Có bảo hiểm y tế hoặc hình thức bảo hiểm khác

* NoDocbcCost (Không đến bác sĩ do chi phí):

0: Không có thời điểm trong 12 tháng qua cần đến bác sĩ mà không thể do chi phí

1: Có thời điểm trong 12 tháng qua cần đến bác sĩ mà không thể do chi phí

* GenHlth (Tình trạng sức khỏe tổng quát), đánh giá trạng thái sức khỏe tổng quát trên thang đánh giá từ 1 đến 5:

1: Xuất sắc

2: Rất tốt

3: Tốt

4: Trung bình

5: Kém

* MentHlth (Tình trạng sức khỏe tâm thần), số ngày trong vòng 30 ngày qua mà tâm trạng không tốt, bao gồm căng thẳng, trầm cảm và vấn đề về cảm xúc.
* PhysHlth (Tình trạng sức khỏe vật lý), số ngày trong vòng 30 ngày qua mà sức khỏe vật lý không tốt, bao gồm bệnh tật và chấn thương.
* DiffWalk (Khó khăn khi đi bộ hoặc leo cầu thang) , nhận giá trị 0 hoặc 1:

0: Không gặp khó khăn

1: Gặp khó khăn

* Sex (Giới tính), nhận giá trị 0 hoặc 1:

0: Nữ

1: Nam

* Age (Độ tuổi), sử dụng thang đo từ 1 đến 13, mô tả các khoảng độ tuổi, ví dụ:

1: 18-24 tuổi

2: 25-29 tuổi

3: 30-34 tuổi

4: 35-39 tuổi

5: 40-44 tuổi

6: 45-49 tuổi

7: 50-54 tuổi

8: 55-59 tuổi

9: 60-64 tuổi

10: 65-69 tuổi

11: 70-74 tuổi

12: 75-79 tuổi

13: 80 tuổi trở lên

* Education (Trình độ học vấn), sử dụng thang đo từ 1 đến 6, mô tả trình độ học vấn:

1: Không đi học hoặc chỉ học mẫu giáo

2: Học tiểu học

3: Học trung học cơ sở

4: Học trung học phổ thông

5: Có bằng đại học

6: Có bằng cao học

* Income (Thu nhập), sử dụng thang đo từ 1 đến 11, mô tả mức thu nhập:

1: Dưới 10.000 đô la

2: Dưới 10.000 đô la

3: Dưới 10.000 đô la

4: Dưới 10.000 đô la

5: Dưới 35.000 đô la

6: Dưới 50.000 đô la

7: Dưới 75.000 đô la

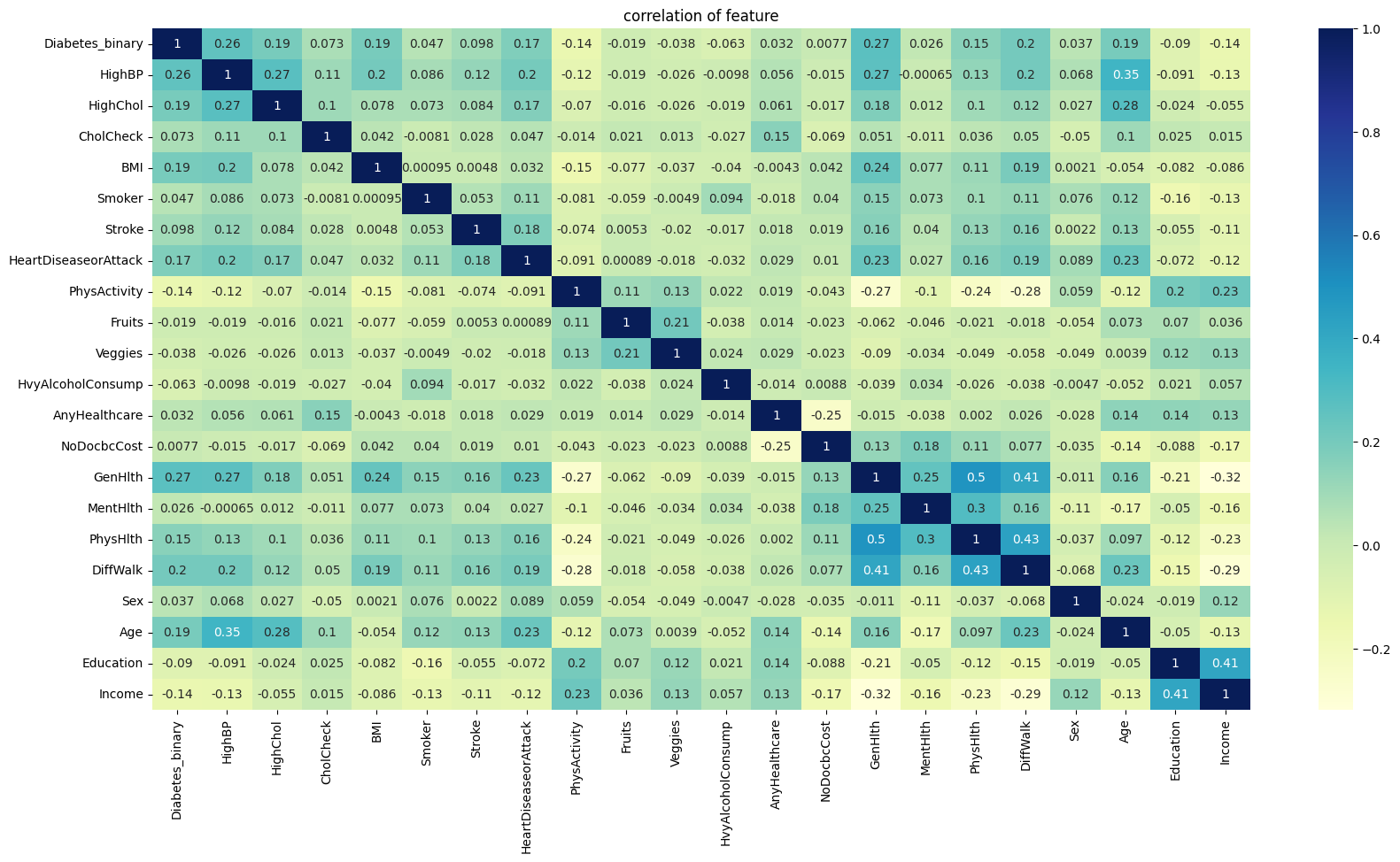
8: Dưới 100.000 đô la

9: Dưới 150.000 đô la

10: Dưới 200.000 đô la

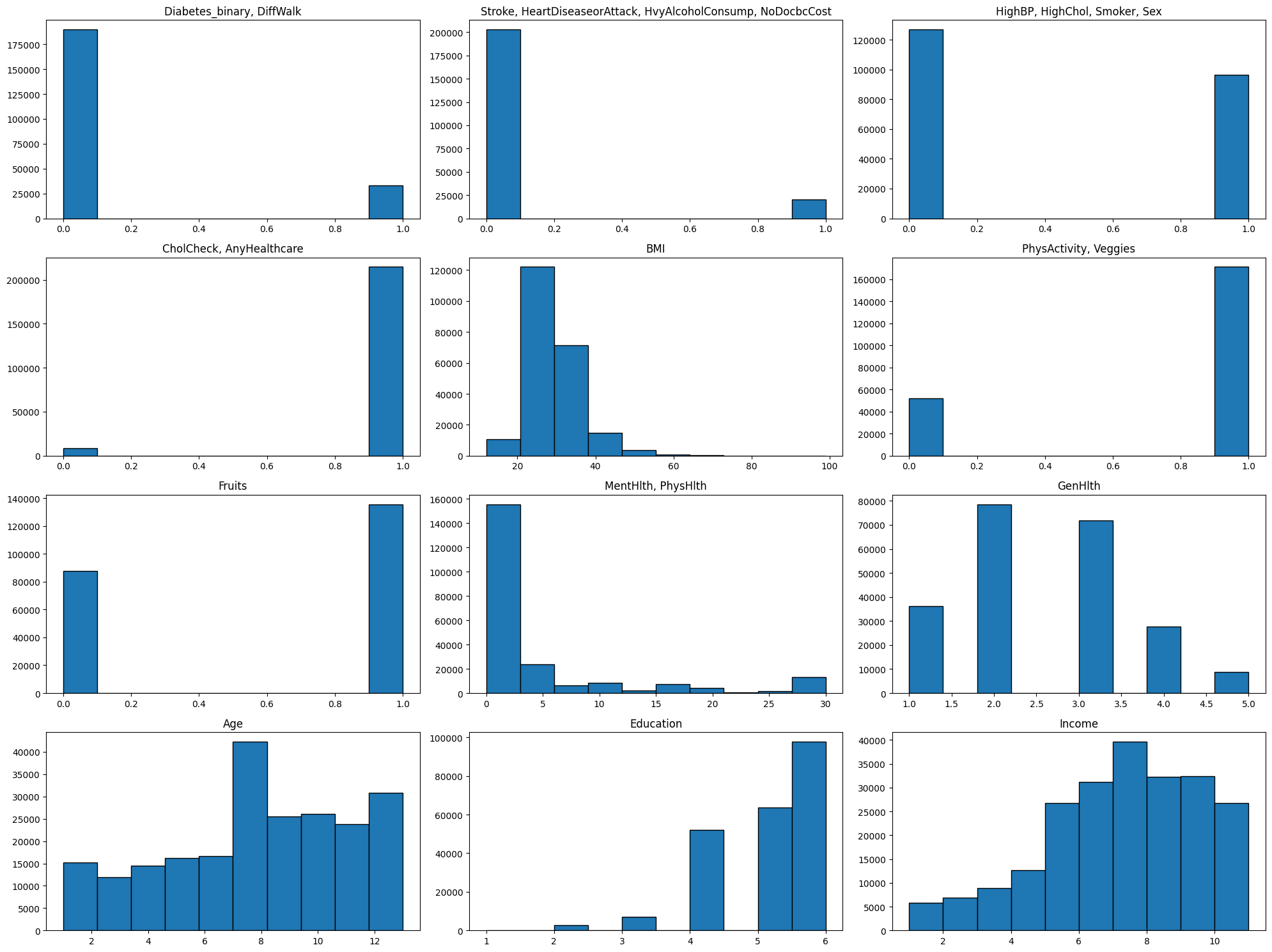
11: 200.000 đô la trở lên

#### Thống kê

****

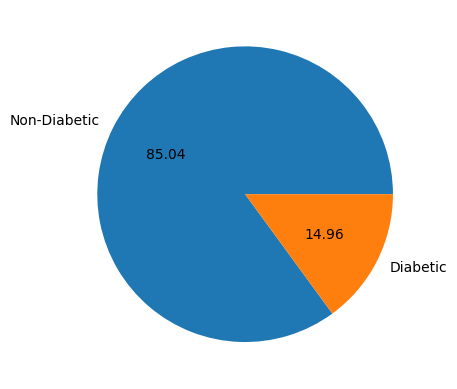
Biểu đồ I.1: Biểu đồ tương quan giữa các thuộc tính

* Biểu đồ trên thể hiện sự tương quan giữa các cột.
* Các cặp: (cột GenHlth, cột PhysHlth), (cột PhysHlth, cột DiffWalk), (cột GenHlth, cột DiffWalk) có mức tương quan dương cao, thể hiện mối quan hệ tích cực. Trong trường hợp của biểu đồ, nếu GenHlth tăng lên, có khả năng PhysHlth và DiffWalk cũng sẽ tăng lên. Điều này có thể diễn giải rằng sức khỏe tổng quát (GenHlth) có mối liên quan tích cực đến sức khỏe thể chất cụ thể (PhysHlth) và khả năng đi lại khó khăn (DiffWalk).
* Các cặp: (cột GenHlth, cột Income), (cột DiffWalk, cột Income) có mức tương quan âm cao, thể hiện mối quan hệ tiêu cực.



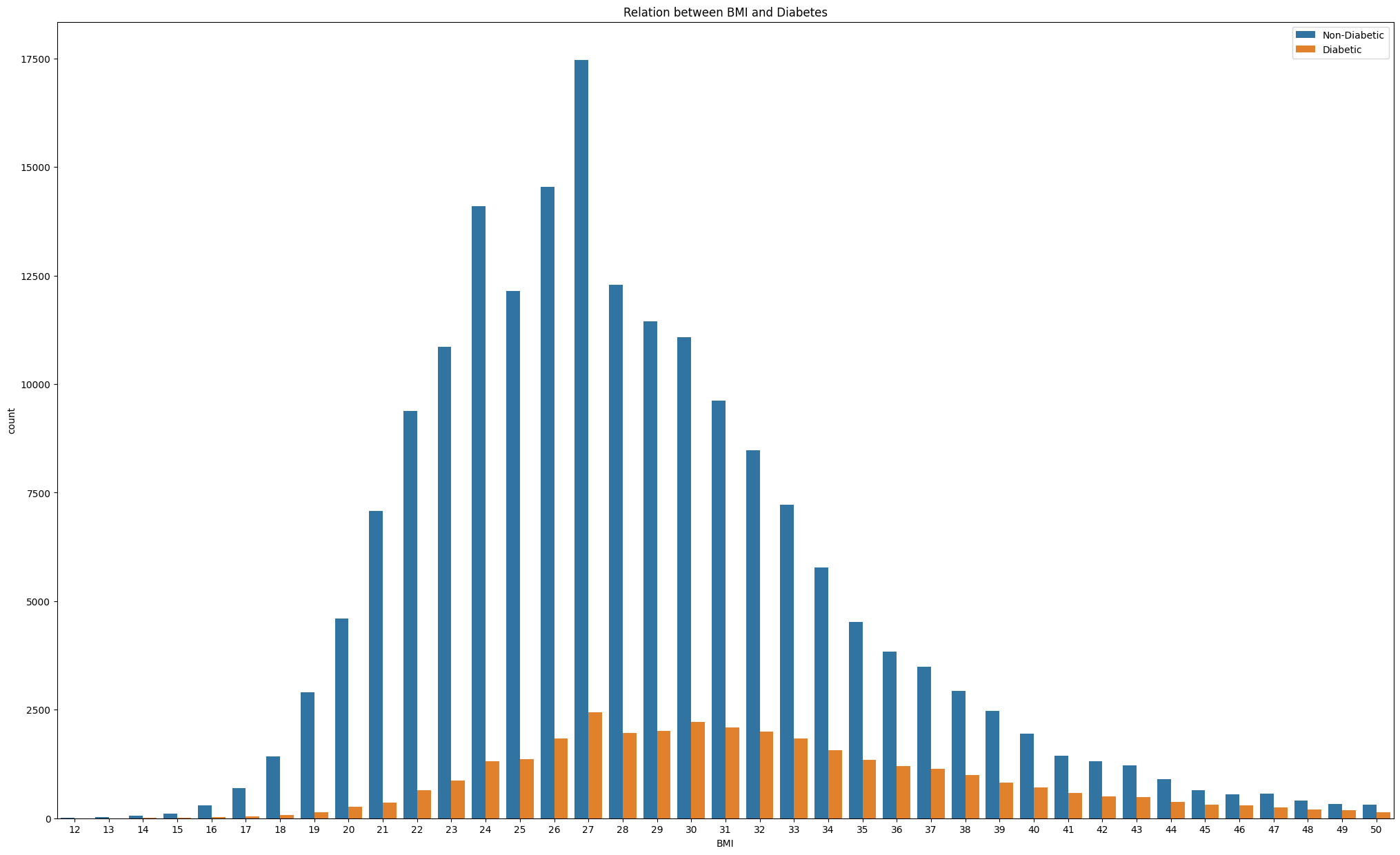
Biểu đồ I.2: Biểu đồ phân phối giá trị trong mỗi thuộc tính

* Mỗi histogram sẽ hiển thị phân phối của giá trị trong mỗi cột của tập dữ liệu. Biểu đồ histogram chỉ ra sự phân phối của dữ liệu, xem xét các giá trị được phân bố trong khoảng giá trị khác nhau.



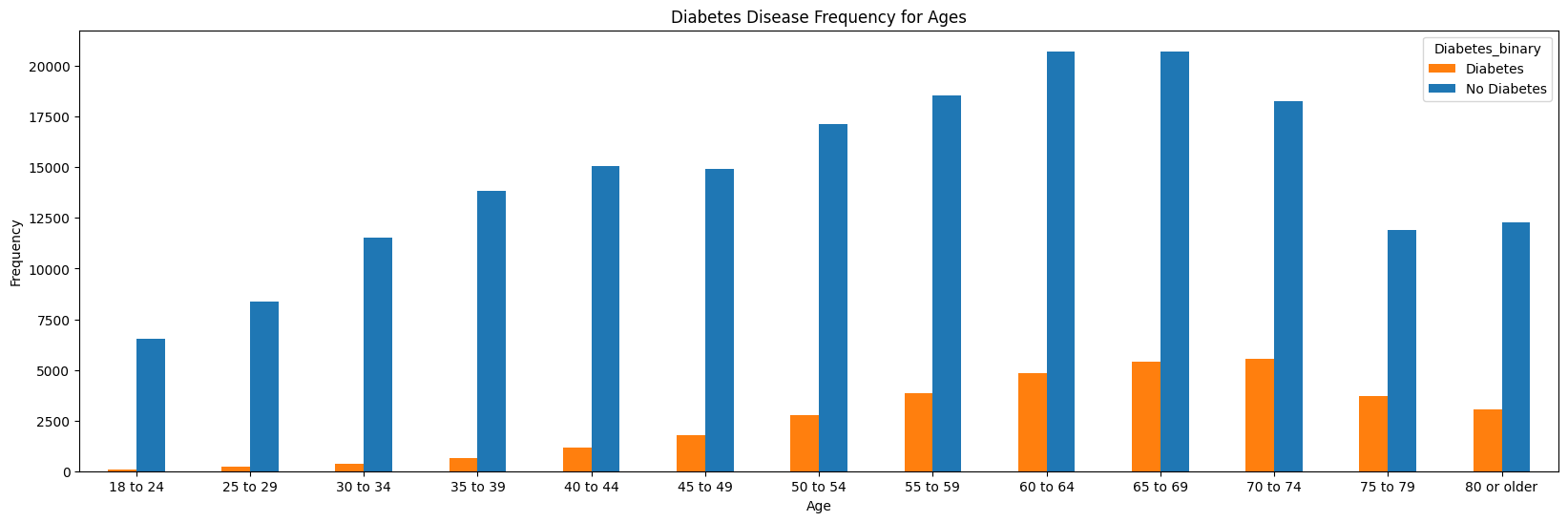
Biểu đồ I.3: Biểu đồ thể hiện tỉ lệ mắc bệnh tiểu đường

* Chúng ta có thể thấy người khỏe mạnh nhiều hơn người mắc bệnh tiểu đường. Và chúng ta cũng có thể thấy rằng 14,96% người mắc bệnh tiểu đường.



Biểu đồ I.4: Biểu đồ thể hiện quan hệ giữa BMI và bệnh tiểu đường

* Như chúng ta có thể thấy những người có chỉ số BMI từ 24-35 có nhiều khả năng mắc bệnh tiểu đường hơn. Và đỉnh điểm tập trung cao nhất của người bệnh tiểu đường là ở chỉ số BMI 27.



Biểu đồ I.5: Biểu đồ thể hiện tần suất bệnh tiểu đường theo độ tuổi

* Chúng tôi biết rằng khi tuổi tác tăng lên, nguy cơ mắc bệnh tiểu đường cũng thường tăng lên. Nhìn từ trên có thể nói, độ tuổi trung bình của người mắc bệnh tiểu đường lớn hơn so với người không mắc bệnh tiểu đường.
* Áp dụng Chi-square là một phép toán thống kê quan hệ giữa hai biến phân loại, sử dụng để kiểm tra xem các quan sát thuộc các nhóm khác nhau có phân phối đồng đều hay không. Cho thấy rằng "Fruits", "Veggies", "CholCheck", "AnyHealthcare", "NoDocbcCost" không có ý nghĩa thống kê và có thể bị xóa khỏi tập dữ liệu. Điều này sẽ giúp giảm tính chiều của dữ liệu và cải thiện hiệu suất của các mô hình học máy.

#### Đánh giá chất lượng

* Bộ dữ liệu có 236.378 hàng và độ đa dạng với 22 thuộc tính
* Bộ dữ liệu có 8333 (3,5%) hàng trùng lặp
* CholCheck bị mất cân bằng cao (77,3%)
* Đột quỵ bị mất cân bằng cao (76,3%)
* HeartDiseaseorAttack bị mất cân bằng cao (57,5%)
* HvyAlcoholConsump bị mất cân bằng cao (66,4%)
* AnyHealthcare bị mất cân bằng cao (77,0%)
* NoDocbcCost bị mất cân bằng cao (65,8%)
* MentHlth có 147300 (62,3%) giá trị 0
* PhysHlth có 156784 (66,3%) giá trị 0

### Tập dữ liệu bệnh tiểu đường của phụ nữ gốc Ấn Độ Pima

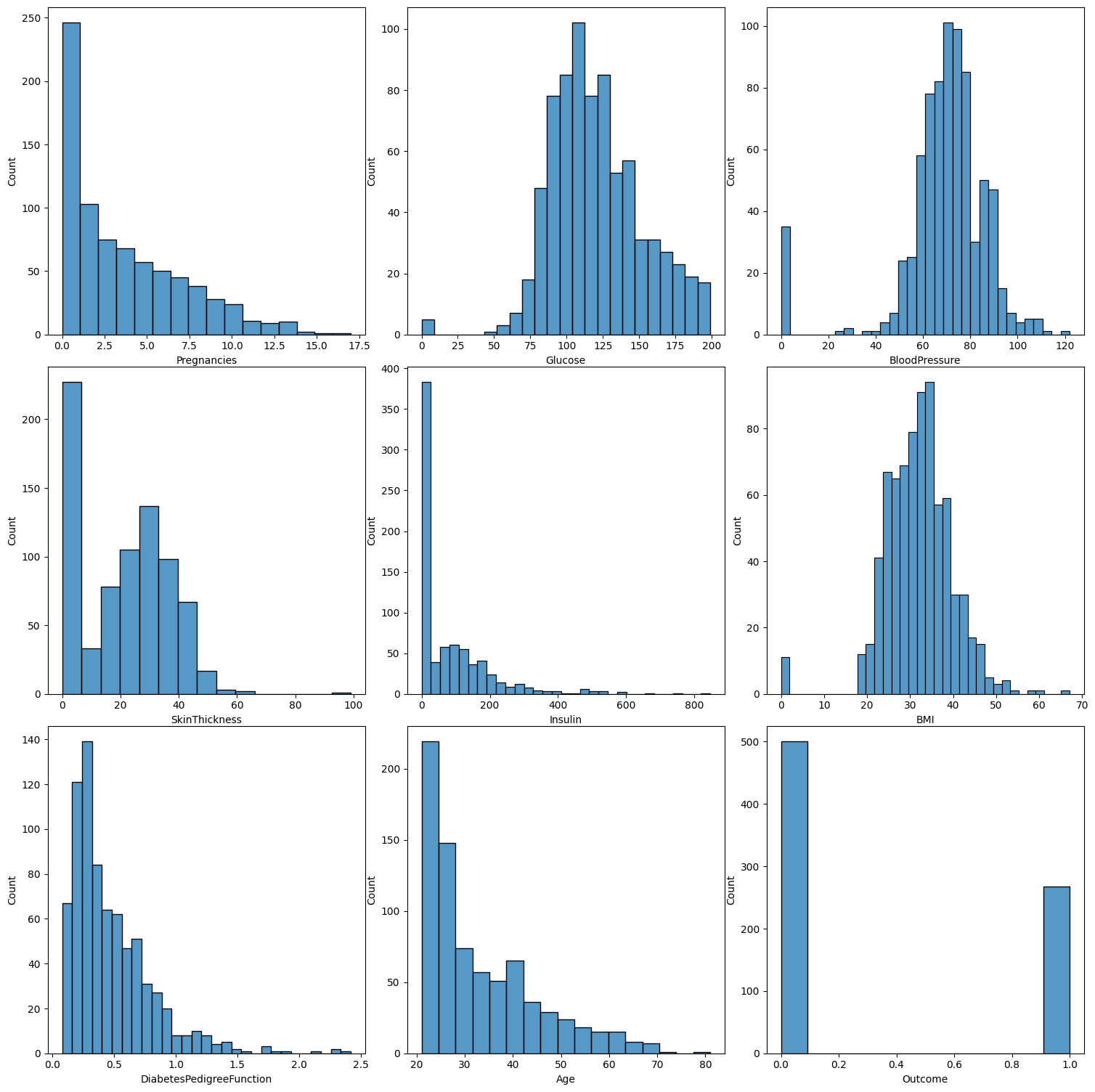
#### Tập mẫu

Dữ liệu sức khỏe của phụ nữ gốc Ấn Độ Pima, ít nhất 21 tuổi, có phản hồi từ 768 cá nhân, bao gồm nhiều thuộc tính liên quan đến sức khỏe.

#### Thuộc tính

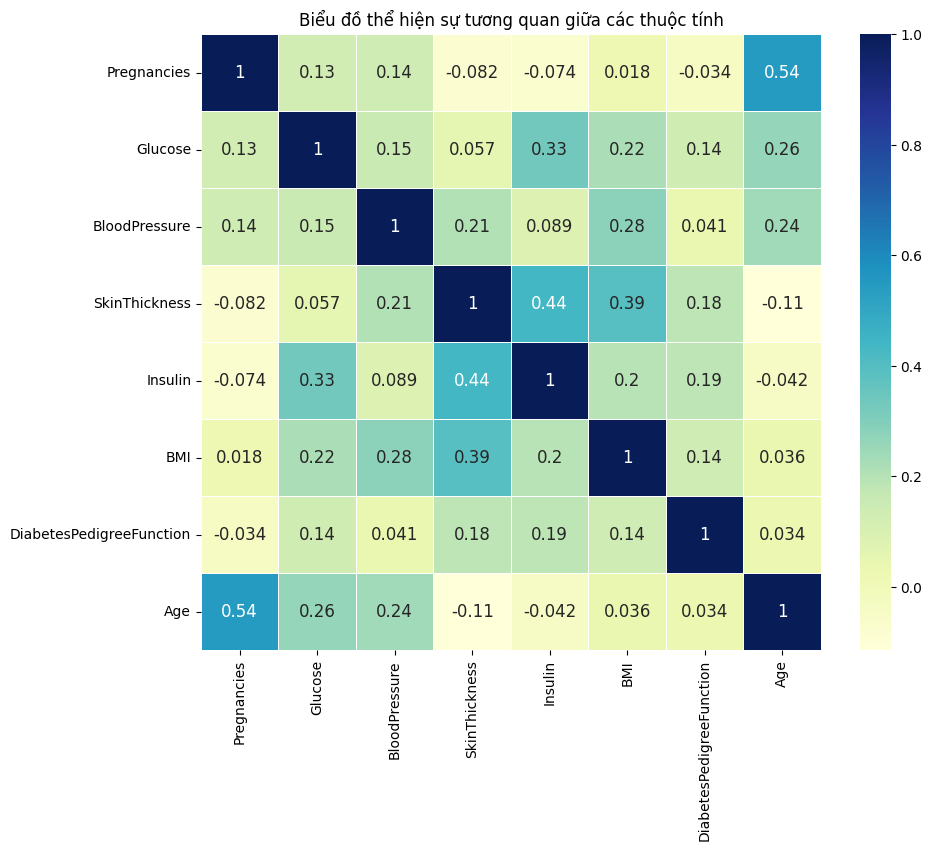
* Pregnancies: Số lần phụ nữ đã mang thai.
* Glucose: Nồng độ đường trong huyết thanh của phụ nữ. Trong máu của con người luôn có một lượng Glucose nhất định để đảm bảo việc cung cấp năng lượng cho các hoạt động thường ngày: 90 - 130 mg/dl (tức 5 - 7,2 mmol/l) ở thời điểm trước bữa ăn, dưới 180 mg/dl (tức 10 mmol/l) ở thời điểm sau ăn khoảng 1 - 2 tiếng, 100 - 150 mg/l (tức 6 - 8,3 mmol/l) ở thời điểm trước khi đi ngủ.
* BloodPressure: Đo lường huyết áp. Huyết áp tối đa thường thay đổi tùy theo độ tuổi, thường từ 90 đến 140 mmHg. Huyết áp tối thiểu dao động trong khoảng từ 50 đến 90 mmHg.
* SkinThickness: Độ dày của da tại vùng cơ triceps, thường được đo bằng cách đo độ dày của lớp mỡ dưới da ở vùng cơ triceps (múi triceps) của cánh tay.
* Insulin: Mức độ insulin trong máu. Insulin là một hormon được sản xuất bởi các tế bào Beta của đảo tụy được sinh ra giúp ổn định lượng đường trong máu và vận chuyển glucose đi nuôi dưỡng các tế bào trong cơ thể.
* BMI (Chỉ số khối cơ thể): Đo lường mỡ cơ thể dựa trên chiều cao và cân nặng.
* DiabetesPedigreeFunction: Một hàm tính điểm xác suất mắc bệnh tiểu đường dựa trên lịch sử gia đình.
* Age: Tuổi của phụ nữ.
* Outcome: Biến mục tiêu chỉ ra liệu phụ nữ có mắc bệnh tiểu đường hay không (1 cho người mắc, 0 cho người không mắc)

#### Thống kê

****

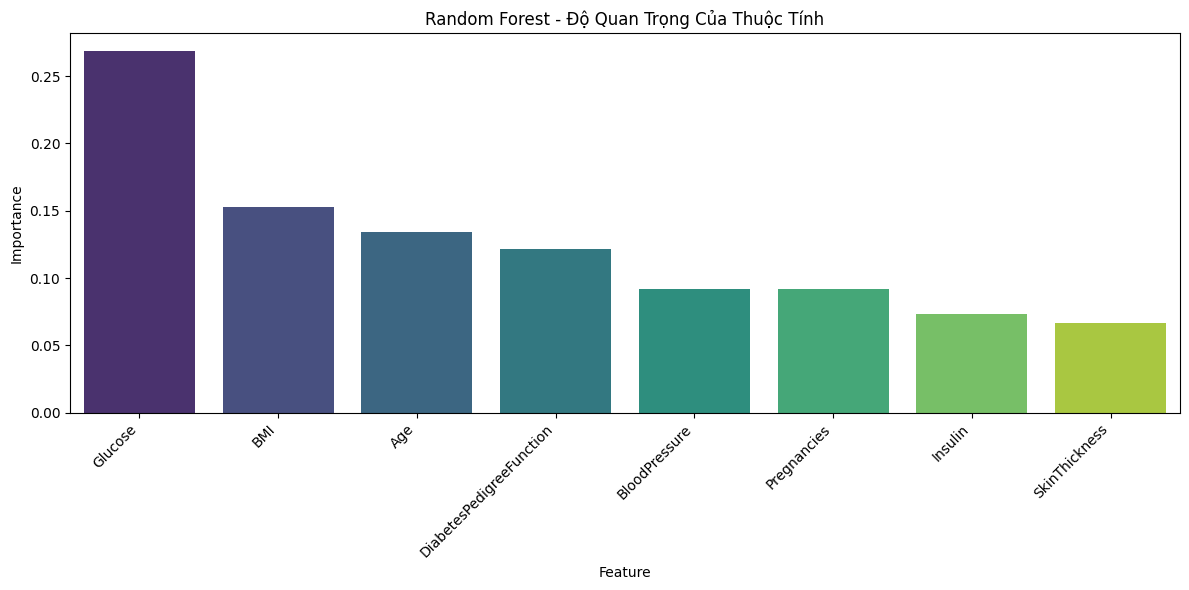
Biểu đồ I.6: Biểu đồ thể hiện sự phân bố của các thuộc tính

Biểu đồ trên là tập hợp các biểu đồ tần suất (histogram) thể hiện sự phân bố của các thuộc tính trong tập dữ liệu.

****

Biểu đồ I.7: Biểu đồ nhiệt thể hiện mối tương quan giữa các thuộc tính

Biểu đồ nhiệt trên thể hiện mối tương quan giữa các thuộc tính. Hầu hết các mối tương quan đều là dương. Các cặp thuộc tính (Age, Pregnancies), (Insulin, SkinThickness), (BMI, SkinThickness), (Insulin, Glucose) có mối tương quan cao với nhau. Có một số ít mối tương quan âm, thể hiện mối tương quan tiêu cực.

****

Biểu đồ I.8: Biểu đồ thể hiện mức độ quan trọng của các thuộc tính

Biểu đồ thể hiện mức độ quan trọng của các thuộc tính trong việc dự đoán bệnh tiểu đường bằng mô hình Random Forest. Mức độ quan trọng được biểu thị bằng giá trị "Importance" trên trục Y, với giá trị càng cao thể hiện thuộc tính đó càng quan trọng. Các thuộc tính được sắp xếp theo thứ tự từ quan trọng nhất đến ít quan trọng nhất trên trục X.

Glucose: Đây là thuộc tính quan trọng nhất. Điều này cho thấy nồng độ Glucose trong máu đóng vai trò quan trọng trong việc dự đoán nguy cơ mắc bệnh tiểu đường.

BMI: Chỉ số khối cơ thể (BMI) cũng là một yếu tố quan trọng. Người có BMI cao có nguy cơ mắc bệnh tiểu đường cao hơn.

Tuổi: Tuổi tác cũng ảnh hưởng đến nguy cơ mắc bệnh tiểu đường, với người cao tuổi có nguy cơ cao hơn. Tuy nhiên, mức độ ảnh hưởng của tuổi tác thấp hơn so với Glucose và BMI.

#### Đánh giá chất lượng

Pregnancies có 111 (14.5%) giá trị 0

BloodPressure có 35 (4.6%) giá trị 0

SkinThickness có 227 (29.6%) giá trị 0

Insulin có 374 (48.7%) giá trị 0

BMI có 11 (1.4%) giá trị 0

Trong tập dữ liệu có một số giá trị 0 xuất hiện trong các thuộc tính như “Pregnancies”, "BloodPressure," "SkinThickness," "Insulin," và "BMI." Đặc biệt là thuộc tính Insulin và SkinThickness, ta thấy đa số giá trị của nó bằng 0, còn các thuộc tính khác thì giá trị 0 không đáng kể. Trong biểu đồ “Random Forest - Độ quan trọng của các thuộc tính”, thuộc tính Insulin và SkinThickness có mức độ quan trọng thấp nhất, kết hợp với có nhiều giá trị không hợp lý thì ta có thể loại bỏ cột này trong quá trình huấn luyện để có thể làm tăng hiệu suất của mô hình.

### Tập dữ liệu bệnh tiểu đường tại IRAQ

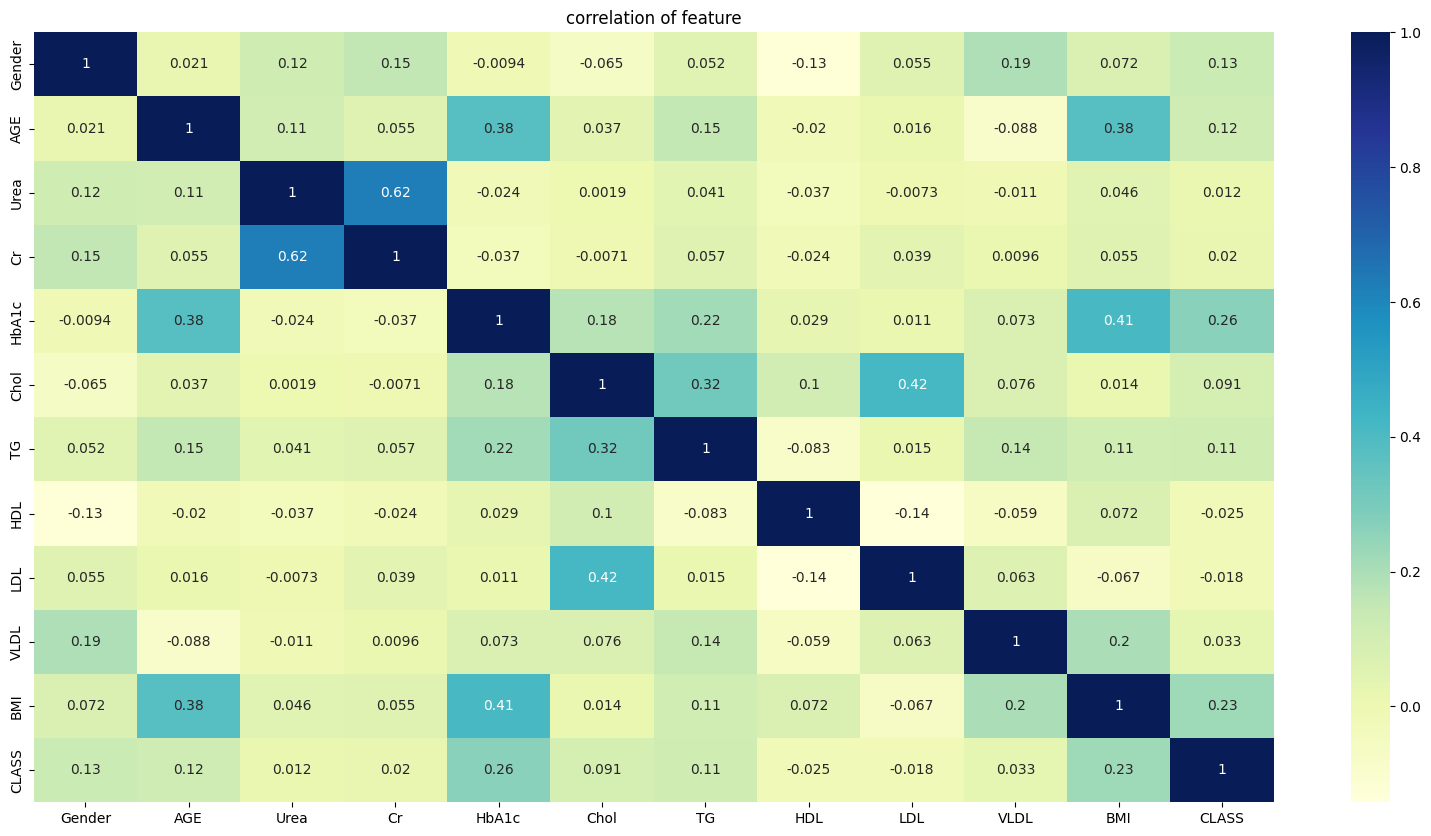
#### Tập mẫu

Dữ liệu bệnh tiểu đường tại Bệnh viện Y Tế Thành phố và Trung tâm Chuyên khoa Nội tiết và Bệnh viện Giảng dạy Tiểu đường-Al-Kindy, từ 1000 cá nhân.

#### Thuộc tính

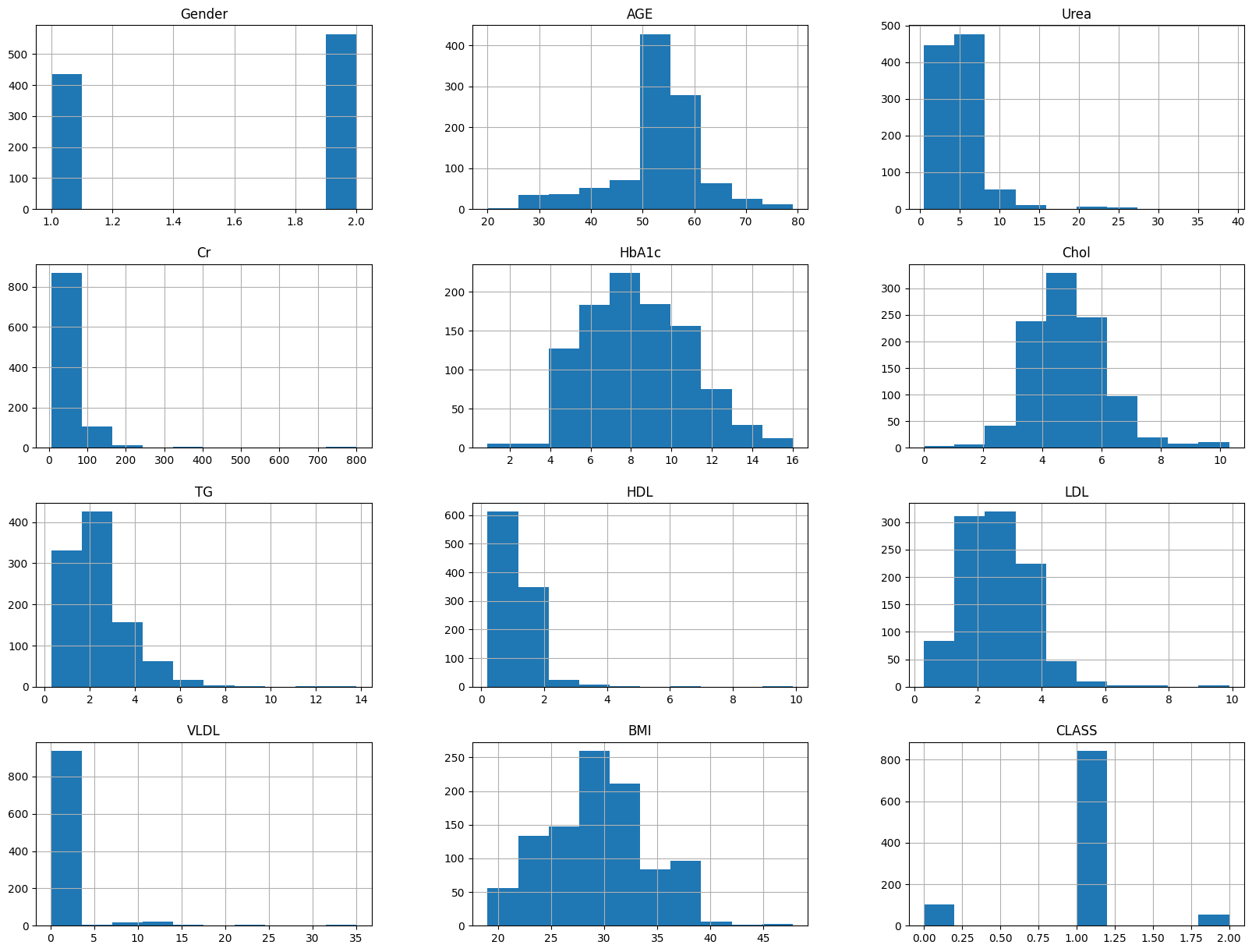
* No. of Patient: Số thứ tự hoặc mã số duy nhất của bệnh nhân trong hệ thống.
* Gender: Giới tính của bệnh nhân, có thể là Nam (M) hoặc Nữ (F).
* Age: Tuổi của bệnh nhân. Có giá trị từ 20 tuổi đến 79 tuổi.
* Urea: Mức độ urea trong máu của bệnh nhân, một chỉ số phản ánh chức năng thận và cung cấp thông tin về cân bằng nước và điện giữa các tế bào và mô mà nó chảy qua. Có giá trị từ 0.5 đến 38.9
* Creatinine ratio (Cr): Tỷ lệ creatinin trong huyết thanh của bệnh nhân. Có giá trị từ 6 đến 800
* HBA1C: Mức độ HbA1c (Glycated Hemoglobin), một chỉ số quan trọng để đánh giá kiểm soát đường huyết của bệnh nhân trong thời gian dài. Có giá trị từ 0.9 đến 16
* Cholesterol (Chol): Mức độ cholesterol trong máu của bệnh nhân, một chỉ số quan trọng cho sức khỏe tim mạch. Có giá trị từ 0 đến 10.3
* Body Mass Index (BMI): Là chỉ số mỡ cơ thể, được tính toán dựa trên cân nặng và chiều cao. Có giá trị từ 19 đến 47.8
* Một loạt các chỉ số lipid được đo sau một thời gian đói, bao gồm tổng cholesterol, cholesterol LDL (Low-Density Lipoprotein), cholesterol VLDL (Very Low-Density Lipoprotein), triglycerides (mỡ trong máu) và cholesterol HDL (High-Density Lipoprotein). Các chỉ số này cung cấp thông tin chi tiết về sức khỏe tim mạch của bệnh nhân.
* Class: Diabetic (Bệnh nhân bị tiểu đường), Non-Diabetic (Bệnh nhân không bị tiểu đường), hoặc Predict-Diabetic (Bệnh nhân có nguy cơ mắc bệnh tiểu đường).

#### Thống kê



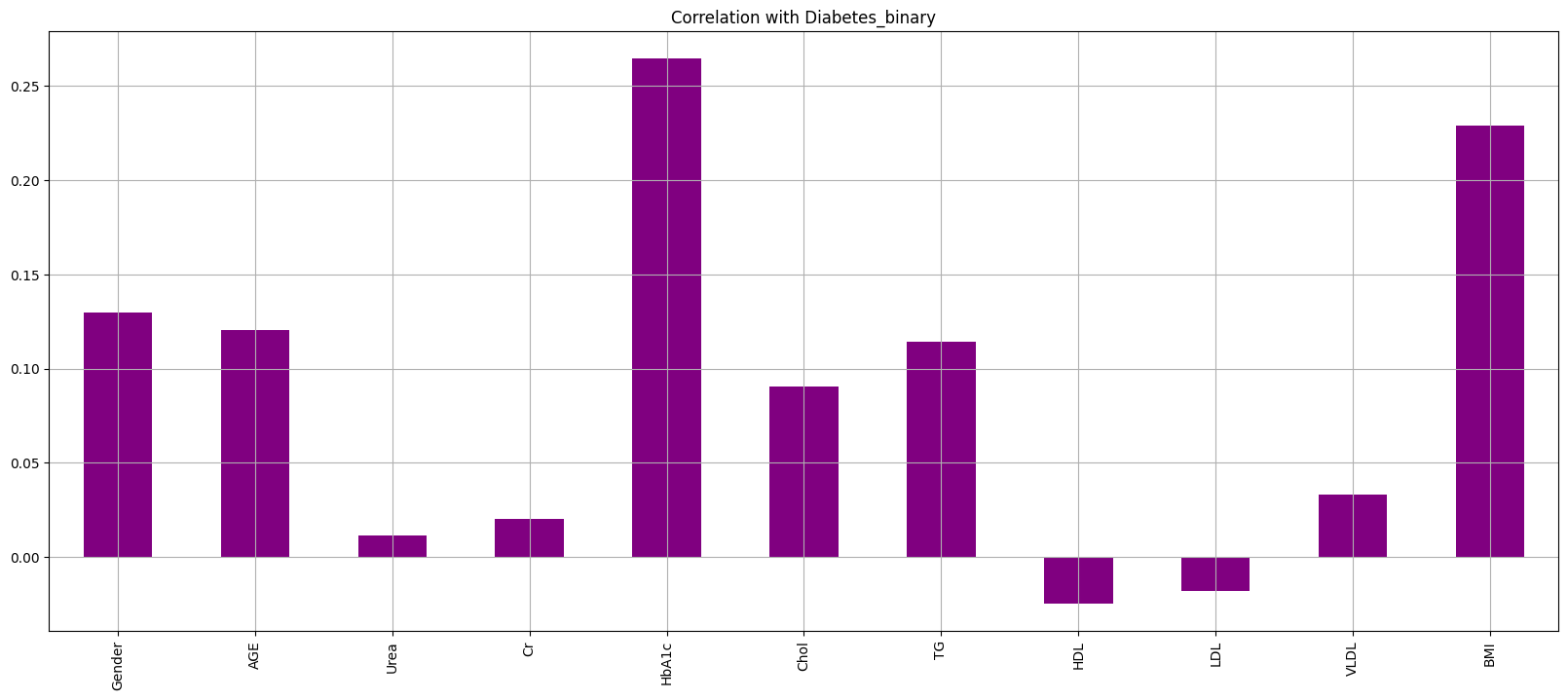
Biểu đồ I.9: Biểu đồ thể hiện mối tương quan giữa các thuộc tính

Hầu hết các mối tương quan đều là dương. Các cặp thuộc tính (HbA1c, CLASS), (BMI, CLASS), (Urea, Cr), (Chol, LDL), (HbA1c, BMI), (HbA1c, AGE), (BMI, AGE) có mối tương quan dương cao với nhau. Mối tương quan càng cao thì chúng sẽ ảnh hưởng, tác động với nhau lớn.



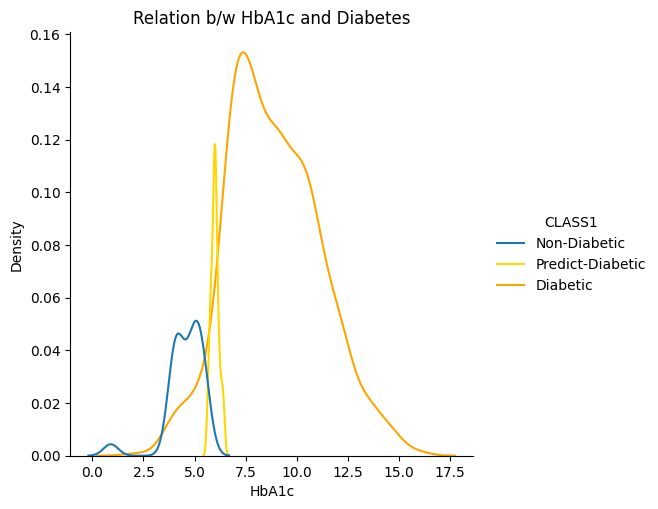
Biểu đồ I.10: Biểu đồ thể hiện phân phối của giá trị trong mỗi thuộc tính

Mỗi histogram sẽ hiển thị phân phối của giá trị trong mỗi cột của tập dữ liệu. Biểu đồ histogram giúp bạn hiểu rõ hơn về phân phối của dữ liệu, giá trị được phân bố trong các khoảng giá trị khác nhau.



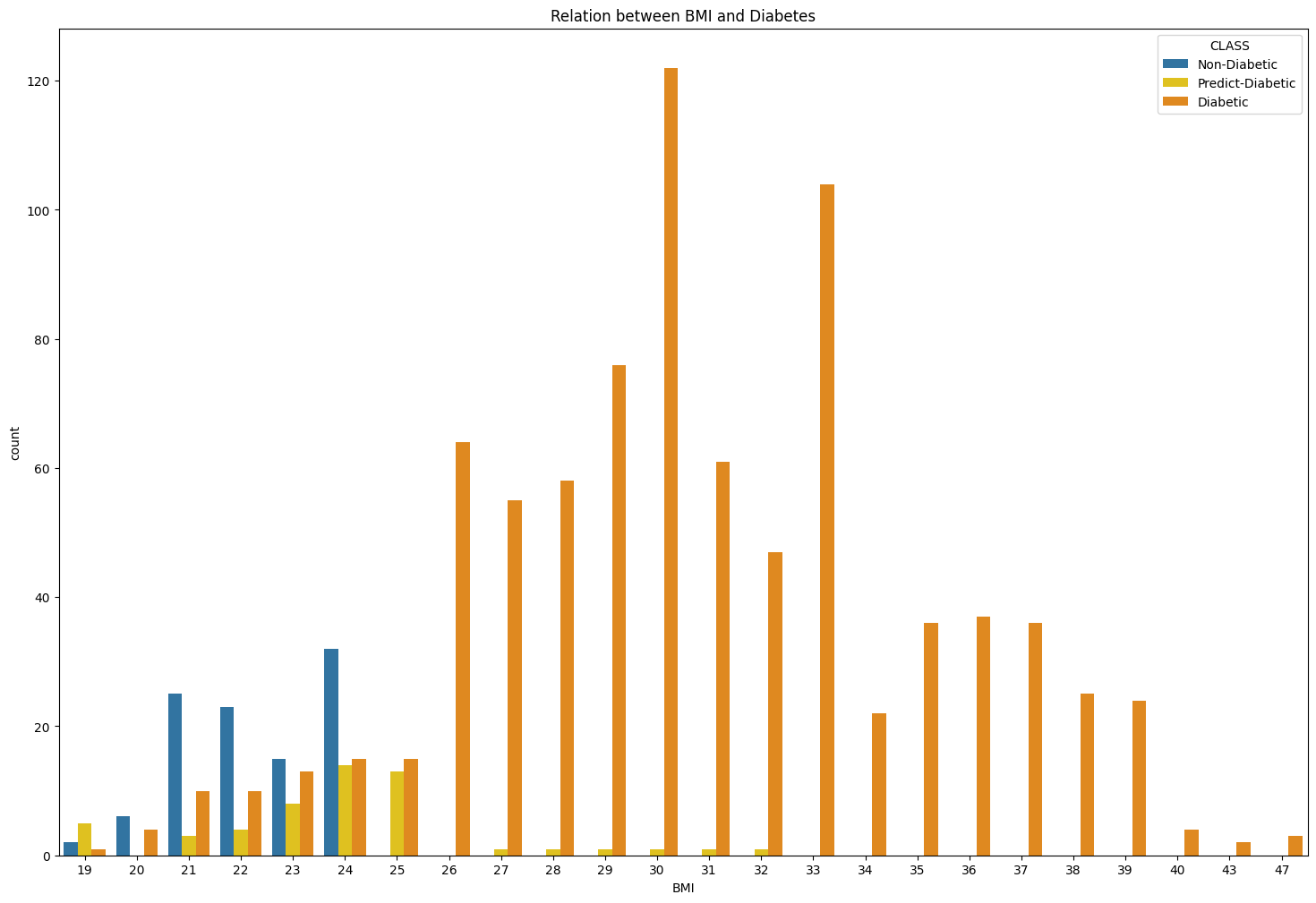
Biểu đồ I.11: Biểu đồ thể hiện mức độ tương quan giữa cột 'CLASS' và các cột khác nhau

* Các kết quả từ biểu đồ cho thấy mức độ tương quan giữa cột 'CLASS' và các cột khác nhau:
* Các cột 'Urea', 'Cr', 'VLDL', 'HDL' và 'LDL' có mức độ tương quan thấp với 'CLASS' . Điều này có thể được hiểu là những cột này ít ảnh hưởng đến khả năng mắc bệnh tiểu đường.
* Các cột 'HbA1C', 'BMI', 'Gender', 'AGE', 'Chol', 'TG' có mức độ tương quan đáng kể với 'CLASS'. Điều này có thể cho thấy rằng những yếu tố này có ảnh hưởng đáng kể đến khả năng mắc bệnh tiểu đường.



Biểu đồ I.12: Biểu đồ thể hiện quan hệ giữa HbA1c và Diabetes

Qua biểu đồ trên ta thấy tính quan trọng của chỉ số HbA1C, đối với người không bị tiểu đường, chỉ số này thấp hơn 6.5%, người có thể bị tiểu đường nằm trong khoảng 5.5-6.5%, người bị tiểu đường có chỉ số này nhỏ hơn, bằng hoặc lớn hơn 6.5%. Thực tế đúng như vậy, HbA1C ở mức 6.5% được khuyến nghị là điểm giới hạn của bệnh tiểu đường và giá trị nhỏ hơn 6,5% không loại trừ bệnh tiểu đường.



Biểu đồ I.13: Biểu đồ thể hiện liên hệ giữa BMI và bệnh tiểu đường

Qua biểu đồ về tính liên hệ giữa BMI và bệnh tiểu đường, ta thấy rằng khi chỉ số này lớn hơn 25 thì có nguy cơ mắc bệnh tiểu đường cao và cao nhất ở 30. Điều này cho thấy càng béo thì càng dễ bị tiểu đường.

#### Đánh giá chất lượng

CLASS tương quan cao với HbA1c.

Cr tương quan cao với Urea.

HbA1c tương quan cao với CLASS.

TG tương quan cao với VLDL.

Urea tương quan cao với Cr.

VLDL tương quan cao với TG.

CLASS bị mất cân bằng (65.0%)

## So sánh

* Điểm giống nhau:
* Liên quan đến y tế: Tất cả ba tập dữ liệu đều có liên quan đến y tế và sức khỏe của con người. Có chứa thông tin về các chỉ số sinh học quan trọng trong việc xác định bệnh tiểu đường như glucose, huyết áp, BMI, cholesterol, và các biến liên quan khác.
* Chứa thông tin về đặc điểm nhân khẩu học: Cả ba tập dữ liệu đều chứa thông tin về tuổi, giới tính, và một số thông tin về đặc điểm nhân khẩu học khác như giáo dục và thu nhập.
* Cả ba tập dữ liệu đều có thông tin về nhãn bệnh tiểu đường, lượng dữ liệu đủ lớn, giúp cho việc huấn luyện và đánh giá mô hình dự đoán.
* Điểm khác nhau:
* Phạm vi và mục tiêu của dữ liệu: Mỗi tập dữ liệu có mục tiêu và phạm vi riêng.
* Loại biến: Các tập dữ liệu có thể chứa các loại biến khác nhau.

Ví dụ: tập dữ liệu thứ ba chứa thông tin về các chỉ số sinh học khác như urea, creatinine trong khi tập dữ liệu thứ 2 không có.

* Số lượng mẫu và chiều sâu của dữ liệu: Các tập dữ liệu có số lượng mẫu và chiều sâu khác nhau.
* Cả 3 bộ dữ liệu đều có những ưu điểm và hạn chế riêng:
* Ở bộ dữ liệu thứ nhất: ưu điểm là kích thước lớn với 236378 mẫu, 22 thuộc tính khác nhau, do đó nó có sự đa dạng và tiềm năng sử dụng rộng rãi. Nhưng dù nhiều thuộc tính thì tập dữ liệu lại không có thuộc tính về chỉ số glucose, một thuộc tính khá quan trọng trong xác định bệnh tiểu đường, trong lĩnh vực y tế.
* Ở bộ dữ liệu thứ hai: ưu điểm là nó có khá đầy đủ các thuộc tính trong việc xác định bệnh tiểu đường như glucose, BMI, điểm xác suất mắc bệnh tiểu đường dựa trên lịch sử gia đình, huyết áp, độ tuổi,... Nhưng bộ dữ liệu này chỉ tập trung về một giới tính là nữ.
* Ở bộ dữ liệu thứ ba: ưu điểm là dữ liệu gồm các thông tin về y tế, phân tích trong phòng thí nghiệm, các chỉ số như Cr, BMI, Urea, Chol, LDL, VLDL, TG, HDL và HBA1C, lượng dữ liệu đủ lớn,...Nhưng bộ dữ liệu này về kích thước mẫu theo nhãn ‘CLASS’ chưa hợp lý: (Không bị bệnh tiểu đường: 103, mắc bệnh tiểu đường 844, có khả năng bị bệnh: 53)
* Kết luận 3 tập dữ liệu trên rất có tiềm năng trong việc xây dựng mô hình dự đoán bệnh tiểu đường.Cách thu thập và nguồn gốc của dữ liệu: Các tập dữ liệu được thu thập từ các nguồn **khác nhau** và có phương pháp thu thập dữ liệu khác nhau.

Tóm lại, điểm giống nhau chủ yếu nằm ở sự liên quan đến sức khỏe và thông tin y tế, trong khi điểm khác nhau chủ yếu nằm ở mục tiêu nghiên cứu, phạm vi và nguồn gốc dữ liệu. Điều này làm nổi bật sự đa dạng và tiềm năng sử dụng rộng rãi của ba tập dữ liệu.

## Kết luận

Cả 3 bộ dữ liệu đều có những ưu điểm và hạn chế riêng:

- Ở bộ dữ liệu thứ nhất: ưu điểm là kích thước lớn với 236378 mẫu, 22 thuộc tính khác nhau, do đó nó có sự đa dạng và tiềm năng sử dụng rộng rãi. Nhưng dù nhiều thuộc tính thì tập dữ liệu lại không có thuộc tính về chỉ số glucose, một thuộc tính khá quan trọng trong xác định bệnh tiểu đường, trong lĩnh vực y tế.

- Ở bộ dữ liệu thứ hai: ưu điểm là nó có khá đầy đủ các thuộc tính trong việc xác định bệnh tiểu đường như glucose, BMI, điểm xác suất mắc bệnh tiểu đường dựa trên lịch sử gia đình, huyết áp, độ tuổi,... Nhưng bộ dữ liệu này chỉ tập trung về một giới tính là nữ.

- Ở bộ dữ liệu thứ ba: ưu điểm là dữ liệu gồm các thông tin về y tế, phân tích trong phòng thí nghiệm, các chỉ số như Cr, BMI, Urea, Chol, LDL, VLDL, TG, HDL và HBA1C, lượng dữ liệu đủ lớn,...Nhưng bộ dữ liệu này về kích thước mẫu theo nhãn ‘CLASS’ chưa hợp lý: (Không bị bệnh tiểu đường: 103, mắc bệnh tiểu đường 844, có khả năng bị bệnh: 53)

Kết luận: 3 tập dữ liệu trên rất có tiềm năng trong việc xây dựng mô hình dự đoán bệnh tiểu đường.

# PHẦN CUỐI KỲ

Sau quá trình phân tích, so sánh và đánh giá, nhóm chọn tập dữ liệu bệnh tiểu đường tại US để tiến hành phân tích sâu hơn trong phần báo cáo cuối kỳ.

Dữ liệu được thu thập bởi Behavioral Risk Factor Surveillance System (BRFSS), một hệ thống khảo sát điện thoại dựa trên tiểu bang được tiến hành hàng năm bởi Trung tâm Kiểm soát và Phòng ngừa Dịch bệnh (CDC) của Hoa Kỳ. Do đó, dữ liệu này được coi là đáng tin cậy, giúp cho kết quả nghiên cứu được công nhận và chấp nhận rộng rãi hơn.

Kích thước lớn với 236378 mẫu, 22 thuộc tính khác nhau, do đó nó có sự đa dạng và tiềm năng sử dụng rộng rãi. Các thuộc tính trong bộ dữ liệu này cung cấp một cái nhìn toàn diện về các yếu tố có thể ảnh hưởng đến tiểu đường và sức khỏe tổng thể. Bộ dữ liệu đã được làm sạch, loại bỏ các giá trị bị thiếu và các ngoại lệ, giúp cho việc phân tích và xử lý dữ liệu dễ dàng hơn. Bộ dữ liệu chưa cân bằng lớp.Vì vậy, dữ liệu phản ánh thực tế tỷ lệ mắc bệnh trong dân số, nhưng có thể gây ra các vấn đề khi huấn luyện mô hình do sự thiếu cân bằng. Vì vậy để nắm bắt thực tế tỷ lệ mắc bệnh trong dân số, thì việc sử dụng bộ dữ liệu không cân bằng lớp là lựa chọn phù hợp.

## 1. Thông tin chi tiết về dữ liệu

### Tập dữ liệu bệnh tiểu đường tại US

Name: Diabetes Health Indicators Dataset

File name: diabetes\_binary\_health\_indicators\_BRFSS2021.csv

Author(s): JULLIEN NAZREEN

Downloadable

Links: [Diabetes Health Indicators Dataset (kaggle.com)](https://www.kaggle.com/datasets/julnazz/diabetes-health-indicators-dataset/data?select=diabetes_binary_health_indicators_BRFSS2021.csv)

### Tập mẫu

* Bộ dữ liệu "diabetes\_binary\_health\_indicators\_BRFSS2021.csv là một tập dữ liệu rõ ràng gồm 236.378 câu trả lời khảo sát. Biến mục tiêu Diabetes\_binary có 2 lớp. 0 là không mắc bệnh tiểu đường và 1 là tiền tiểu đường hoặc tiểu đường. Tập dữ liệu này có 21 biến đặc trưng và không cân bằng.

### Thuộc tính

* Diabetes\_binary (Tiểu đường), nhận giá trị 0 hoặc 1:

0: Không mắc bệnh tiểu đường

1: Tiền tiểu đường hoặc đang mắc bệnh tiểu đường

* HighBP (Huyết áp cao), nhận giá trị 0 hoặc 1:

0: Không có huyết áp cao

1: Có huyết áp cao

* HighChol (Cholesterol cao), nhận giá trị 0 hoặc 1:

0: Không có cholesterol cao

1: Có cholesterol cao

* CholCheck (Kiểm tra cholesterol), nhận giá trị 0 hoặc 1:

0: Không kiểm tra cholesterol trong vòng 5 năm

1: Đã kiểm tra cholesterol trong vòng 5 năm

* BMI (Chỉ số khối cơ thể), nhận các giá trị lớn hơn 0, chỉ số này đo lường sự tương quan giữa cân nặng và chiều cao của một người.
* Smoker (Hút thuốc), nhận giá trị 0 hoặc 1:

0: Không hút thuốc

1: Đã hút ít nhất 100 điếu thuốc trong cả đời

* Stroke (Tai biến), nhận giá trị 0 hoặc 1:

0: Không bị tai biến

1: Đã bị tai biến

* HeartDiseaseorAttack (Bệnh tim mạch hoặc đau tim), nhận giá trị 0 hoặc 1:

0: Không mắc bệnh tim mạch hoặc đau tim

1: Đã mắc bệnh tim mạch hoặc đã trải qua đau tim

* PhysActivity (Hoạt động vận động), nhận giá trị 0 hoặc 1:

0: Không tham gia hoạt động vận động trong 30 ngày qua (không tính công việc)

1: Tham gia hoạt động vận động trong 30 ngày qua (không tính công việc)

* Fruits (Tiêu thụ trái cây), nhận giá trị 0 hoặc 1:

0: Không tiêu thụ ít nhất 1 loại trái cây mỗi ngày

1: Tiêu thụ ít nhất 1 loại trái cây mỗi ngày

* Veggies (Tiêu thụ rau củ):

0: Không tiêu thụ ít nhất 1 loại rau củ mỗi ngày

1: Tiêu thụ ít nhất 1 loại rau củ mỗi ngày

* HvyAlcoholConsump (Tiêu thụ rượu nặng):

0: Không tiêu thụ rượu nặng (nam giới uống hơn 14 đơn vị mỗi tuần, nữ giới uống hơn 7 đơn vị mỗi tuần)

1: Tiêu thụ rượu nặng

* AnyHealthcare (Bảo hiểm y tế):

0: Không có bảo hiểm y tế hoặc bất kỳ hình thức bảo hiểm nào khác

1: Có bảo hiểm y tế hoặc hình thức bảo hiểm khác

* NoDocbcCost (Không đến bác sĩ do chi phí):

0: Không có thời điểm trong 12 tháng qua cần đến bác sĩ mà không thể do chi phí

1: Có thời điểm trong 12 tháng qua cần đến bác sĩ mà không thể do chi phí

* GenHlth (Tình trạng sức khỏe tổng quát), đánh giá trạng thái sức khỏe tổng quát trên thang đánh giá từ 1 đến 5:

1: Xuất sắc

2: Rất tốt

3: Tốt

4: Trung bình

5: Kém

* MentHlth (Tình trạng sức khỏe tâm thần), số ngày trong vòng 30 ngày qua mà tâm trạng không tốt, bao gồm căng thẳng, trầm cảm và vấn đề về cảm xúc.
* PhysHlth (Tình trạng sức khỏe vật lý), số ngày trong vòng 30 ngày qua mà sức khỏe vật lý không tốt, bao gồm bệnh tật và chấn thương.
* DiffWalk (Khó khăn khi đi bộ hoặc leo cầu thang) , nhận giá trị 0 hoặc 1:

0: Không gặp khó khăn

1: Gặp khó khăn

* Sex (Giới tính), nhận giá trị 0 hoặc 1:

0: Nữ

1: Nam

* Age (Độ tuổi), sử dụng thang đo từ 1 đến 13, mô tả các khoảng độ tuổi, ví dụ:

1: 18-24 tuổi

2: 25-29 tuổi

3: 30-34 tuổi

4: 35-39 tuổi

5: 40-44 tuổi

6: 45-49 tuổi

7: 50-54 tuổi

8: 55-59 tuổi

9: 60-64 tuổi

10: 65-69 tuổi

11: 70-74 tuổi

12: 75-79 tuổi

13: 80 tuổi trở lên

* Education (Trình độ học vấn), sử dụng thang đo từ 1 đến 6, mô tả trình độ học vấn:

1: Không đi học hoặc chỉ học mẫu giáo

2: Học tiểu học

3: Học trung học cơ sở

4: Học trung học phổ thông

5: Có bằng đại học

6: Có bằng cao học

* Income (Thu nhập), sử dụng thang đo từ 1 đến 11, mô tả mức thu nhập:

1: Dưới 10.000 đô la

2: Dưới 10.000 đô la

3: Dưới 10.000 đô la

4: Dưới 10.000 đô la

5: Dưới 35.000 đô la

6: Dưới 50.000 đô la

7: Dưới 75.000 đô la

8: Dưới 100.000 đô la

9: Dưới 150.000 đô la

10: Dưới 200.000 đô la

11: 200.000 đô la trở lên

## So sánh và kết luận sâu

### Bệnh tiểu đường là gì? Nguyên nhân. Tỉ lệ người trưởng thành mắc bệnh trong tập dữ liệu thể hiện như thế nào?

Bệnh đái tháo đường (hay còn gọi là bệnh tiểu đường) là bệnh rối loạn chuyển hóa không đồng nhất, có đặc điểm tăng glucose huyết do khiếm khuyết về tiết insulin, về tác động của insulin hoặc cả hai. Tăng glucose mạn tính trong thời gian dài gây nên những rối loạn chuyển hóa carbohydrate, protide, lipide, gây tổn thương ở nhiều cơ quan khác nhau, đặc biệt ở tim và mạch máu, thận, mắt, thần kinh[[1]](#footnote-1)[1].

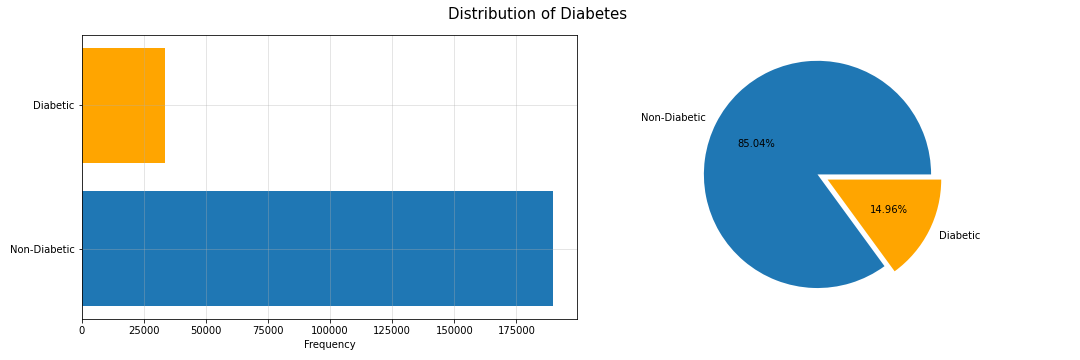
Phân loại đái tháo đường gồm:

Đái tháo đường type 1: Đây là bệnh xảy ra do cơ thể phát sinh phản ứng tự miễn ngừng sản xuất insulin. Do đó, người mắc bệnh lý này chỉ có một cách điều trị duy nhất là dùng insulin ngoại sinh trong cả phần đời còn lại của họ.

Đái tháo đường type 2: Ở bệnh lý này cơ thể sinh ra các kháng thể kháng insulin hay nói dễ hiểu hơn là mặc dù cơ thể vẫn tạo ra insulin nhưng kháng thể kháng insulin sẽ làm giảm đi lượng insulin được tiết vào máu, dẫn đến thiếu hụt insulin.

Đái tháo đường thai kỳ (là đái tháo đường được chẩn đoán trong 3 tháng giữa hoặc 3 tháng cuối của thai kỳ và không có bằng chứng về đái tháo đường type 1, type 2 trước đó): Đây là tình trạng cơ thể ít nhạy cảm trước insulin và có thể khỏi bệnh ngay sau sinh.

Nguyên nhân chủ yếu là do lối sống kém lành mạnh, nạp quá nhiều đường và chất béo, không bổ sung đầy đủ chất xơ, hạn chế vận động. Bệnh tiểu đường có thể dẫn đến nhiều biến chứng nguy hiểm, do đó cần phát hiện sớm để có thể can thiệp điều trị hiệu quả ngay từ giai đoạn đầu.

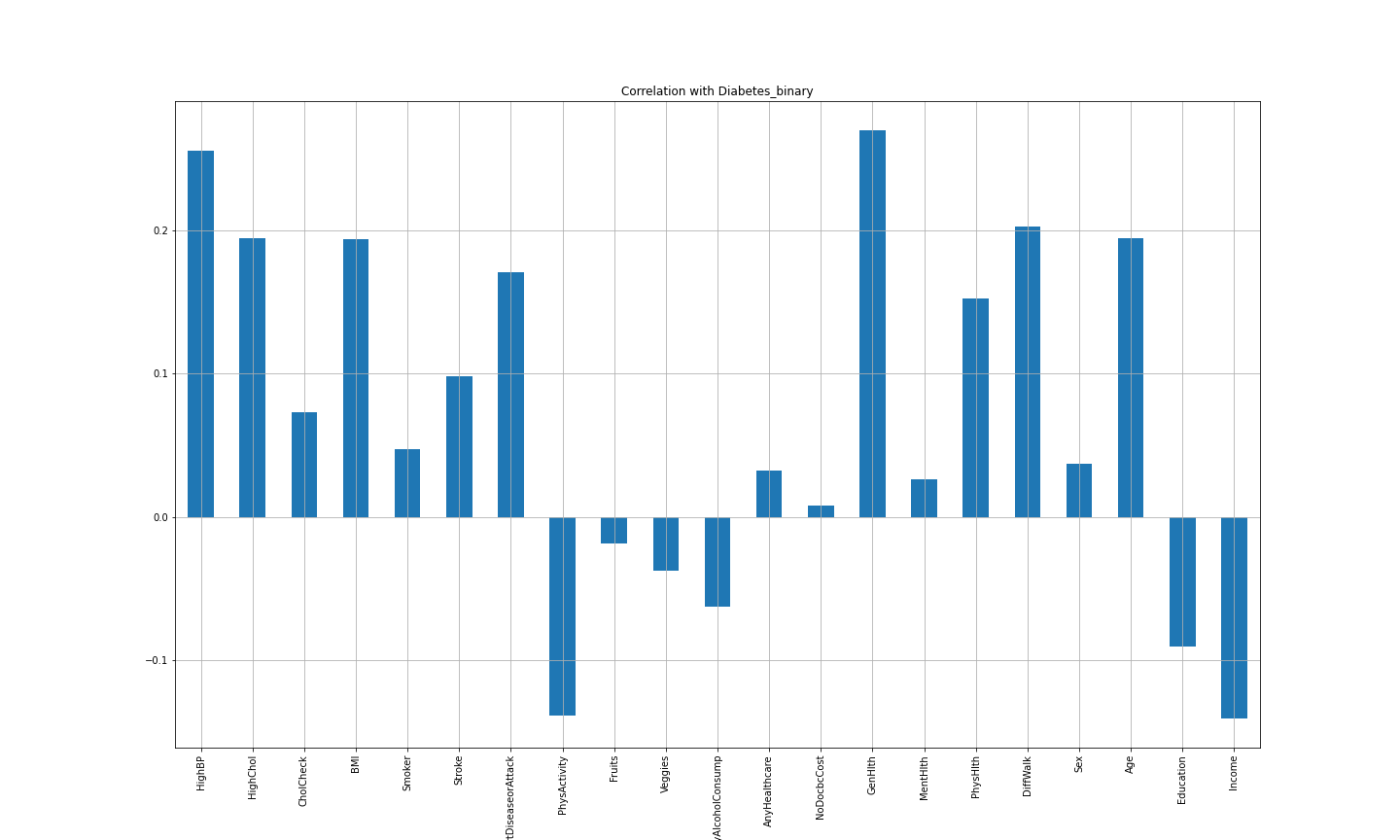


Biểu đồ II.1: Biều đồ thể hiện phân phối và tỉ lệ về bệnh tiểu đường

Trong tập dữ liệu có 14,96% người trưởng thành mắc bệnh tiểu đường. “Theo thống kê của Liên đoàn Đái tháo đường thế giới (IDF), năm 2021 trên thế giới ước tính có khoảng 537 triệu người trưởng thành đang sống chung với bệnh tiểu đường, tương ứng với tỷ lệ cứ 10 người lớn độ tuổi 20 - 79 tuổi có 1 người mắc tiểu đường”[[2]](#footnote-2)[1]. Có thể thấy tỉ lệ người trưởng thành mắc bệnh của tập dữ liệu cao hơn không quá nhiều so với thống kê ở trên.

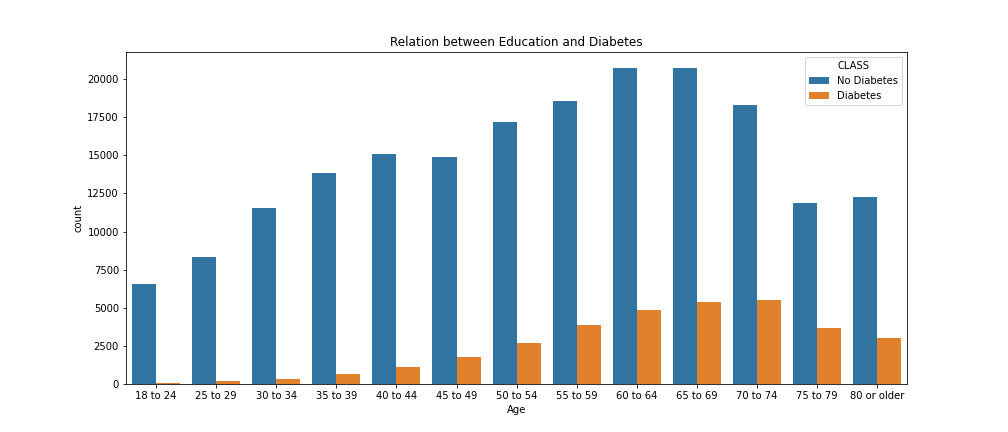
### Các thuộc tính có mối tương quan với bệnh tiểu đường

Vậy những yếu tố hay thuộc tính nào có mối tương quan với bệnh tiểu đường? Biểu đồ dưới đây sẽ thể hiện điều đó:



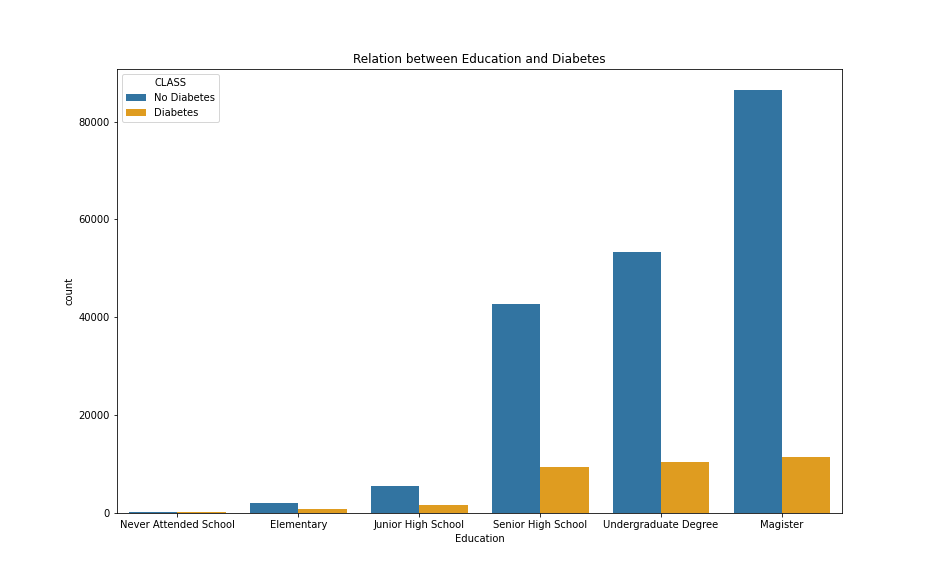
Biểu đồ II.2: Biểu đồ thể hiện mức độ tương quan của các thuộc tính với bệnh tiểu đường

Biểu đồ trên cho thấy các thuộc tính HighBP, HighChol, PhysActivity, GenHlth, DiffWalk, Age, BMI, Education, Income có mối tương quan khá cao với biến Diabetes\_binary. Ta sẽ phân tích một vài thuộc tính trong số đó.



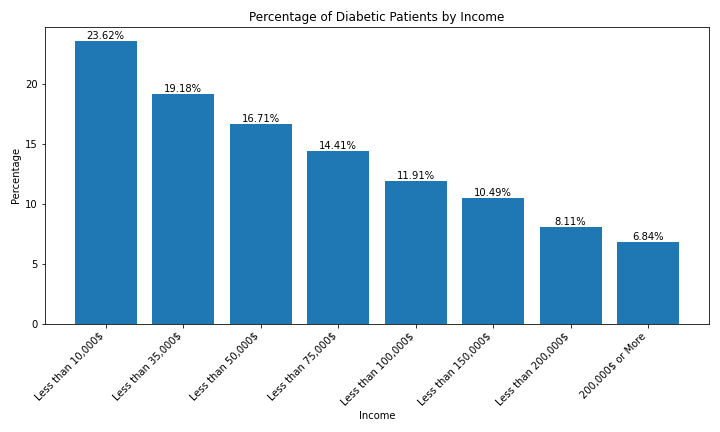
Biểu đồ II.3: Biểu đồ thể hiện tần suất bệnh tiểu đường theo độ tuổi

Từ biểu đồ có thể nhận thấy, từ độ tuổi 18 đã ghi nhận những trường hợp mắc bệnh tiểu đường. “Theo Báo điện tử Chính phủ, bệnh tiểu đường có thể khởi phát từ người vài tháng tuổi, độ tuổi hay gặp nhất là từ 10-14 tuổi”[[3]](#footnote-3)[2]. Ở độ tuổi càng cao thì tỉ lệ mắc bệnh tiểu đường càng cao, độ tuổi từ 60 đến 74 là độ tuổi có tỉ lệ mắc bệnh tiểu đường cao nhất.

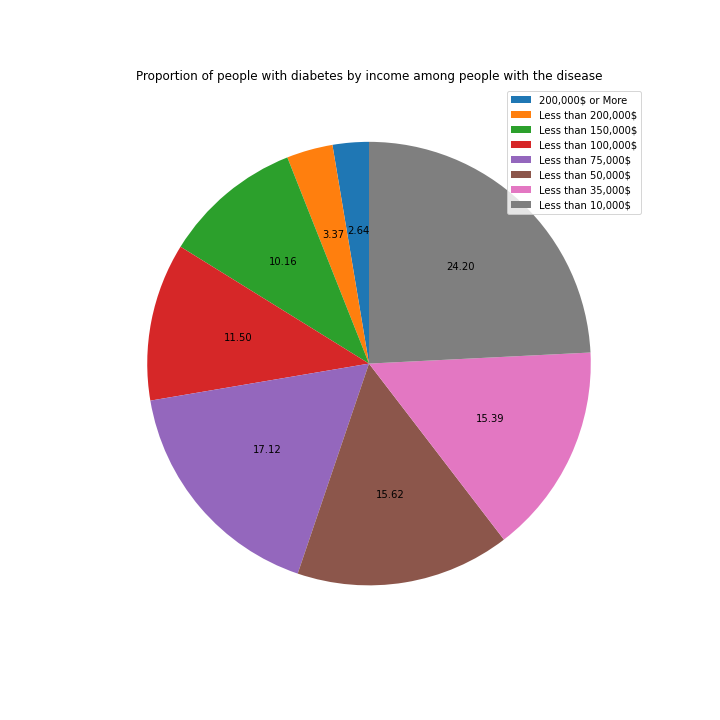


Biểu đồ II.4: Biểu đồ thể hiện mối quan hệ giữa Education (trình độ học vấn) và bệnh tiểu đường

Vậy tại sao trình độ học vấn (Education) lại có liên quan đến bệnh tiểu đường? Ta dễ dàng nhận thấy mặc dù trình độ học vấn càng cao thì số lượng mắc bệnh tiểu đường càng cao nhưng tỉ lệ lại thấp đi. Chúng ta có thể coi đây là một dấu hiệu chỉ ra rằng những người có trình độ học vấn cao hơn thì nắm bắt thông tin nhiều hơn và cảnh giác hơn với căn bệnh này.



Biểu đồ II.5: Biểu đồ thể hiện tỉ lệ mắc bệnh tiểu đường trong từng nhóm thu nhập (Income)



Biểu đồ II.6: Biểu đồ thể hiện tỉ lệ mắc bênh tiểu đường theo thu nhập trong nhóm đã mắc bệnh tiểu đường

Cả hai biểu đồ trên đều chỉ ra rằng những người có thu nhập càng cao thì tỉ lệ mắc bệnh tiểu đường càng thấp. Đây có thể là do những người có thu nhập càng cao thì lại càng quan tâm đến sức khỏe hơn, có mức sống tốt hơn những người có thu nhập thấp cho nên nguy cơ mắc bệnh sẽ giảm đi.

### Hoạt động về thể chất, thói quen ăn uống, lối sống không lành mạnh có ảnh hưởng thế nào đến nguy cơ mắc bệnh tiểu đường

Để đánh giá các hoạt động về thể chất, thói quen ăn uống, lối sống không lành mạnh có ảnh hưởng như thế nào đến nguy cơ mắc bệnh tiểu đường đối với tập dữ liệu, ta sẽ thực hiện như sau:

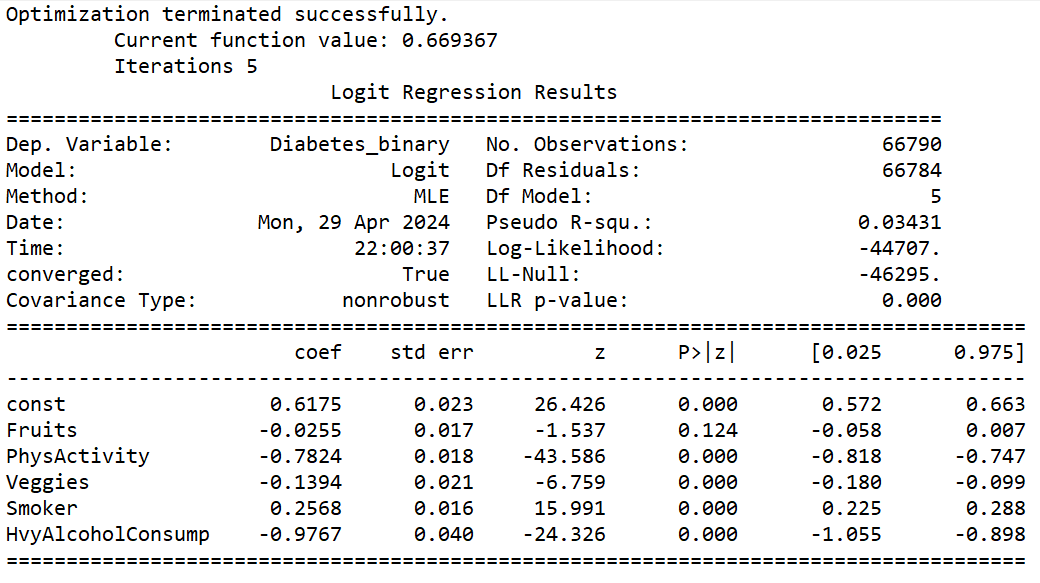
Các biến độc lập dùng để đánh giá là 'Fruits', 'PhysActivity', 'Veggies', 'Smoker' và 'HvyAlcoholConsump'

Biến phụ thuộc: Diabetes\_binary

Ta đặt một giả thuyết null như sau: không có ảnh hưởng của các biến độc lập đến biến phụ thuộc (Nghĩa là các hoạt động về thể chất, thói quen ăn uống, lối sống không lành mạnh không có ảnh hưởng đến nguy cơ mắc bệnh tiểu đường)

Ta sẽ sử dụng mô hình hồi quy logistic đa biến. Mô hình này sẽ đánh giá ảnh hưởng của các biến độc lập đến khả năng mắc bệnh tiểu đường, và cũng cho phép kiểm định giả thuyết về sự ảnh hưởng của từng biến.

Đây là kết quả sau khi huấn luyện mô hình:



Hệ số của "Fruits" (-0.0255) là âm, nhưng không đạt mức ý nghĩa thống kê với p\_value (P>|z|) ≈ 0.124 (> 0.05). Điều này có nghĩa rằng không có sự liên hệ đáng kể giữa việc tiêu thụ trái cây và nguy cơ mắc bệnh tiểu đường.

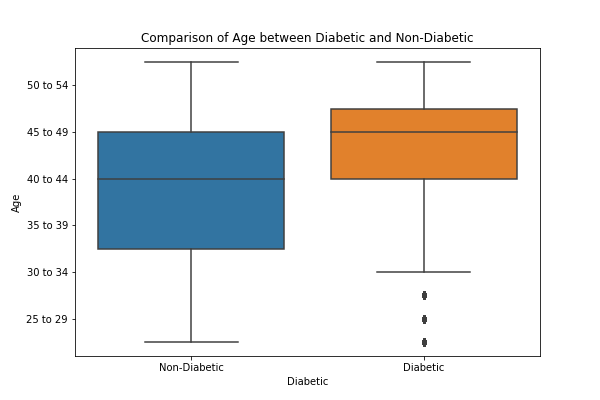
Hệ số của “PhysActivity” (-0.7824), “Veggies” (-0.1394) và “HvyAlcoholConsump” (-0.9767) là âm và đều đạt mức ý nghĩa thống kê với giá trị p\_value < 0.05. Điều này ngụ ý rằng việc tham gia hoạt động thể chất và tiêu thụ nhiều rau cải liên quan đến sự giảm nguy cơ mắc bệnh tiểu đường.

Hệ số của “Smoker” (0.2568) là dương và cũng đạt mức ý nghĩa thống kê với giá trị p\_value < 0.05. Điều này chỉ ra rằng việc hút thuốc liên quan đến tăng nguy cơ mắc bệnh tiểu đường.

Do đó, ta có thể bác bỏ giả thuyết null và như vậy ta đã chứng minh được các hoạt động về thể chất, thói quen ăn uống, lối sống không lành mạnh có ảnh hưởng đến nguy cơ mắc bệnh tiểu đường.

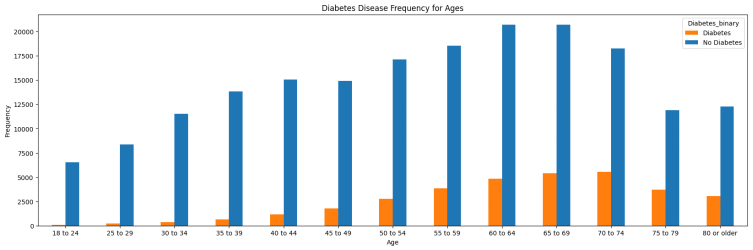
Thật vậy, theo trang tin tức Bệnh viện đa khoa tỉnh Hà Tĩnh, “việc bỏ bữa sáng, lười vận động, ăn thực phẩm chế biến sẵn, thiếu ngủ, căng thẳng có thể khiến bạn nạp nhiều đường, hormone insulin giảm hoạt động… dẫn đến bệnh tiểu đường”[[4]](#footnote-4)[3]. “Theo các nghiên cứu khoa học cho thấy, hút thuốc lá là một trong những nguyên nhân gây ra bệnh tiểu đường type 2. Nguy cơ mắc bệnh tiểu đường type 2 ở những người hút thuốc lá nhiều hơn từ 30 - 40% so với người không hút”[[5]](#footnote-5)[4]. Đây cũng chính là lí do người ta gọi bệnh tiểu đường là “Bệnh lối sống”.

### Độ tuổi càng cao thì càng dễ mắc bệnh tiểu đường đúng không? So sánh với thực trạng tại Việt Nam hiện nay

****

Biểu đồ II.7: Biểu đồ so sánh độ tuổi giữa nhóm mắc bệnh và không mắc bệnh tiểu đường

Chúng ta biết rằng khi tuổi tác tăng lên, nguy cơ mắc bệnh tiểu đường cũng thường tăng lên. Nhìn biểu đồ trên có thể nói, độ tuổi trung bình của người mắc bệnh tiểu đường nằm trong khoảng 45-49 tuổi.



Biểu đồ II.8: Biểu đồ thể hiện tần suất bệnh tiểu đường theo độ tuổi

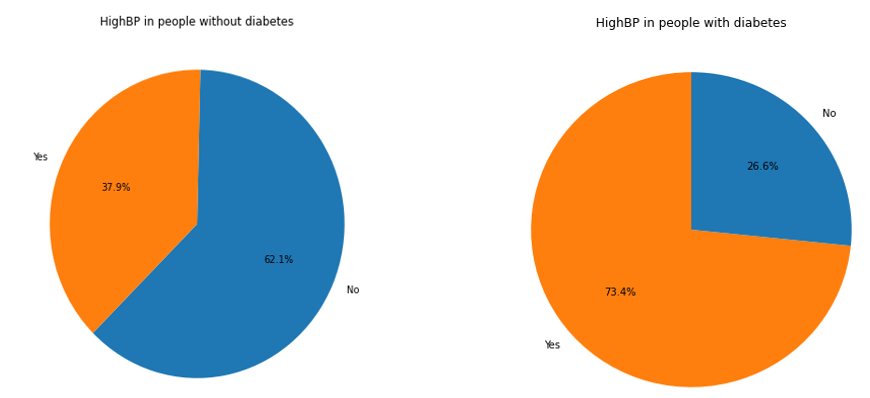
Theo quan niệm trước đây: “Tiểu đường là bệnh của người cao tuổi”. Hiện tại, quan niệm này đã không còn đúng, bởi đối tượng mắc tiểu đường đang ngày càng trẻ hóa. Độ tuổi trung bình của người bị tiểu đường khi bắt đầu mắc bệnh là từ 45 – 64 tuổi. Độ tuổi thường xuyên được chẩn đoán mắc tiểu đường type 1 là từ khi mới sinh đến 30 tuổi. Theo thống kê, tỷ lệ mắc bệnh cao nhất trong độ tuổi từ 10 – 15.

Bệnh tiểu đường type 2 thường diễn ra âm thầm và khó phát hiện hơn tiểu đường type 1. Tỷ lệ mắc bệnh tiểu đường type 2 là 90 – 95% tổng số ca bệnh. Độ tuổi khởi phát bệnh thường ở tuổi 40 và gia tăng nhanh chóng ở những người trong nhóm tuổi từ 45 – 65 tuổi.

Độ tuổi được chẩn đoán mắc bệnh tiểu đường thai kỳ nhiều nhất là trên 25 tuổi. Tỷ lệ mắc bệnh càng cao với các bà mẹ lớn tuổi.

### Bệnh tiểu đường có thể là nguồn gốc cho những căn bệnh khác hay không? Dữ liệu có thể hiện được điều đó hay không?

#### Tăng huyết áp

****

Biểu đồ II.9: Biểu đồ thể hiện mối tương quan giữa bệnh tiểu đường và cao huyết áp

Dựa theo 2 biểu đồ trên, ta thấy rằng đối với những người không bị bệnh tiểu đường, thì người không cao huyết áp chiếm 62.1%, người cao huyết áp chiếm 37.9%. Đối với những người bị bệnh tiểu đường, thì người không cao huyết áp chỉ chiếm 26.6%, người cao huyết áp lên đến 73.4%. Tỉ lệ này đã cho ta thấy bệnh tiểu đường có mối tương quan cao với tăng huyết áp. Những người tiểu đường thường sẽ có chỉ số huyết áp rất cao.

Các nghiên cứu đã chỉ ra rằng, có ít nhất 1 trong 3 bệnh nhân mắc bệnh tiểu đường loại 1 sẽ bị tăng huyết áp.

Khi tăng huyết áp và bệnh tiểu đường cùng tồn tại, 2 bệnh này có xu hướng ảnh hưởng lẫn nhau khiến tình trạng sức khỏe trở nên tồi tệ. Sự “kết hợp” này có thể nguy hiểm đến tính mạng, làm tăng nguy cơ đau tim hoặc đột quỵ...

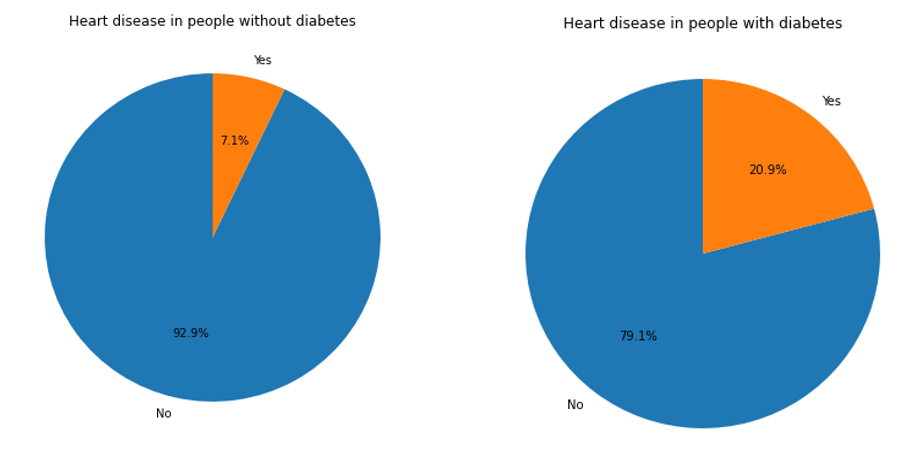
Bệnh tiểu đường có thể làm tăng huyết áp khi:

* Làm giảm khả năng co giãn của mạch máu
* Tăng lượng chất lỏng trong cơ thể
* Thay đổi cách cơ thể quản lý insulin

Tăng huyết áp và tiểu đường thường cùng tồn tại vì chúng có chung các yếu tố nguy cơ bao gồm: thừa cân, chế độ ăn uống không lành mạnh, hút thuốc lá và lối sống ít hoạt động thể chất.

https://medihub.vn/bai-viet/moi-lien-quan-giua-tieu-duong-va-huyet-ap-cao/

#### Bệnh đau tim

****

Biểu đồ II.10: Biểu đồ thể hiện mối tương quan giữa bệnh tiểu đường và bệnh đau tim

Dựa theo 2 biểu đồ trên, ta thấy rằng đối với những người không bị bệnh tiểu đường, thì người mắc bệnh đau tim chiếm 7.1%, người không mắc bệnh đau tim chiếm 92.9%. Đối với những người bị bệnh tiểu đường, thì người mắc bệnh đau tim chiếm 20.9%, người không mắc bệnh đau tim chiếm 79.1%. Ở những người bị bệnh tiểu đường, phần trăm bị bệnh tim cao hơn hẳn so với những người không mắc bệnh (20.9% so với 7.1%). Những người tiểu đường thường có nguy cơ cao mắc bệnh tim.

Thực tế đã chỉ ra rằng , theo thời gian, lượng đường trong máu cao do bệnh tiểu đường có thể phá hủy các mạch máu và dây thần kinh có thức năng kiểm soát tim và các mạch máu trên cơ thể bạn. Bạn bị tiểu đường càng lâu thì khả năng mắc bệnh tim càng cao.

Người lớn mắc bệnh tiểu đường có nguy cơ tử vong cao do biến chứng của bệnh tim hoặc đột quỵ cao gần gấp đôi so với người không mắc bệnh tiểu đường.

Tuy nhiên, một số biện pháp có thể kiểm soát bệnh tiểu đường cũng giúp bạn làm giảm nguy cơ mắc bệnh tim mạch hoặc đột quỵ

Một số lý do khiến bệnh tiểu đường làm tăng nguy cơ mắc bệnh đau tim.

Tăng huyết áp: Người mắc bệnh tiểu đường thường có nguy cơ cao hơn mắc tăng huyết áp. Khi hai yếu tố này kết hợp, nguy cơ mắc bệnh tim mạch tăng lên.

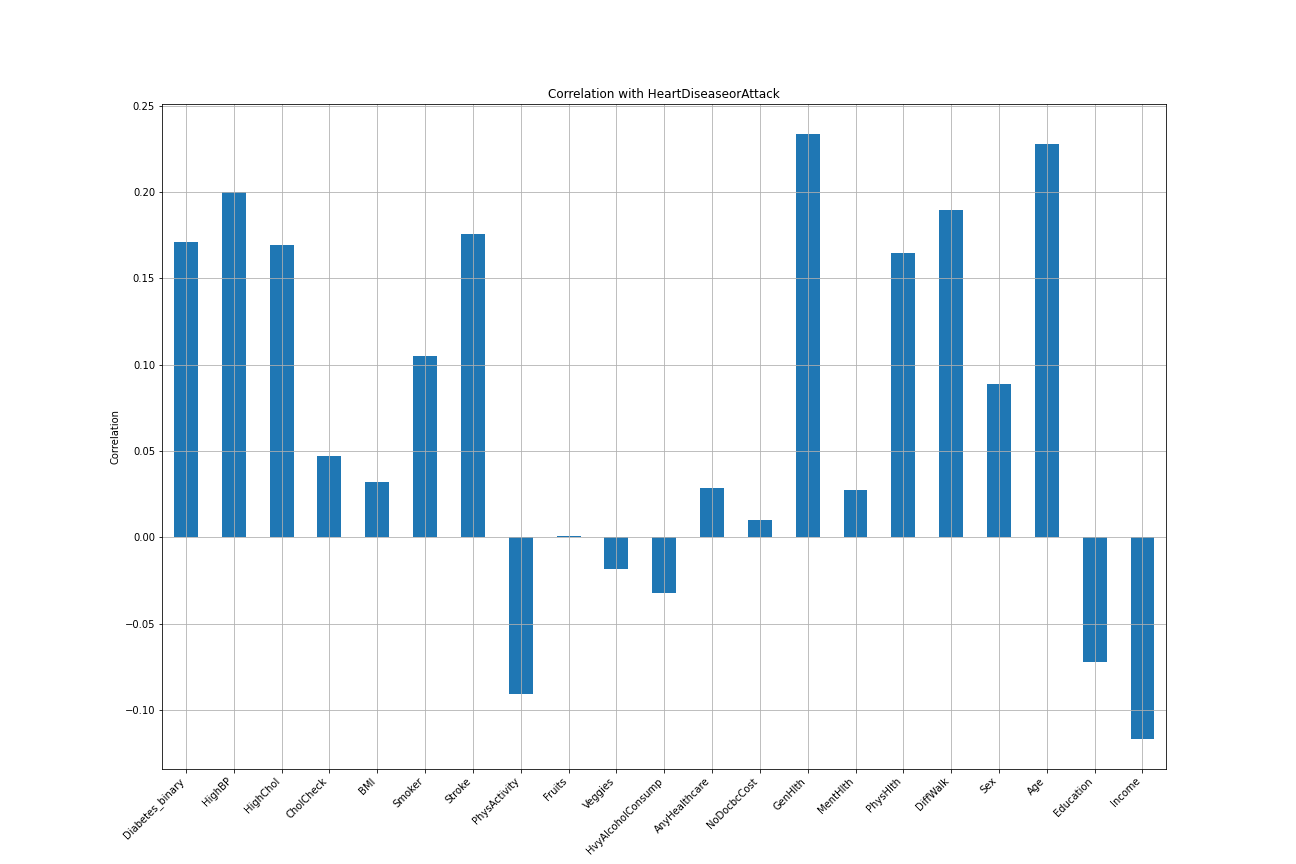
Cholesterol bất thường và triglycerides cao: Người mắc bệnh tiểu đường thường có mức cholesterol LDL (“xấu”) cao, mức cholesterol HDL (“tốt”) thấp và triglycerides cao. Điều này thường xảy ra ở người mắc bệnh tim mạch sớm.

Béo phì: Béo phì là một yếu tố nguy cơ lớn cho bệnh tim mạch và liên quan mật thiết đến kháng insulin.

Thiếu hoạt động vận động: Thiếu hoạt động vận động là một yếu tố nguy cơ khác cho kháng insulin và bệnh tim mạch.

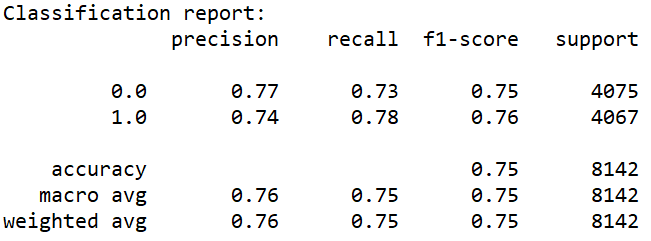
Kiểm soát đường huyết không tốt: Khi đường huyết không được kiểm soát tốt (quá cao), nguy cơ mắc bệnh tim mạch tăng lên.

Hút thuốc: Hút thuốc làm tăng nguy cơ mắc bệnh tim mạch và đột quỵ, bất kể có bị tiểu đường hay không.



Biểu đồ II.11: Biểu đồ thể hiện sự tương quan của HeartDiseaseorAttack với các thuộc tính khác

Sử dụng thuật toán Logistic Regression, độ chính xác dự đoán: 75.71%



#### Đột quỵ

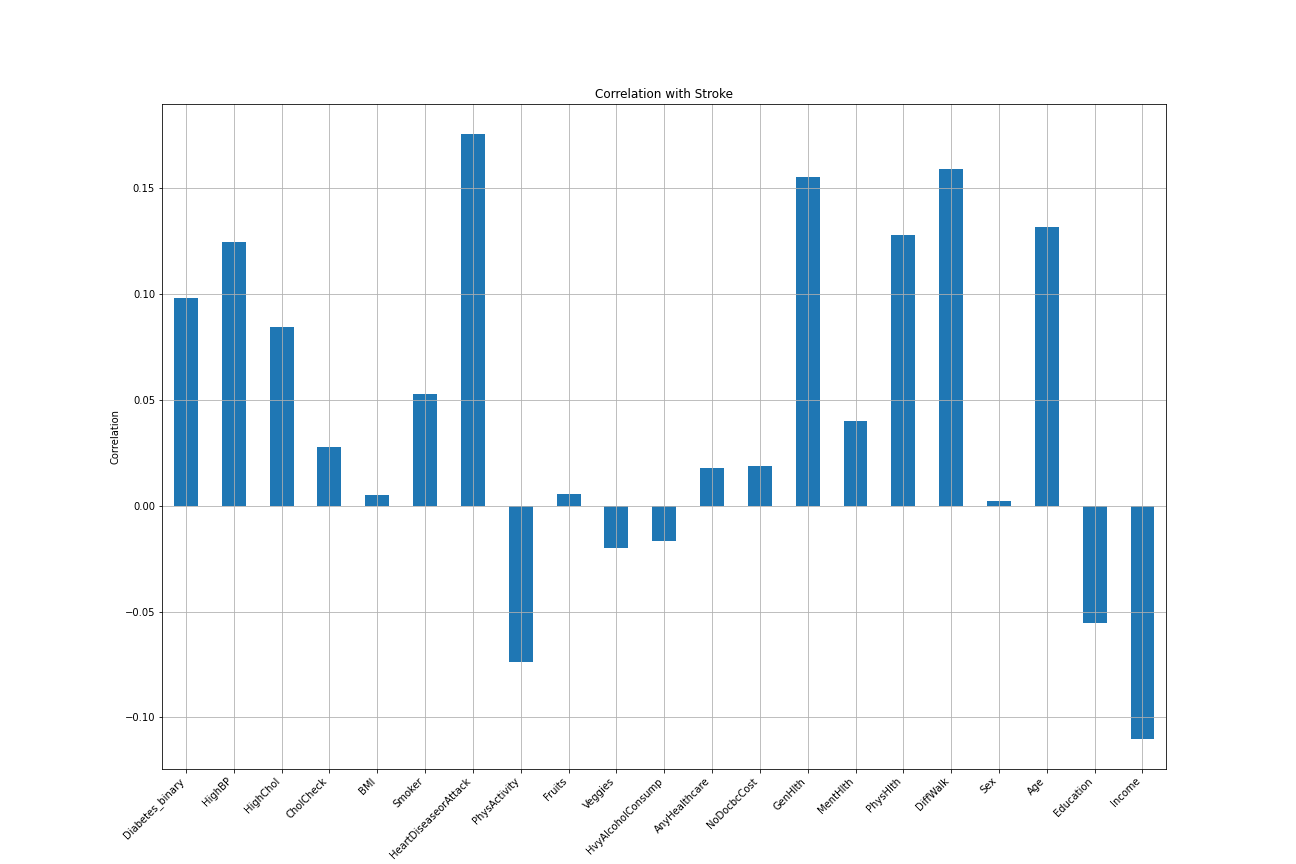
“Đột quỵ xảy ra khi nguồn cung cấp máu cho một phần não của bạn đột nhiên bị gián đoạn. Sau đó mô não bị tổn thương. Hầu hết các cơn đột quỵ xảy ra do cục máu đông chặn một mạch máu trong não hoặc cổ. Đột quỵ có thể gây ra các vấn đề về vận động, đau, tê và các vấn đề về khả năng suy nghĩ, ghi nhớ hoặc nói. Một số người cũng có vấn đề về cảm xúc, chẳng hạn như trầm cảm sau một cơn đột quỵ.

Nếu bạn bị tiểu đường, khả năng bạn bị đột quỵ cao gấp 1,5 lần so với những người không mắc bệnh tiểu đường. Nhưng bạn có thể giảm thiểu rủi ro bằng cách chăm sóc sức khỏe của mình một cách hợp lý.”[[6]](#footnote-6)[5]

Một số yếu tố nguy cơ gây tai biến ở người tiểu đường đó là:

* Lười vận động: Những người lười vận động thì nguy cơ tim mạch tăng cao hơn vì tăng sản xuất cholesterol - một loại mỡ trong máu có thể gây tích tụ trên thành mạch máu nếu như bạn bị béo phì do lười vận động.
* Cholesterol cao hơn bình thường: Hầu hết các bệnh nhân đái tháo đường thường bị nhiễm mỡ máu có lượng cholesterol trong máu vượt quá mức cho phép, nhất là cholesterol có hại làm cho quá trình xơ vữa động mạch càng trở nên trầm trọng hơn, nguy cơ tai biến mạch máu não càng cao. Điều quan trọng là các yếu tố này thường xuất hiện đồng thời ở bệnh nhân đái tháo đường.
* Tăng huyết áp: Khi bị tăng huyết áp, tim phải hoạt động nhiều hơn để bơm máu. Tăng huyết áp có thể làm cho tim căng giãn và tổn thương mạch máu và tăng nguy cơ nhồi máu cơ tim, tai biến mạch máu não, biến chứng mắt và thận.
* Hút thuốc lá: Những người hút thuốc lá sẽ làm tăng nguy cơ bị tai biến mạch máu não, vì vậy, việc ngưng thuốc lá đặc biệt quan trọng đối với bệnh nhân đái tháo đường vì hút thuốc lá và đái tháo đường cùng làm hẹp mạch máu.
* Tiền sử gia đình có người bị tai biến mạch máu não: Bạn không thể thay đổi nguy cơ này do yếu tố gia đình.

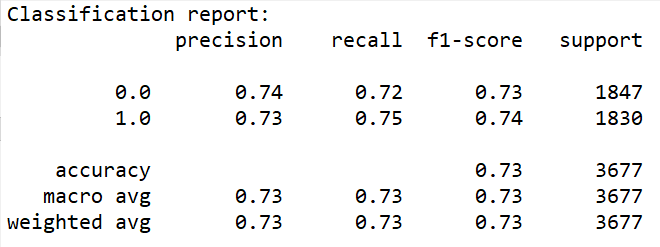
Một khi các yếu tố này kết hợp với bệnh tiểu đường thì sẽ làm tăng nguy cơ tai biến mạch máu não lên nhiều lần. Chính vì vậy, ngoài yếu tố gia đình không thể thay đổi, khi điều trị bệnh tiểu đường không bao giờ được bỏ quên tác động vào các yếu tố nguy cơ này, đặc biệt là thói quen hút thuốc lá, ít vận động, do vậy nguy cơ tai biến là rất có thể nếu không được kiểm soát và điều trị đúng đắn[[7]](#footnote-7)[6].



Biểu đồ II.12: Biểu đồ thể hiện sự tương quan của Stroke với các thuộc tính khác

Dựa theo trên, ta sẽ thử sử dụng dữ liệu trên để dự đoán tai biến. Sử dụng thuật toán Logistic Regression, độ chính xác: 74.46%

Classification report:



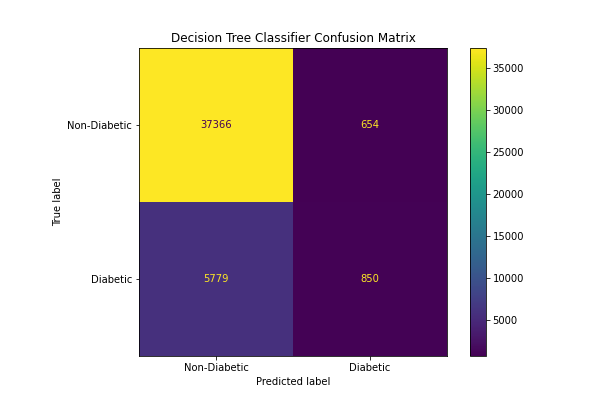
Tuy accuracy đạt được là 96%, nhưng khi có sự mất cân bằng trong số lượng mẫu giữa các lớp, accuracy có thể không phản ánh chính xác hiệu suất của mô hình.

### Việc xây dựng mô hình chẩn đoán bệnh tiểu đường thông qua tập dữ liệu trên có cho ra kết quả tốt với độ chính xác > 90% theo phương thức đánh giá độ chính xác là F1-score hoặc công thức (Số dự đoán đúng / Tổng số \* 100%) hay không? Với mô hình có độ chính xác tốt nhất có nên ứng dụng vào các phòng khám, bệnh viện hay không?

* Độ chính xác với công thức: Độ chính xác = Số dự đoán đúng/Tổng số dự đoán
* Ma trận nhầm lẫn: là một công cụ quan trọng trong machine learning và thống kê, được sử dụng để đánh giá hiệu suất của một mô hình phân loại. Nó thường được sử dụng để so sánh các dự đoán của mô hình với các giá trị thực tế của dữ liệu. Ma trận này bao gồm các ô, trong đó hàng của ma trận biểu thị các nhãn thực tế và cột biểu thị các nhãn được dự đoán bởi mô hình. Các ô trên đường chéo chính của ma trận thường chứa số lượng các trường hợp được phân loại đúng, trong khi các ô còn lại chứa các trường hợp được phân loại sai:
* Báo cáo bao gồm các thông tin sau:
* Precision: (true positive) / (true positive + false positive). Đây là tỷ lệ giữa số lượng các trường hợp dự đoán đúng thuộc một lớp so với tổng số trường hợp được dự đoán là thuộc lớp đó. Nó đo lường mức độ chính xác của các dự đoán tích cực, tức là dự đoán một trường hợp là positive (1.0) mà thực sự là positive
* Recall: (true positive) / (true positive + false negative). Đây là tỷ lệ giữa số lượng các trường hợp dự đoán đúng thuộc một lớp so với tổng số trường hợp thực sự thuộc lớp đó. Nó đo lường khả năng của mô hình phát hiện tất cả các trường hợp tích cực có sẵn trong dữ liệu.
* F1-score: 2 ∗ (precision∗recall​) / (precision+recall). Trung bình điều hòa giữa precision và recall. Nó là một phép đo tổng hợp của hiệu suất phân loại, chủ yếu được sử dụng khi precision và recall có trọng số quan trọng tương đương.
* Support: Số lượng mẫu trong mỗi lớp.
* Macro avg: Trung bình không quan trọng hơn cho mỗi lớp.
* Weighted avg: Trung bình được tính toán dựa trên trọng số của mỗi lớp, phụ thuộc vào số lượng mẫu của từng lớp.

#### Xây dựng mô hình dự đoán khi dữ liệu chưa cân bằng

* Áp dụng Thuật toán Decision Tree Classifier xây dựng mô hình phân loại Bệnh Tiểu đường và Sử dụng RandomizedSearchCV để Tinh chỉnh Tham số.
* Độ chính xác với công thức: Độ chính xác = Số dự đoán đúng/Tổng số dự đoán: 85.59%

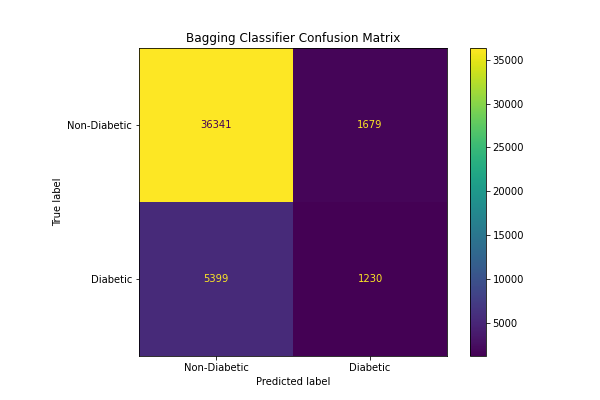


Biểu đồ II.13: Ma trận nhầm lẫn

Báo cáo:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | precision | recall | f1-score | support |
| 0.0 | 0.87 | 0.98 | 0.92 | 38020 |
| 1.0 | 0.57 | 0.13 | 0.21 | 6629 |
| accuracy |  |  | 0.86 | 44649 |
| macro avg | 0.65 | 0.56 | 0.56 | 44649 |
| weighted avg | 0.82 | 0.86 | 0.82 | 44649 |

* Áp dụng Thuật toán Bagging Classifier xây dựng mô hình phân loại Bệnh Tiểu đường và Sử dụng RandomizedSearchCV để Tinh chỉnh Tham số:
* Độ chính xác với công thức: Độ chính xác = Số dự đoán đúng/Tổng số dự đoán: 84.15%

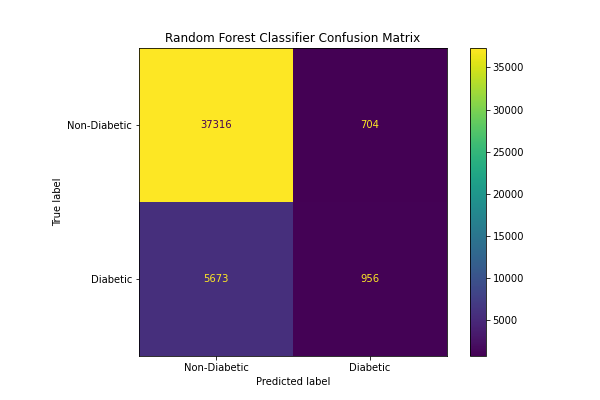


Biểu đồ II.14: Ma trận nhầm lẫn

Báo cáo

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | precision | recall | f1-score | support |
| 0.0 | 0.87 | 0.96 | 0.91 | 38020 |
| 1.0 | 0.42 | 0.19 | 0.26 | 6629 |
| accuracy |  |  | 0.84 | 44649 |
| macro avg | 0.65 | 0.57 | 0.58 | 44649 |
| weighted avg | 0.80 | 0.84 | 0.81 | 44649 |

* Áp dụng Thuật toán Random Forest Classifier xây dựng mô hình phân loại Bệnh Tiểu đường và Sử dụng RandomizedSearchCV để Tinh chỉnh Tham số.
* Độ chính xác với công thức: Độ chính xác = Số dự đoán đúng/Tổng số dự đoán: 85.72%

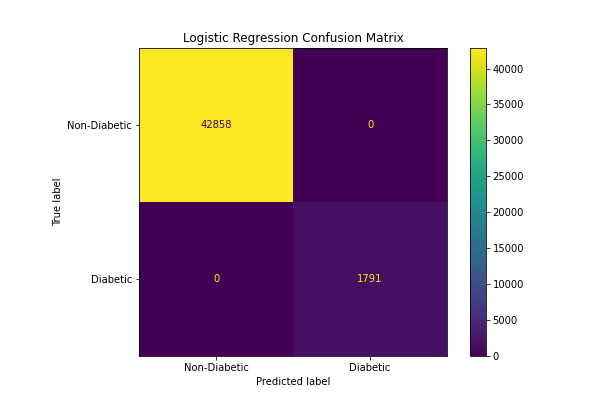


Ma trận nhầm lẫn

Báo cáo

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | precision | recall | f1-score | support |
| 0.0 | 0.87 | 0.98 | 0.92 | 38020 |
| 1.0 | 0.58 | 0.14 | 0.23 | 6629 |
| accuracy |  |  | 0.86 | 44649 |
| macro avg | 0.72 | 0.56 | 0.58 | 44649 |
| weighted avg | 0.82 | 0.86 | 0.82 | 44649 |

* Thuật toán Linear Regression
* Độ chính xác với công thức: Độ chính xác = Số dự đoán đúng/Tổng số dự đoán: 85.52%

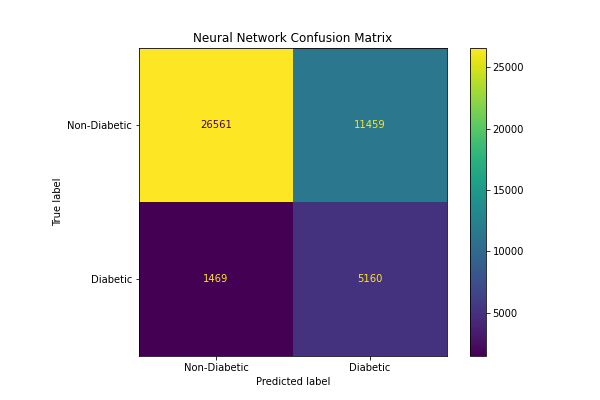


Biểu đồ II.15: Ma trận nhầm lẫn

Báo cáo

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | precision | recall | f1-score | support |
| 0.0 | 0.87 | 0.98 | 0.92 | 38020 |
| 1.0 | 0.55 | 0.15 | 0.23 | 6629 |
| accuracy |  |  | 0.86 | 44649 |
| macro avg | 0.71 | 0.56 | 0.58 | 44649 |
| weighted avg | 0.82 | 0.86 | 0.82 | 44649 |

* Thuật toán Neural network:
* Độ chính xác với công thức: Độ chính xác = Số dự đoán đúng/Tổng số dự đoán với threshold = 0.5 là: 71.05%



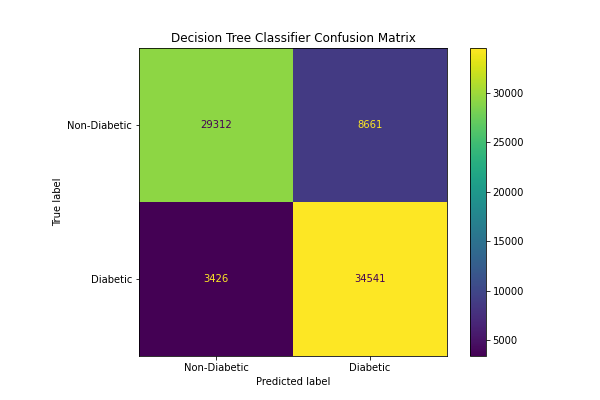
Biểu đồ II.16: Ma trận nhầm lẫn

Báo cáo:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | precision | recall | f1-score | support |
| 0.0 | 0.95 | 0.70 | 0.80 | 38020 |
| 1.0 | 0.31 | 0.78 | 0.44 | 6629 |
| accuracy |  |  | 0.71 | 44649 |
| macro avg | 0.63 | 0.74 | 0.62 | 44649 |
| weighted avg | 0.85 | 0.71 | 0.75 | 44649 |

#### Xây dựng mô hình với dữ liệu được cân bằng thông qua xử lý thêm dữ liệu

* Áp dụng Thuật toán Decision Tree Classifier xây dựng mô hình phân loại Bệnh Tiểu đường và Sử dụng RandomizedSearchCV để Tinh chỉnh Tham số.
* Độ chính xác với công thức: Độ chính xác = Số dự đoán đúng/Tổng số dự đoán: 84.08%

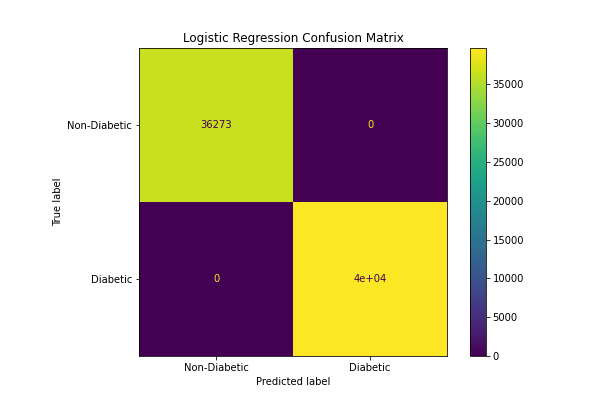


Biểu đồ II.17: Ma trận nhầm lẫn

Báo cáo:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | precision | recall | f1-score | support |
| 0.0 | 0.90 | 0.77 | 0.83 | 37973 |
| 1.0 | 0.80 | 0.91 | 0.85 | 37967 |
| accuracy |  |  | 0.84 | 75940 |
| macro avg | 0.85 | 0.84 | 0.84 | 75940 |
| weighted avg | 0.85 | 0.84 | 0.84 | 75940 |

* Thuật toán Linear Regression
* Độ chính xác với công thức: Độ chính xác = Số dự đoán đúng/Tổng số dự đoán: 73.5%

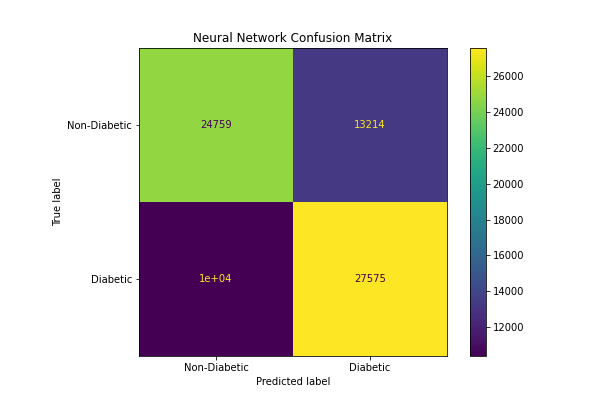


Biểu đồ II.18: Ma trận nhầm lẫn

Báo cáo

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | precision | recall | f1-score | support |
| 0.0 | 0.75 | 0.71 | 0.73 | 37973 |
| 1.0 | 0.72 | 0.76 | 0.74 | 37967 |
| accuracy |  |  | 0.74 | 75940 |
| macro avg | 0.74 | 0.74 | 0.73 | 75940 |
| weighted avg | 0.74 | 0.74 | 0.73 | 75940 |

* Thuật toán Neural network:
* Độ chính xác với công thức: Độ chính xác = Số dự đoán đúng/Tổng số dự đoán với threshold = 0.5 là: 68.91%



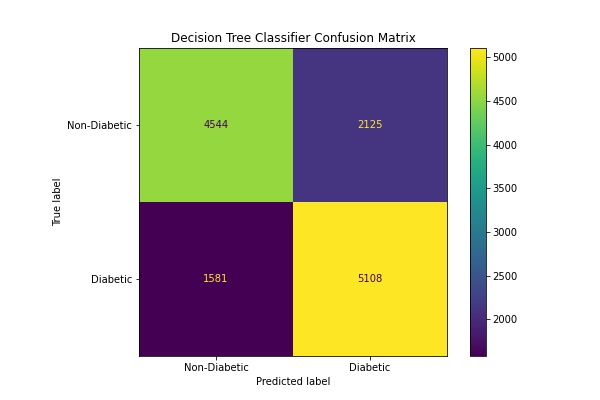
Biểu đồ II.19: Ma trận nhầm lẫn

Báo cáo:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | precision | recall | f1-score | support |
| 0.0 | 0.70 | 0.65 | 0.68 | 37973 |
| 1.0 | 0.68 | 0.73 | 0.70 | 37967 |
| accuracy |  |  | 0.69 | 75940 |
| macro avg | 0.69 | 0.69 | 0.69 | 75940 |
| weighted avg | 0.69 | 0.69 | 0.69 | 75940 |

#### Xây dựng mô hình với dữ liệu được cân bằng bằng cách lựa chọn ra từ tập dữ liệu ban đầu

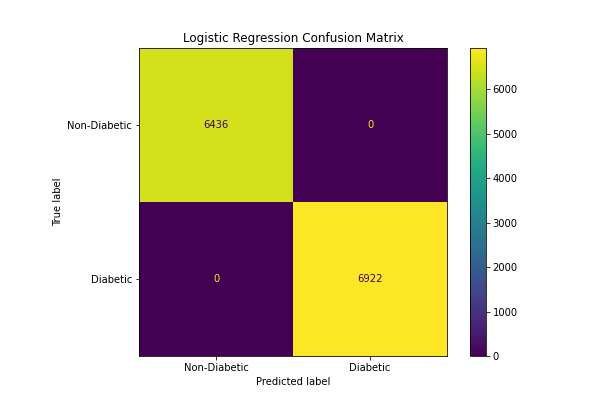
* Áp dụng Thuật toán Decision Tree Classifier xây dựng mô hình phân loại Bệnh Tiểu đường và Sử dụng RandomizedSearchCV để Tinh chỉnh Tham số.
* Độ chính xác với công thức: Độ chính xác = Số dự đoán đúng/Tổng số dự đoán: 72.26%



Báo cáo:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | precision | recall | f1-score | support |
| 0.0 | 0.74 | 0.68 | 0.71 | 6669 |
| 1.0 | 0.71 | 0.76 | 0.73 | 6689 |
| accuracy |  |  | 0.72 | 13358 |
| macro avg | 0.72 | 0.72 | 0.72 | 13358 |
| weighted avg | 0.72 | 0.72 | 0.72 | 13358 |

* Thuật toán Linear Regression
* Độ chính xác với công thức: Độ chính xác = Số dự đoán đúng/Tổng số dự đoán: 72.95%

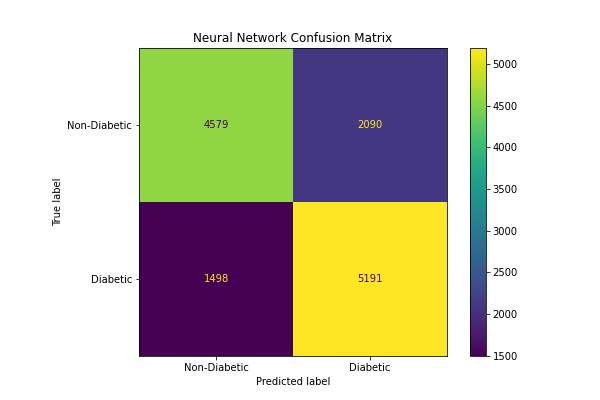


Ma trận nhầm lẫn

Báo cáo

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | precision | recall | f1-score | support |
| 0.0 | 0.74 | 0.71 | 0.72 | 6669 |
| 1.0 | 0.72 | 0.75 | 0.73 | 6689 |
| accuracy |  |  | 0.73 | 13358 |
| macro avg | 0.73 | 0.73 | 0.73 | 13358 |
| weighted avg | 0.73 | 0.73 | 0.73 | 13358 |

* Thuật toán Neural network:
* Độ chính xác với công thức: Độ chính xác = Số dự đoán đúng/Tổng số dự đoán với threshold = 0.5 là: 73.14%



Ma trận nhầm lẫn

Báo cáo:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | precision | recall | f1-score | support |
| 0.0 | 0.75 | 0.69 | 0.72 | 6669 |
| 1.0 | 0.71 | 0.78 | 0.74 | 6689 |
| accuracy |  |  | 0.73 | 13358 |
| macro avg | 0.73 | 0.73 | 0.73 | 13358 |
| weighted avg | 0.73 | 0.73 | 0.73 | 13358 |

Kết luận: Qua 5 thuật toán trên việc xây dựng mô hình chẩn đoán bệnh tiểu đường thông qua tập dữ liệu trên có kết quả với độ chính xác < 90% theo phương thức đánh giá độ chính xác là F1-score hoặc công thức (Số dự đoán đúng / Tổng số \* 100%), cho thấy rằng tập dữ liệu này vẫn chưa đủ tốt để thực hiện đào tạo các mô hình và ứng dụng vào các các bệnh viện, phòng khám, vì độ chính xác không được cao, có thể gây ảnh hưởng lớn đến người thăm khám.

### Những chính sách cần đề ra để phòng tránh bệnh tiểu đường.

Để phòng tránh bệnh tiểu đường, cần triển khai một số chính sách và biện pháp như sau:

* **Chương trình giáo dục và tăng nhận thức**: Tăng cường giáo dục về nguy cơ và cách phòng tránh tiểu đường thông qua các chương trình giáo dục sức khỏe cộng đồng, chiến dịch truyền thông, và các hoạt động tại các cơ sở y tế.
* **Sản phẩm thực phẩm và quảng cáo**: Quy định nghiêm ngặt quảng cáo các sản phẩm thực phẩm không lành mạnh và đồ uống có đường, đồng thời khuyến khích tiêu dùng các sản phẩm thực phẩm có lợi cho sức khỏe.
* **Tăng cường hoạt động thể chất**: Xây dựng và hỗ trợ các chương trình thúc đẩy hoạt động thể chất đều đặn ở cả trẻ em và người lớn, bao gồm cả việc tích hợp hoạt động thể dục vào cuộc sống hàng ngày và việc cải thiện cơ sở hạ tầng thể chất.
* **Quản lý cân nặng và dinh dưỡng**: Thúc đẩy việc duy trì cân nặng lành mạnh thông qua chính sách và chương trình hỗ trợ việc thực hiện chế độ ăn uống cân đối và lành mạnh.
* **Kiểm soát căn bệnh liên quan**: Tăng cường kiểm soát các yếu tố rủi ro liên quan đến bệnh tiểu đường như huyết áp cao, cholesterol cao, và béo phì thông qua các chương trình sàng lọc và điều trị.
* **Hỗ trợ hành vi thay đổi lối sống**: Cung cấp các chương trình hỗ trợ để giúp người dân thay đổi lối sống không lành mạnh, bao gồm cả hỗ trợ hút thuốc, giảm căng thẳng, và quản lý stress.
* **Tăng cường tiếp cận chăm sóc y tế**: Cải thiện tiếp cận và chất lượng chăm sóc y tế cho những người có nguy cơ cao và những người đã mắc bệnh tiểu đường, bao gồm cả việc cung cấp dịch vụ chăm sóc đa ngành nghề và quản lý bệnh tốt hơn.
* **Chính sách thuế và giá cả**: Áp dụng chính sách thuế và giá cả để ưu tiên các sản phẩm và dịch vụ tạo động lực cho lối sống lành mạnh và phòng tránh bệnh tiểu đường.
* **Nghiên cứu và phát triển**: Hỗ trợ nghiên cứu và phát triển các phương pháp mới để phòng tránh, chẩn đoán sớm và điều trị bệnh tiểu đường.
* **Hợp tác toàn cầu**: Hợp tác với cộng đồng quốc tế để chia sẻ kinh nghiệm, thông tin và tài nguyên để tăng cường phòng tránh và quản lý bệnh tiểu đường[[8]](#footnote-8)[7][[9]](#footnote-9)[8].

## Kết luận

Dữ liệu được thu thập bởi Behavioral Risk Factor Surveillance System (BRFSS), một hệ thống khảo sát điện thoại dựa trên tiểu bang được tiến hành hàng năm bởi Trung tâm Kiểm soát và Phòng ngừa Dịch bệnh (CDC) của Hoa Kỳ. Do đó, dữ liệu này được coi là đáng tin cậy, giúp cho kết quả nghiên cứu được công nhận và chấp nhận rộng rãi hơn. Kích thước lớn với 236378 mẫu, 22 thuộc tính khác nhau, do đó nó có sự đa dạng và tiềm năng sử dụng rộng rãi. Các thuộc tính trong bộ dữ liệu này cung cấp một cái nhìn toàn diện về các yếu tố có thể ảnh hưởng đến tiểu đường và sức khỏe tổng thể. Bộ dữ liệu đã được làm sạch, loại bỏ các giá trị bị thiếu và các ngoại lệ, giúp cho việc phân tích và xử lý dữ liệu dễ dàng hơn. Bộ dữ liệu chưa cân bằng lớp.Vì vậy, dữ liệu phản ánh thực tế tỷ lệ mắc bệnh trong dân số, nhưng có thể gây ra các vấn đề khi huấn luyện mô hình do sự thiếu cân bằng. Vì vậy để nắm bắt thực tế tỷ lệ mắc bệnh trong dân số, thì việc sử dụng bộ dữ liệu không cân bằng lớp là lựa chọn phù hợp. Sau khi phân tích đánh giá bộ dữ liệu và sử dụng mô hình để dự đoán bệnh tiểu đường, ta thấy rằng không có mô hình nào đạt được độ chính xác > 90% theo cả hai phương thức đánh giá độ chính xác là F1-score và công thức (Số dự đoán đúng / Tổng số \* 100%). Các thuật toán đã thử là Decision Tree Classifier, Bagging Classifier, Random Forest Classifier, Linear Regression và Neural Network. Dù có sử dụng các phương pháp khác nhau như cân bằng dữ liệu thông qua xử lý hoặc lựa chọn thuật toán khác nhau, không có mô hình nào đạt được mục tiêu độ chính xác > 90%. Điều này có thể ám chỉ rằng tập dữ liệu hoặc phương pháp huấn luyện cần được cải thiện hoặc mở rộng hơn để đạt được mục tiêu đó.

# TÀI LIỆU THAM KHẢO

1. Bệnh viện Tâm Trí Đồng Tháp, “Nguyên nhân và triệu chứng của bệnh tiểu đường,” 20/02/2023. [Trực tuyến]. Địa chỉ: <https://bvtamtridongthap.com.vn/vn/nguyen-nhan-va-trieu-chung-cua-benh-tieu-duong.html> [Truy cập 29/4/2024].
2. Báo Điện tử Chính phủ, “Ghi nhận nhiều trẻ từ 7-18 tuổi mắc đái tháo đường,” 12/02/2023. [Trực tuyến]. Địa chỉ: <https://baochinhphu.vn/ghi-nhan-nhieu-tre-tu-7-18-tuoi-mac-dai-thao-duong-102230212170507779.htm> [Truy cập ngày 29/4/2024].
3. Bệnh viện đa khoa tỉnh Hà Tĩnh, “5 thói quen làm tăng nguy cơ mắc bệnh tiểu đường,” 05/10/2023. [Trực tuyến]. Địa chỉ: <https://bvdkht.vn/news/view/5-thoi-quen-lam-tang-nguy-co-mac-benh-tieu-duong/> [Truy cập ngày 29/4/2024].
4. Nhà thuốc Long Châu, “Bệnh tiểu đường có liên quan đến hút thuốc lá không?,” 17/02/2022. [Trực tuyến]. Địa chỉ: <https://nhathuoclongchau.com.vn/bai-viet/benh-tieu-duong-co-lien-quan-den-hut-thuoc-la-khong-50643.html> [Truy cập ngày 29/4/2024].
5. VINMEC, “Các biến chứng nguy hiểm của bệnh tiểu đường,”. [Trực tuyến]. Địa chỉ: <https://www.vinmec.com/vi/tin-tuc/thong-tin-suc-khoe/suc-khoe-tong-quat/cac-bien-chung-nguy-hiem-cua-benh-tieu-duong/> [Truy cập ngày 29/4/2024].
6. [VINMEC, “Tai biến mạch máu não ở người tiểu đường,”. [Trực tuyến]. Địa chỉ: <https://www.vinmec.com/vi/tim-mach/thong-tin-suc-khoe/tai-bien-mach-mau-nao-o-nguoi-tieu-duong/> [Truy cập ngày 29/4/2024].

Bộ Y tế, “Cách phòng tránh bệnh tiểu đường,” 06/12/2015. [Trực tuyến]. Địa chỉ: <https://moh.gov.vn/chuong-trinh-muc-tieu-quoc-gia/-/asset_publisher/7ng11fEWgASC/content/cach-phong-tranh-benh-tieu-uong?inheritRedirect=false> [Truy cập ngày 29/4/2024].

1. Bộ Y tế, “12 cách đơn giản phòng bệnh đái tháo đường,” 09/01/2020. [Trực tuyến]. Địa chỉ: <https://moh.gov.vn/chuong-trinh-muc-tieu-quoc-gia/-/asset_publisher/7ng11fEWgASC/content/12-cach-on-gian-phong-benh-ai-thao-uong> [Truy cập ngày 29/4/2024].

1. [1] Bệnh viện Tâm Trí Đồng Tháp, “Nguyên nhân và triệu chứng của bệnh tiểu đường,” 20/02/2023. [Trực tuyến]. Địa chỉ: <https://bvtamtridongthap.com.vn/vn/nguyen-nhan-va-trieu-chung-cua-benh-tieu-duong.html> [Truy cập 29/4/2024]. [↑](#footnote-ref-1)
2. [1] Bệnh viện Tâm Trí Đồng Tháp, “Nguyên nhân và triệu chứng của bệnh tiểu đường,” 20/02/2023. [Trực tuyến]. Địa chỉ: <https://bvtamtridongthap.com.vn/vn/nguyen-nhan-va-trieu-chung-cua-benh-tieu-duong.html> [Truy cập 29/4/2024]. [↑](#footnote-ref-2)
3. [2] Báo Điện tử Chính phủ, “Ghi nhận nhiều trẻ từ 7-18 tuổi mắc đái tháo đường,” 12/02/2023. [Trực tuyến]. Địa chỉ: <https://baochinhphu.vn/ghi-nhan-nhieu-tre-tu-7-18-tuoi-mac-dai-thao-duong-102230212170507779.htm> [Truy cập ngày 29/4/2024]. [↑](#footnote-ref-3)
4. [3] Bệnh viện đa khoa tỉnh Hà Tĩnh, “5 thói quen làm tăng nguy cơ mắc bệnh tiểu đường,” 05/10/2023. [Trực tuyến]. Địa chỉ: <https://bvdkht.vn/news/view/5-thoi-quen-lam-tang-nguy-co-mac-benh-tieu-duong/> [Truy cập ngày 29/4/2024]. [↑](#footnote-ref-4)
5. [4] Nhà thuốc Long Châu, “Bệnh tiểu đường có liên quan đến hút thuốc lá không?,” 17/02/2022. [Trực tuyến]. Địa chỉ: <https://nhathuoclongchau.com.vn/bai-viet/benh-tieu-duong-co-lien-quan-den-hut-thuoc-la-khong-50643.html> [Truy cập ngày 29/4/2024]. [↑](#footnote-ref-5)
6. [5] VINMEC, “Các biến chứng nguy hiểm của bệnh tiểu đường,”. [Trực tuyến]. Địa chỉ: <https://www.vinmec.com/vi/tin-tuc/thong-tin-suc-khoe/suc-khoe-tong-quat/cac-bien-chung-nguy-hiem-cua-benh-tieu-duong/> [Truy cập ngày 29/4/2024]. [↑](#footnote-ref-6)
7. [6] VINMEC, “Tai biến mạch máu não ở người tiểu đường,”. [Trực tuyến]. Địa chỉ: <https://www.vinmec.com/vi/tim-mach/thong-tin-suc-khoe/tai-bien-mach-mau-nao-o-nguoi-tieu-duong/> [Truy cập ngày 29/4/2024]. [↑](#footnote-ref-7)
8. [7] Bộ Y tế, “Cách phòng tránh bệnh tiểu đường,” 06/12/2015. [Trực tuyến]. Địa chỉ: <https://moh.gov.vn/chuong-trinh-muc-tieu-quoc-gia/-/asset_publisher/7ng11fEWgASC/content/cach-phong-tranh-benh-tieu-uong?inheritRedirect=false> [Truy cập ngày 29/4/2024]. [↑](#footnote-ref-8)
9. [8] Bộ Y tế, “12 cách đơn giản phòng bệnh đái tháo đường,” 09/01/2020. [Trực tuyến]. Địa chỉ: <https://moh.gov.vn/chuong-trinh-muc-tieu-quoc-gia/-/asset_publisher/7ng11fEWgASC/content/12-cach-on-gian-phong-benh-ai-thao-uong> [Truy cập ngày 29/4/2024]. [↑](#footnote-ref-9)