ĐẠI HỌC QUỐC GIA TP. HỒ CHÍ MINH

TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN

BÁO CÁO ĐỒ ÁN

MÔN XỬ LÝ SỐ LIỆU THỐNG KÊ

Project 2: CDC Diabetes Health Indicators

Giảng viên môn học: Tô Đức Khánh

Nhóm 11

22280004 - Trương Bình Ba

22280018 - Chiêm Huỳnh Giao

22280020 - Nguyễn Ngọc Bảo Hân

22280037 - Nguyễn Thị Xuân Hương

22280052 - Phan Thị Ngọc Linh

22280088 - Hồ Trần Anh Thư

Tp. Hồ Chí Minh, 13 tháng 01 năm 2025

Mục lục

A. MỞ ĐẦU

Tiểu đường – kẻ thù thầm lặng, đang lặng lẽ lan rộng và gây ra những hệ lụy nghiêm trọng đối với sức khỏe cộng đồng trên toàn cầu. Không chỉ làm suy giảm chất lượng cuộc sống, căn bệnh này còn kéo theo hàng loạt biến chứng nguy hiểm như bệnh tim, suy thận, mù lòa và thậm chí là cắt cụt chi. Điều đáng lo ngại hơn, tỷ lệ mắc bệnh vẫn không ngừng gia tăng, đặc biệt tại các quốc gia đang phát triển.

Giữa bối cảnh đó, việc nghiên cứu và phân tích các yếu tố nguy cơ trở thành một nhiệm vụ cấp bách, không chỉ để cải thiện hiểu biết về căn bệnh mà còn để phát triển những công cụ dự đoán hiệu quả. Mục tiêu là làm sáng tỏ bản chất đa chiều của các yếu tố nguy cơ bệnh tiểu đường, từ các khía cạnh y sinh học đến kinh tế - xã hội. Từ đó đưa ra các chiến lược, chẩn đoán sớm để thay đổi lối sống và điều trị hiệu quả hơn.

B. NỘI DUNG

I. MỤC TIÊU CẦN ĐẠT ĐƯỢC:

- Khám phá sự đóng góp của các biến đặc trưng lên biến mục tiêu và tìm hiểu mối quan hệ giữa các biến.
- Sử dụng AB testing để kiểm chứng các giả thiết đặt ra trong quá trình làm.
- Đề xuất các chiến lược thay đổi lối sống làm tăng nguy cơ mắc bệnh tiểu đường.
- Xây dựng được mô hình dự đoán tốt nhất các nguy cơ có tác động lớn đến nguy cơ mắc bệnh, hỗ trợ con người trong việc điều trị bệnh tiểu đường.

II. SƠ LƯỢC VÀ MÔ TẢ BỘ DỮ LIỆU

1. Sơ lược về bộ dữ liệu

diabetes _ 012 _ health _ indications _ BRFSS2015.csv là một tập dữ liệu sạch gồm 253680 phản hồi khảo sát cho BRFSS2015 của CDC.

- Biến mục tiêu ‘diabetes_012’ có 3 lớp.
  - 0 là không mắc bệnh tiểu đường (“Non-diabete”)
  - 1 là tiền tiểu đường (“Pre-diabete”)
  - 2 là tiểu đường. (“Diabete”)
- Có sự mất cân bằng nghiêm trọng giữa các lớp trong tập dữ liệu này.
- Tập dữ liệu gồm 21 biến đặc trưng.

Mô tả của các biến trong bộ dữ liệu, với các mốc thời gian (30 ngày, 5 năm) quanh thời gian mà bộ dữ liệu này được phát hành.

- **diabetes_012:** Không tiểu đường (0), tiền tiểu đường (1), tiểu đường (2)
- **high_bp:** Tình trạng cao huyết áp
- **high_chol:** Tình trạng cholesterol cao
- **chol_check:** Đã kiểm tra cholesterol trong vòng 5 năm gần đây hay chưa (gần với thời điểm bộ dữ liệu này được phát hành)
- **bmi:** Chỉ số bmi của cơ thể

- **smoker:** Đã hút được ít nhất 100 điếu thuốc hay chưa
- **stroke:** Đã từng đột quỵ hay chưa
- **heart_disorderor_attack:** Tình trạng mắc bệnh tim mạch vành/nhồi máu cơ tim
- **phys_activity:** Tình trạng tập thể dục thể thao trong 30 ngày vừa qua
- **fruits:** Có tiêu thụ trái cây mỗi ngày hay không
- **veggies:** Có tiêu thụ rau củ mỗi ngày hay không
- **hvy_alcohol_consump:** Tình trạng nghiện rượu (nam: 14 lần uống/tuần, nữ: 7 lần uống/tuần)
- **any_healthcare:** Có bất kỳ loại bảo hiểm sức khỏe nào không
- **no_docbc_cost:** Trong 1 năm vừa qua, có bao giờ không thể đi khám vì chi phí quá cao không?
- **gen_hlth:** Thang đo sức khỏe tổng quát từ 1 -> 5 (1: rất tốt, ..., 5: rất tệ)
- **men_hlth:** Trong 30 ngày vừa qua, có bao nhiêu ngày mà tinh thần bạn không ổn định?
- **phys_hlth:** Trong 30 ngày vừa qua, có bao nhiêu ngày mà sức khỏe thể chất của bạn không ổn định?
- **diff_walk:** Có gặp khó khăn trong việc đi đứng hay leo cầu thang không?
- **sex:** Giới tính
- **age:** 14 nhóm tuổi khác nhau (Từ 1 đến 8: 18 đến 54 tuổi, Từ 9 đến 12: 60 đến 79 tuổi, 13: >= 80 tuổi)
- **education:** Bậc học tập cao nhất (1: Không đi học, 2: Lớp 1 đến Lớp 8, 3: Lớp 9 đến Lớp 11, 4: Lớp 12 hay đã tốt nghiệp THPT, 5: Học ĐH/CĐ năm 1 đến năm 3, 6: Năm 4 hoặc năm cao hơn hoặc đã tốt nghiệp)

Nhận xét:

+ Bộ dữ liệu khá lớn với nhiều biến đặc trưng.
+ Có số lượng các nhãn lệch khá lớn, số lượng nhãn **`Non-diabete`** là 213703, nhiều gấp khoảng 5 lần tổng số lượng của hai biến **`Pre-diabete`** và **`Diabete`** cộng lại (hai nhãn sau lần lượt có số lượng là 4631 và 35346).
+ Ngoại trừ 3 biến là `bmi`, `men_hlth`, `phys_hlth` là biến định lượng thì 18 biến còn lại đều là biến định tính.

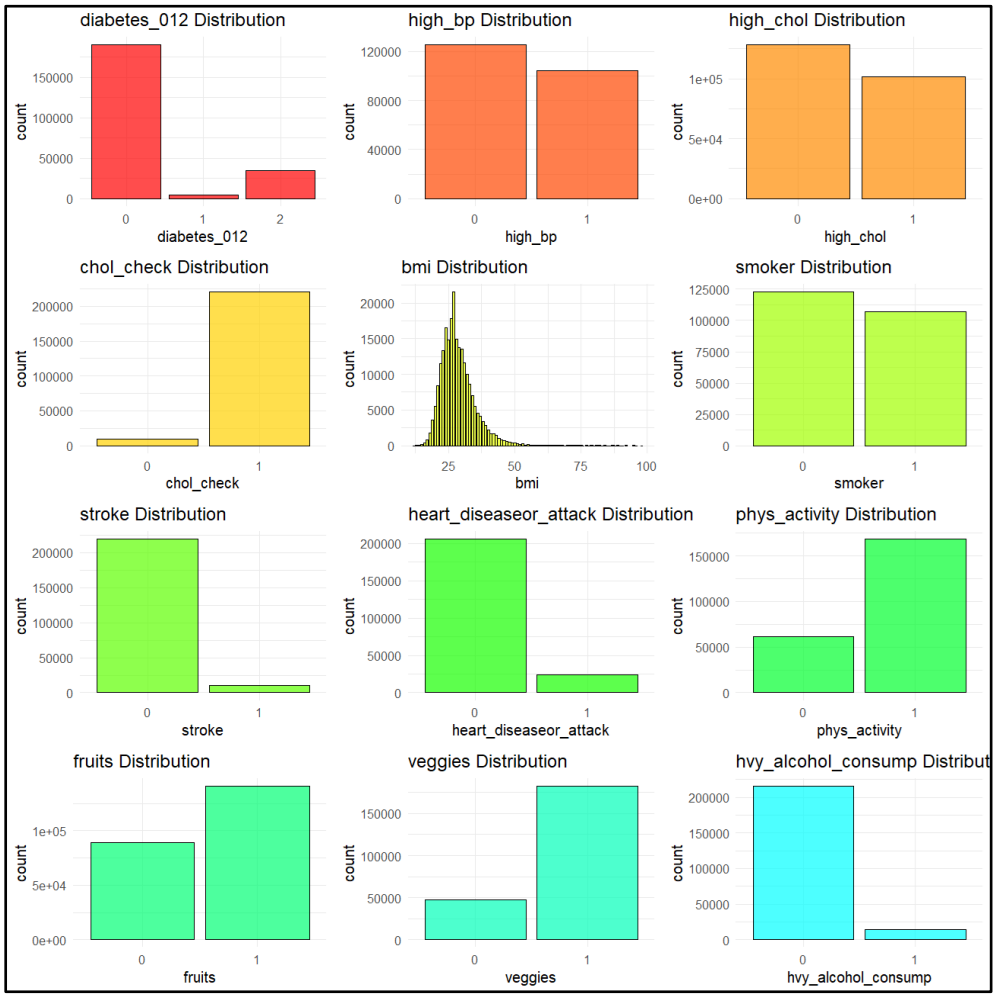
2. Bảng tổng hợp và mô tả dữ liệu (các biến định lượng)

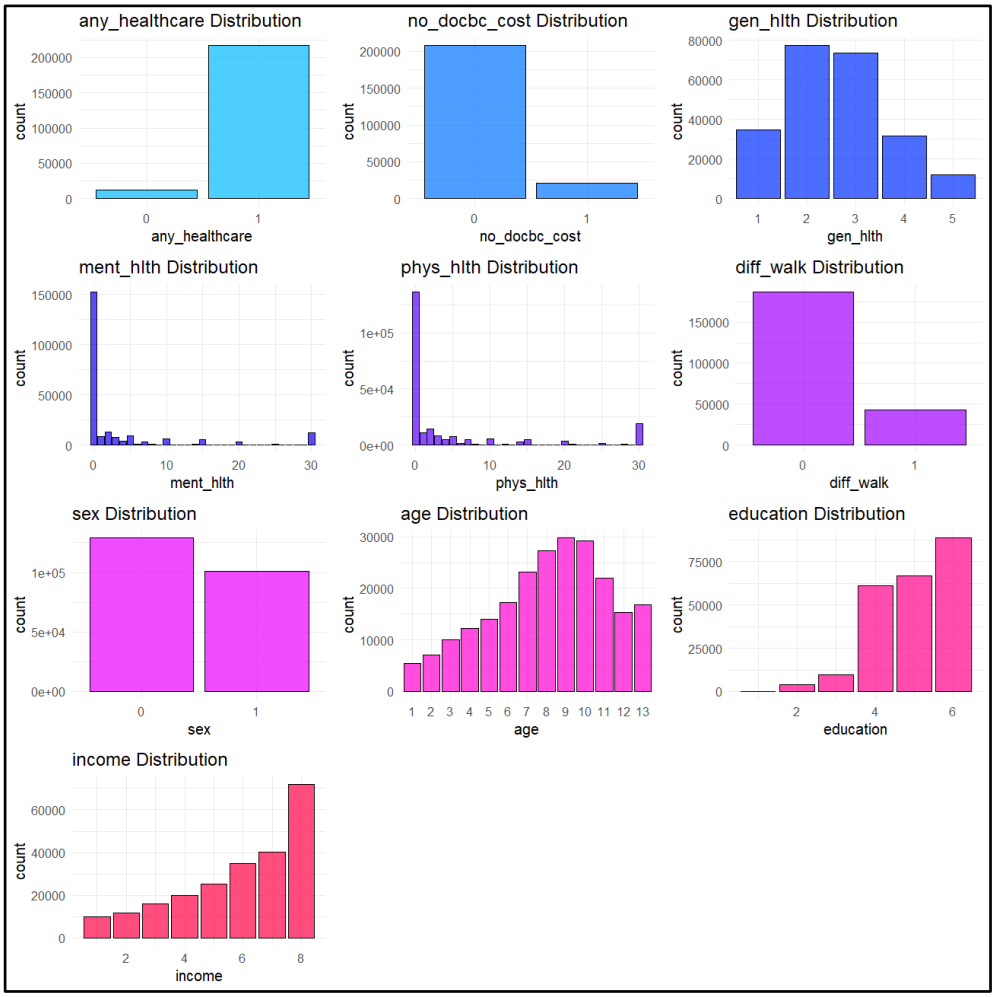
A tibble:3 × 7

bien	gtnn	gtln	tv	tb	dlc	iqr
<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
bmi	12	98	27	28.685670	6.786360	8
ment_hlth	0	30	0	3.505373	7.713725	2
phys_hlth	0	30	0	4.675178	9.0465681	4

3. Biểu đồ về dữ liệu

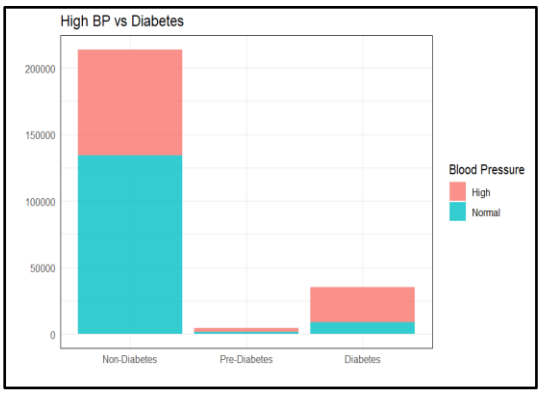
Phân phối mẫu của các biến trong dữ liệu:

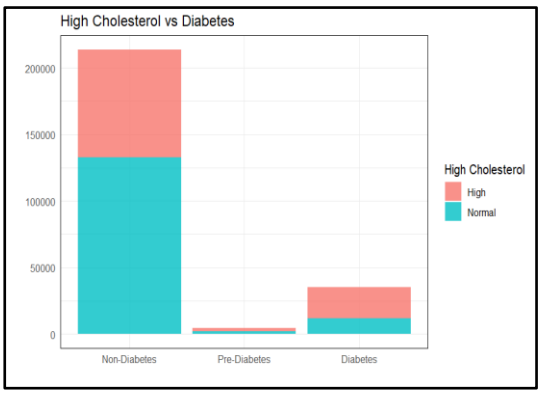




1.1. Biểu diễn cho biến định tính

a) Biến `high_bp` và `high_chol`

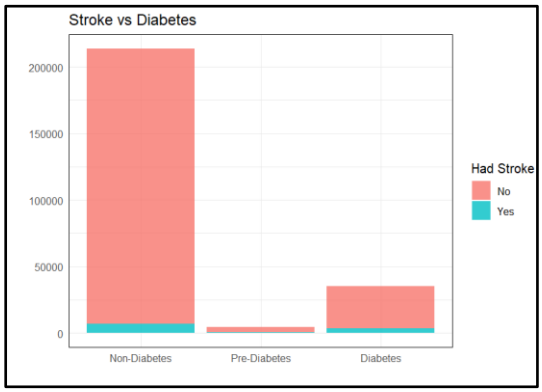


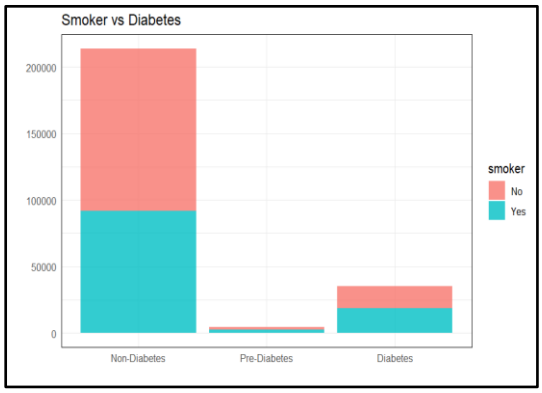


Nhận xét:

- Huyết áp cao có vẻ phổ biến ở cả ba nhóm, nhưng đặc biệt vượt trội ở người **tiểu đường** (chiếm 75,24%). Tương tự như huyết áp, tỷ lệ cholesterol cao - 66,95% - trong nhóm **Diabetes**.
- Huyết áp cao và cholesterol cao có mối liên hệ rõ rệt với nguy cơ và diễn biến bệnh tiểu đường.

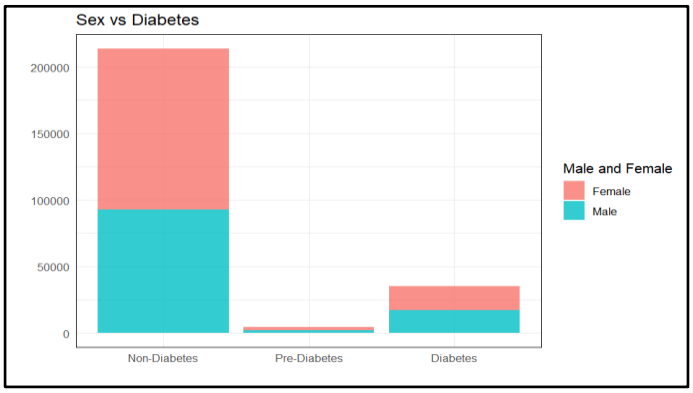
b) `Smoker`, `Stroke`, `Sex`





Nhận xét:

- Trong những người bị bệnh tiểu đường, số người không bị đột quỵ chiếm đa số, trong khi số người đã từng bị đột quỵ là rất ít. Do đó, ta có thể thấy đột quỵ có thể **không phải** là yếu tố nhận biết bệnh tiểu đường.
- Trong các nhóm, tỷ lệ người hút thuốc gần ngang bằng với người không hút thuốc. Điều này cho thấy hút thuốc **có khả năng** làm tăng rủi ro ảnh hưởng đến sức khỏe người bị tiểu đường.

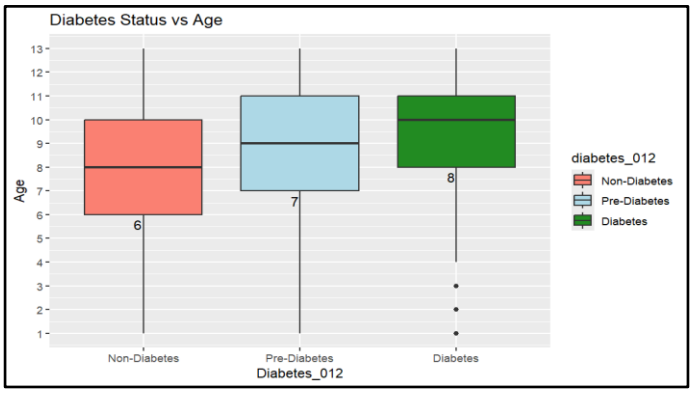


Nhận xét:

- Số lượng nữ mắc bệnh tiểu đường lớn hơn nam nhưng không đáng kể. Cả hai giới đều có nguy cơ mắc bệnh tiểu đường gần tương đương. Từ đó có thể thấy, giới tính **không phải** là yếu tố quyết định trực tiếp đến nguy cơ mắc tiểu đường.

c) Biến `Age` và `Heart Diseaseor Attack` đối với `Diabetes_012`

- Biến **`Age`**



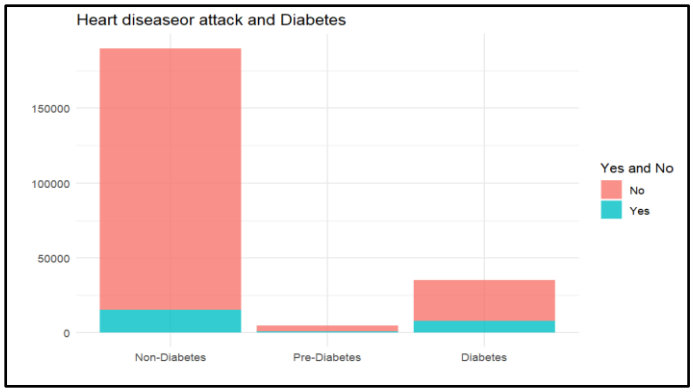
Nhận xét:

- Bệnh tiểu đường phổ biến hơn ở những nhóm tuổi cao, đặc biệt là từ 55 tuổi trở lên. Đây là giai đoạn cơ thể bắt đầu giảm khả năng chuyển hóa glucose do sự suy giảm chức năng insulin và sự tích tụ các yếu tố nguy cơ khác như béo phì, huyết áp cao, và lối sống ít vận động.
- Những người dưới 35 tuổi ít có nguy cơ mắc bệnh tiểu đường hơn, điều này có thể do sự trao đổi chất trong cơ thể còn tốt, hệ miễn dịch còn khỏe, và ít chịu ảnh hưởng từ các yếu tố nguy cơ mãn tính (như huyết áp, cholesterol cao).

Kết luận:

Chủ động kiểm soát sức khỏe từ sớm và triển khai các biện pháp phòng ngừa, chăm sóc cho các nhóm tuổi khác nhau sẽ mang lại hiệu quả cao trong việc kiểm soát và giảm thiểu tác động của bệnh tiểu đường.

- Biến **`Heart_diseaseor_attack`**



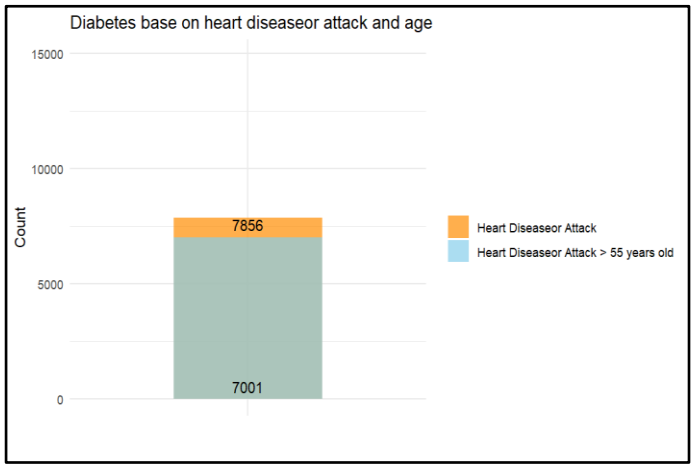
Nhận xét:

Tiểu đường và bệnh tim mạch có mối liên hệ chặt chẽ. Người mắc bệnh tiểu đường (cả tiền tiểu đường và tiểu đường) có nguy cơ cao hơn bị bệnh mạch vành, suy tim, hoặc nhồi máu cơ tim do lượng đường trong máu cao kéo dài làm tổn thương các mạch máu và hệ tuần hoàn.

Kết luận:

Cần có các biện pháp cải thiện, thay đổi lối sống, kiểm soát y tế và hỗ trợ tâm lý sẽ là chìa khóa giúp giảm tác động của bệnh.

Đặt vấn đề: Liệu những người cao tuổi và có bệnh nền (bệnh tim) thì có nguy cơ mắc bệnh tiểu đường cao hơn hay không?



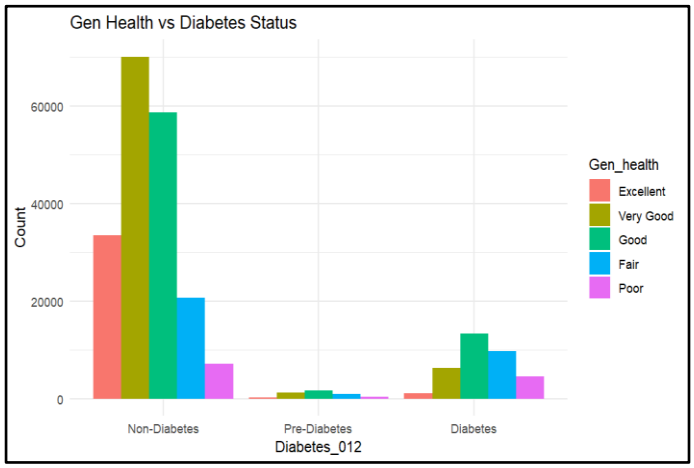
Nhận xét:

- Nhóm người vừa bị tiểu đường vừa bị bệnh tim phần lớn là người ở độ tuổi từ 55 trở lên. Có thể thấy được rằng người lớn tuổi có bệnh nền thì nguy cơ bị tiểu đường rất cao. Khi tuổi tác tăng cao, cơ thể trải qua nhiều thay đổi sinh học như giảm hiệu quả chuyển hóa glucose, tăng đề kháng insulin, và suy giảm chức năng các cơ quan quan trọng.

Kết luận:

- Thay đổi lối sống lành mạnh như chế độ ăn uống cân đối, tập thể dục đều đặn và kiểm tra sức khỏe định kỳ, có thể giảm nguy cơ tiến triển tiểu đường cũng như các biến chứng liên quan.

d) `Gen_hlth`



Nhận xét:

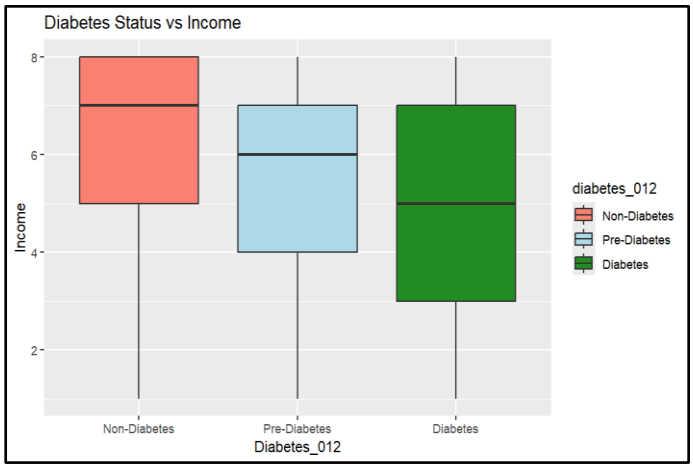
- Những người có sức khỏe tổng quan tốt (Very Good/Good) ít mắc bệnh tiểu đường hơn.
- Nhóm người tự đánh giá sức khỏe của mình ở mức "Fair" hoặc "Poor" có nguy cơ cao mắc bệnh tiểu đường. Đây là dấu hiệu cho thấy cần thực hiện các biện pháp can thiệp sớm như kiểm tra sức khỏe thường xuyên, thay đổi lối sống, và cải thiện chế độ dinh dưỡng.

Kết luận:

- Biểu đồ này nhấn mạnh rằng sức khỏe tổng quan có mối liên hệ rất chặt chẽ với nguy cơ mắc bệnh tiểu đường, giúp đưa ra những đánh giá tình trạng bệnh và đưa ra các phương pháp thích hợp.

e) `Income` và `Education`

- **`Income`**



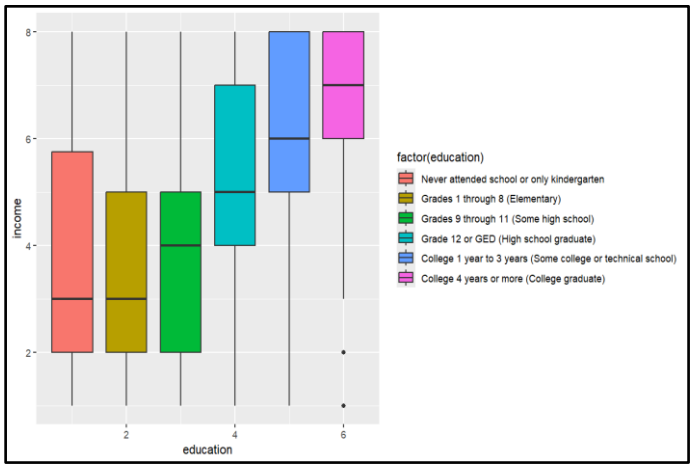
Nhận xét:

- Nhóm **Non-Diabetes** (không mắc tiểu đường) có mức thu nhập trung bình cao nhất. Điều này cho thấy những người có thu nhập cao hơn thường có khả năng tiếp cận các dịch vụ y tế, chế độ ăn uống và lối sống lành mạnh hơn, giúp giảm nguy cơ mắc tiểu đường.

Kết luận:

- Thu nhập có mối liên hệ nghịch với nguy cơ mắc tiểu đường: thu nhập càng thấp thì nguy cơ mắc tiểu đường và từng bị tiểu đường càng cao.
- Từ điều này, các chương trình chăm sóc sức khỏe và giáo dục y tế cần tập trung vào các nhóm thu nhập thấp để giảm nguy cơ mắc tiểu đường trong cộng đồng.

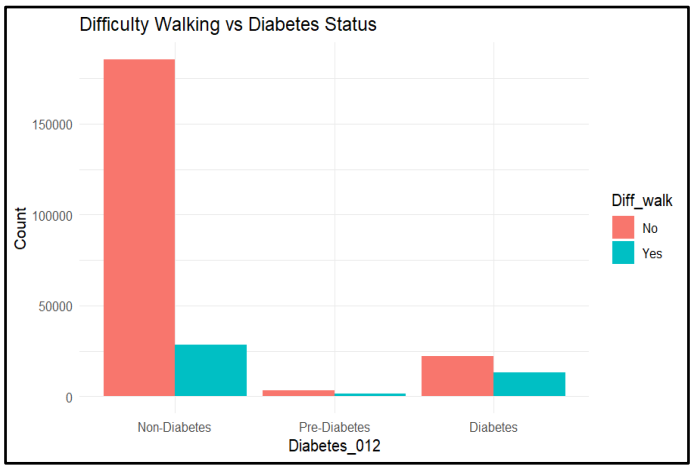
Hệ số tương quan của Income và Education là 0,419 nên ta vẽ biểu dồ xét mối quan hệ giữa chúng:



Nhận xét:

- Nhìn chung, trình độ học vấn càng cao thì thu nhập trung bình càng tăng. Điều này phản ánh rằng giáo dục đóng vai trò quan trọng trong việc nâng cao mức thu nhập.

f) `Diff_walk`



Nhận xét:

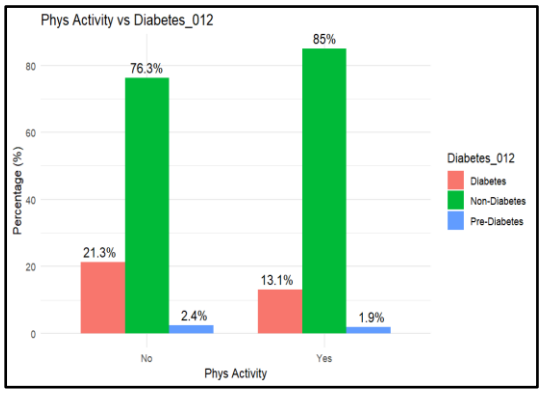
- Trong nhóm Non-Diabetes (không tiểu đường), phần lớn mọi người không gặp khó khăn khi đi bộ. Đây là nhóm có tỷ lệ "No" cao nhất so với các nhóm khác.

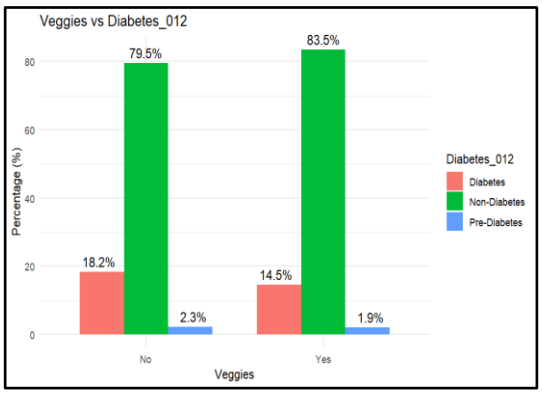
- Tỷ lệ "No" trong từng nhóm giảm dần khi chuyển từ nhóm Pre-Diabetes (tiền tiểu đường) và nhóm Diabetes (bị tiểu đường).

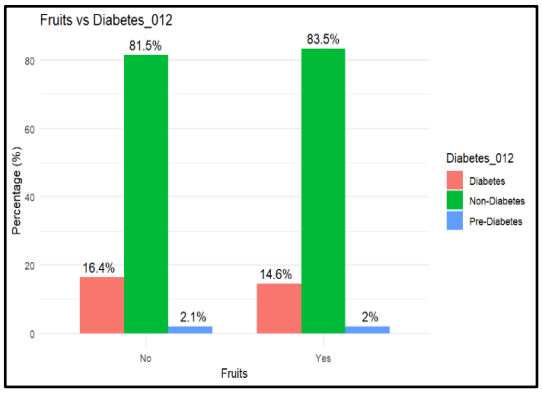
Kết luận:

Gặp khó khăn trong việc đi lại hoặc leo cầu thang có thể là một trong những triệu chứng liên quan đến bệnh tiểu đường, đặc biệt khi bệnh tiến triển. Việc tăng cường hoạt động thể chất và duy trì cân nặng lành mạnh có thể giúp giảm nguy cơ chuyển sang nhóm **Diabetes** và giảm thiểu nguy cơ gặp khó khăn khi đi bộ.

g) Các nhóm biến liên quan đến lối sống bao gồm: 'Phys_activity', 'Veggies', 'Fruits'.







Nhận xét:

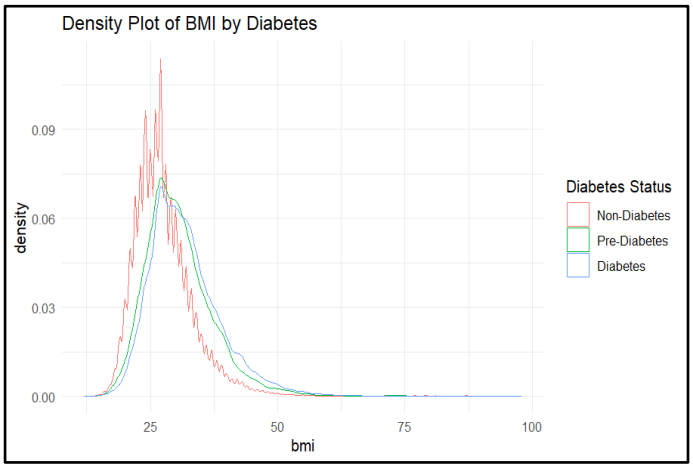
Người có lối sống lành mạnh, tham gia hoạt động thể chất, ăn nhiều rau xanh, trái cây có tỉ lệ tiền tiểu đường và tiểu đường thấp hơn rõ rệt.

Kết luận:

Điều này cho thấy việc tham gia hoạt động thể chất, tăng cường vận động, tiêu thụ lượng đường và mỡ thừa trong cơ thể kết hợp với chế độ ăn nhiều rau củ, trái cây chứa nhiều vitamin, giàu chất xơ là một trong những phương pháp lành mạnh có ảnh hưởng tích cực đến sức khỏe, giúp phòng ngừa, giảm nguy cơ mắc bệnh tiểu đường.

1.2. Biểu diễn cho biến định lượng

a) 'BMI'



Nhận xét:

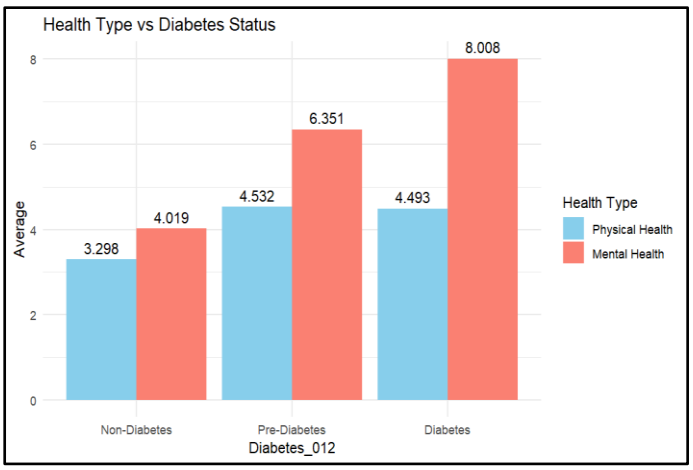
- Nhóm Non-Diabetes có đỉnh mật độ BMI thấp hơn so với nhóm Pre-Diabetes và Diabetes, tập trung nhiều trong khoảng BMI từ 20 đến 30.
- Nhóm Pre-Diabetes và Diabetes có xu hướng chuyển dịch sang BMI cao hơn (khoảng từ 30 đến 40).
- Người mắc bệnh béo phì có nguy cơ mắc tiểu đường cao gấp 8 lần so với những người có trọng lượng cơ thể bình thường hoặc thậm chí gầy.

Kết luận:

BMI cao không chỉ ảnh hưởng đến sức khỏe tổng thể mà còn là một trong những yếu tố rủi ro chính dẫn đến các bệnh mãn tính như tiểu đường. Vì vậy cần có chế độ ăn uống cân bằng, hoạt động thể chất

thường xuyên và giảm tiêu thụ thực phẩm chế biến sẵn nhằm kiểm soát chỉ số BMI ở mức lành mạnh và giảm nguy cơ mắc bệnh.

b) 'Ment Hlth' và 'Phys Hlth'



Nhận xét:

- Nhóm *Non-Diabetes* có trung bình số ngày sức khỏe thể chất không tốt (khoảng 3.3 ngày) thấp hơn so với nhóm *Pre-Diabetes* và *Diabetes* (khoảng 4.5 ngày).
- Nhóm *Diabetes* cũng có số ngày sức khỏe tinh thần không tốt cao nhất (khoảng 8 ngày), cao gần gấp đôi so với nhóm *Non-Diabetes* (4 ngày).
- Sức khỏe tinh thần dường như bị ảnh hưởng nặng nề hơn sức khỏe thể chất ở nhóm *Diabetes*, với sự chênh lệch rõ ràng giữa số ngày trung bình của hai loại sức khỏe.

Kết luận:

Những người mắc bệnh tiểu đường có xu hướng gặp nhiều vấn đề sức khỏe tinh thần hơn. Điều này có thể liên quan đến căng thẳng trong việc quản lý bệnh, lo lắng về biến chứng, hoặc gánh nặng tài chính. Từ đó khuyến khích các chương trình hỗ trợ tâm lý và giáo dục sức khỏe có thể giúp cải thiện chất lượng cuộc sống. Đồng thời các yếu tố như hoạt động thể chất, dinh dưỡng hợp lý và giấc ngủ đủ có thể đóng vai trò quan trọng trong việc cải thiện cả sức khỏe thể chất và tinh thần.

III. ĐỀ XUẤT PHÂN TÍCH VÀ XỬ LÝ SỐ LIỆU

1.Tiền xử lý dữ liệu:

Làm sạch dữ liệu:

- Đổi tên các biến cần thiết
- Loại bỏ dữ liệu trùng lặp
- Phân loại các biến theo loại (biến định lượng, biến định tính) để hỗ trợ trong các tác vụ đạt được các mục tiêu phân tích.

2. Trực quan hóa dữ liệu

- Sử dụng các kỹ thuật vẽ biểu đồ như: histogram, boxplot, density,... để hiểu rõ hơn về phân phối, đặc trưng dữ liệu, từ đó quyết định các phân tích tiếp theo.

3. Kiểm định bằng AB testing

Sử dụng các phương pháp AB testing đã được học để tiến hành kiểm định:

- Resampling method: để kiểm tra sự độc lập của các biến định tính.
- Permutation ANOVA: dùng để kiểm tra sự khác biệt giữa các nhóm trong trường hợp các giả định chuẩn của ANOVA truyền thống (như phân phối chuẩn và phương sai đồng nhất) không được thỏa mãn.

4. Xử lý mất cân bằng dữ liệu

Thử nghiệm cân bằng dữ liệu trên các phương pháp khác nhau:

+ Phương pháp 1: Under Sampling
+ Phương pháp 2: Over Sampling
+ Phương pháp 3: SMOTE
+ Phương pháp 4: Class Weight
+ Phương pháp 5: Kết hợp Under Sampling và SMOTE
+ Phương pháp 6: Kết hợp Under Sampling và Over Sampling

Đánh giá các phương pháp cân bằng dữ liệu dựa vào hiệu suất của mô hình. Bộ dữ liệu cho kết quả hiệu suất mô hình dự đoán được tốt nhất trên tất cả các lớp sẽ được lựa chọn để triển khai trên nhiều mô hình khác nhau.

5. Xây dựng mô hình

Có sự song song trong việc tìm kiếm mô hình giải thích tốt nhất đồng thời quyết định xem sẽ thực hiện trên tập dữ liệu nào để có thể cho ra một kết quả khả quan nhất.

Xây dựng 2 loại mô hình:

- **Phân loại đa nhóm:** đưa ra kết quả dự đoán về tình trạng bệnh tiểu đường của một người (*'Non-diabete', 'Pre-diabete', 'Diabete'*) dựa vào các chỉ số đầu vào được cung cấp, với các mô hình được sử dụng là: Multinominal Logistic Regression, Random Forest, Naive Bayes.
    + Dự đoán dựa trên top 10 biến đầu vào tương quan cao nhất với biến mục tiêu
    + Dự đoán chỉ dựa trên các chỉ số liên quan đến sức khỏe (health indicator). Các biến chỉ số sức khỏe bao gồm: gen_hlth, high_bp, bmi, diff_walk, high_chol, heart_diseaseor_attack, phys_hlth.

Dựa vào kết quả của mô hình phân loại đa nhóm, cùng với sự "lép vế" về số lượng của biến Pre-diabete so với hai biến còn lại, chúng em cân nhắc xây dựng mô hình đơn giản hơn để nhận biết một người có mắc bệnh tiểu đường hay không (không phân biệt "Pre-diabete" và "Diabete") và hướng tới việc xác định được mô hình phân loại nhị phân nào giải thích tốt nhất.

- **Phân loại nhị phân:** dự đoán xem một người có mắc bệnh tiểu đường dựa và các chỉ số được cung cấp lần lượt với 3 mô hình sau: Logistic Regression, Random Forest, Naive Bayes.

IV. PHƯƠNG PHÁP VÀ CHIẾN LƯỢC THỰC HIỆN

1.Khám phá dữ liệu, xem xét mối tương quan giữa các biến với nhau và đối với biến mục tiêu

1.1. Phương pháp

Dùng các biểu đồ phù hợp để trực quan các biến và mối quan hệ giữa các biến với nhau để làm rõ nghi ngờ về mối tương quan giữa các biến, và sự khác biệt trong sự đóng góp giữa các biến đặc trưng lên các biến phân loại.

1.2. Cách thực hiện

- Trực quan phân phối mẫu của các biến trong bộ dữ liệu bằng biểu đồ histogram
- Kiểm tra mối liên hệ giữa các biến với nhau thông qua các biểu đồ trực quan phù hợp.
- Kiểm tra liệu có sự khác biệt đáng kể của 2 biến phân loại *'Pre-diabete'*, *'Diabete'* không bằng cách loại bỏ các dòng có nhãn *'Non-diabete'* trong bộ dữ liệu và trực quan một một số tương quan của các biến đặc trưng với 2 nhãn để kiểm tra.

2. Dùng A/B Testing kiểm chứng các giả thiết đặt ra

2.1. Phương pháp

- Sau khi trực quan hóa dữ liệu, nhận thấy các biến có ảnh hưởng đến biến mục tiêu. Chúng em tiến hành A/B testing lần lượt cho biến định tính và định lượng. Vì dữ liệu của dataset này rất nhiều nên chúng em đã random 10% dữ liệu gốc sau đó đi kiểm định.
- Biến định tính:
    - Kiểm tra sự độc lập bằng cách sử dụng resampling method để khắc phục hạn chế khi Chi-square có ít nhất 1 tần số kỳ vọng $\mathbf{E_{ij}<5}$.
- Biến định lượng:
    - Kiểm tra xem có sự khác nhau về trung bình các biến định lượng giữa các nhóm không.
        - Sử dụng phương pháp resampling ANOVA để khắc phục các hạn chế của ANOVA thông thường.

2.2. Các bước thực hiện

Bước 1: Đặt ra giả thiết:

- Về biến định tính:
  - H_0: biến định tính độc lập với biến mục tiêu
  - H_1: biến định tính không độc lập với biến mục tiêu
- Về biến định lượng:
  - H_0: không có sự khác biệt về trung bình của các biến định lượng giữa ba nhóm biến phân loại.
  - H_1: có sự khác biệt về trung bình các biến định lượng giữa nhóm biến phân loại.

Bước 2: Lấy random 10% dữ liệu từ dataset.

Bước 3: Viết hàm và sử dụng các hàm kiểm định có sẵn trong máy.

Bước 4: Tiến hành kiểm định và nhận xét các kết quả đạt được.

3.Xây dựng mô hình dự đoán nguy cơ mắc bệnh tiểu đường.

3.1. Phương pháp:

Chúng em sử dụng các thư viện trực quan mạnh mẽ như *ggplot2* và *corrplot* để khám phá dữ liệu. Quá trình phân tích này giúp chúng em hiểu rõ hơn về mối liên hệ giữa các đặc trưng trong bộ dữ liệu. Sau bước phân tích khám phá, chúng em bắt đầu quá trình xử lý đặc trưng (feature selection) để tìm ra các biến có khả năng dự báo tốt nhất thông qua *correlation*, tạo cơ sở vững chắc cho bước xây dựng mô hình.

Để chuẩn bị cho bước xây dựng mô hình, chúng em tìm ra cách xử lý tối ưu cho sự mất cân bằng của bộ data được cung cấp bằng cách áp dụng các phương pháp cân bằng dữ liệu (*Under Sampling, Over Sampling, Data Generation, Class Weight*) và áp dụng lên một mô hình *Multinominal Logistic Regression* để tìm ra phương pháp tốt nhất cho bộ dữ liệu. Sau đó, tiếp tục thử nghiệm bộ dữ liệu thu được trên các mô hình khác như Random Forest và Naive Bayes. Sử dụng các chỉ số đánh giá, đặc biệt chú ý đến chỉ số Macro-F1, Kappa của các mô hình thu được sau khi huấn luyện để lựa chọn ra mô hình giải thích tốt nhất cho bộ dữ liệu.

3.2. Các bước thực hiện:

Bước 1: Lựa chọn các biến đặc trưng có mối tương quan cao nhất với biến phân loại để đưa vào mô hình dự đoán.

Bước 2: Chuẩn bị dữ liệu, chia dữ liệu huấn luyện - kiểm tra trước khi bước vào huấn luyện mô hình.

Bước 3: Huấn luyện mô hình trên các bộ dữ liệu khác nhau và các mô hình đề ra.

Bước 4: Đánh giá mô hình bằng các chỉ số đánh giá mô hình phân loại và trực quan kết quả (nếu có).

Bước 5: Lựa chọn bộ dữ liệu phù hợp và mô hình tốt nhất trên bộ dữ liệu được chọn.

V. KẾT QUẢ PHÂN TÍCH

1. AB-TESTING

Các biến định tính: tất cả đều có p_value xấp xỉ bằng 0.

Các biến định lượng: ment_htlh, bmi, phys_htlh có p_value xấp xỉ bằng 0.

Nhận xét:

Sau khi thực hiện A/B testing

- Tất cả các biến định tính đều có p-value xấp xỉ về 0, có thể thấy rằng tất cả các biến độc lập đều có ý nghĩa thống kê trong mối quan hệ với biến mục tiêu (bệnh tiểu đường). Kết quả này cho thấy dữ liệu chứa các biến đều đóng vai trò nhất định trong việc dự đoán bệnh tiểu đường.
- Tất cả các biến định lượng đều có p-value xấp xỉ về 0 cho thấy có sự sai khác về trung bình ment_htlh, bmi, phys_htlh giữa nhóm của biến mục tiêu.

Đưa ra các nhận định trong cuộc sống:

- **Chỉ số cơ thể:** Người có BMI cao, huyết áp cao và cholesterol cao có nguy cơ mắc tiểu đường cao hơn. Quản lý cân nặng và kiểm soát huyết áp/cholesterol là yếu tố quan trọng để phòng ngừa bệnh.
  + Để giảm huyết áp và cholestorol cao cần có lối sống lành mạnh (tập thể dục, ăn ít muối, giảm mỡ động vật) có thể hỗ trợ giảm nguy cơ mắc tiểu đường.
  + Kiểm tra định kỳ các chỉ số huyết áp và cholesterol cho người thuộc nhóm nguy cơ cao về nguy cơ chuyển sang tiền tiểu đường hoặc tiểu đường.

- **Thói quen:** Thực hiện thói quen lành mạnh như tập thể dục thường xuyên và bỏ thuốc lá giúp giảm nguy cơ và nâng cao sức khỏe bản thân.
- **Lối sống:** Người tự đánh giá sức khỏe kém, mắc bệnh tim mạch hoặc từng bị đột quỵ có nguy cơ cao mắc tiểu đường. Quản lý sức khỏe tổng thể là yếu tố quyết định trong việc phòng ngừa. Cần giữ cho tinh thần thoải mái, có lối sống lành mạnh cả về thể chất và tinh thần.
- Nguy cơ mắc tiểu đường tăng theo độ tuổi, đặc biệt là trên 55 tuổi, và người có thu nhập/thực hành giáo dục thấp dễ mắc bệnh hơn.
    + Do đó cần tăng cường kiểm tra sức khỏe định kỳ, cung cấp thông tin về chế độ ăn uống lành mạnh và khuyến khích hoạt động thể chất dễ dàng.
    + Đồng thời, hỗ trợ tiếp cận dịch vụ y tế giá rẻ, thực phẩm lành mạnh và tổ chức các chương trình giáo dục sức khỏe cộng đồng.

Chiến lược cho mô hình:

Tất cả các biến độc lập đều có ý nghĩa thống kê,tuy nhiên cần đảm bảo tận dụng tối đa thông tin từ dữ liệu và tối ưu hóa mô hình phân loại, do đó cần thực hiện một số biện pháp như:

+ Cân bằng dữ liệu (Data Balancing), tránh hiện tượng mô hình thiên lệch (bias) về phía nhóm chiếm đa số (dữ liệu nhóm "Non-Diabetes" áp đảo, mô hình có thể bỏ qua nhóm “ Pre-Diabetes” và “Diabetes”. Đồng thời cải thiện khả năng nhận diện nhóm thiểu số (nhóm mắc bệnh) mà không làm giảm hiệu quả chung của mô hình.
+ Sử dụng các phương pháp đo lường tầm quan trọng của đặc trưng (feature importance), kiểm tra giá trị của các trọng số (coefficients) từ đó đưa ra đánh giá mức độ ảnh hưởng của từng biến.

```
## Correlation:
                          diabetes_type
high_bp                     0.27159642
high_chol                   0.20908491
chol_check                  0.06754648
bmi                        0.22437947
smoker                      0.06291410
stroke                     0.10717867
heart_diseaseor_attack    0.18027169
phys_activity              -0.12194717
```

fruits -0.04219163

veggies -0.05897160

hvy_alcohol_consump -0.05788191

any_healthcare 0.01541038

no_docbc_cost 0.03543569

gen_hlth 0.30258662

ment_hlth 0.07350677

phys_hlth 0.17628674

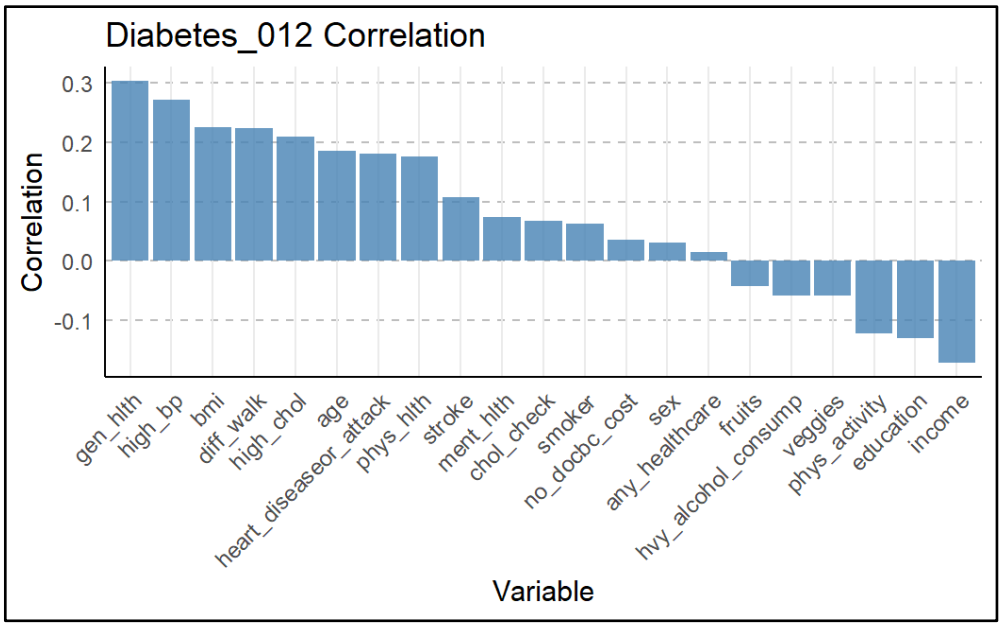
diff_walk 0.22423912

sex 0.03104016

age 0.18502579

education -0.13051692

income -0.17148304



Top 10 biến:

"gen_hlth" "high_bp" "bmi"

"diff_walk" "high_chol" "age"

"heart_diseaseor_attack" "phys_hlth" "income"

"education"

+ Thử nghiệm dữ liệu trên nhiều mô hình khác nhau: Multinominal Logistic Regression, Random Forest, Naive Bayes để tìm ra mô hình phù hợp và hiệu quả nhất.

2. XỬ LÝ MẤT CÂN BẰNG DỮ LIỆU

STT	Phương pháp	Accuracy	Kappa	Macro-F1
1	Under Sampling	0.5032397	0.2548596	0.1245795
2	Over Sampling	0.5143988	0.2715983	0.1327445
3	SMOTE	0.5179986	0.2769978	0.1367898
4	Class weight	0.5154788	0.2732181	0.1337437
5	**Under Sampling w SMOTE**	**0.5147588**	**0.2721382**	**0.1343537**
6	Under Sampling + Over Sampling	0.5158387	0.2737581	0.1339996

Nhận xét:

So với dữ liệu gốc, các phương pháp cân bằng dữ liệu đã giải quyết vấn đề overfitting, có thể dự đoán cả ba lớp. Chỉ số hiệu suất mô hình không quá khác biệt giữa các phương pháp cân bằng dữ liệu, do đó có thể chọn một phương pháp để tiến hành huấn luyện mô hình.

Tuy nhiên, nhóm đã chọn phương pháp thứ năm, tương ứng với bộ dữ liệu *"Kết hợp Under Sampling về trung bình cỡ mẫu ba lớp và SMOTE tạo dữ liệu mới"*. Nguyên nhân là do bộ dữ liệu có sự mất cân bằng nghiêm trọng, với cỡ mẫu của lớp ít quan sát nhất chỉ chiếm hơn 2% cỡ mẫu lớp nhiều quan sát nhất, do đó việc chỉ dùng các phương pháp riêng biệt sẽ làm dữ liệu bị mất tính đa dạng hoặc bị lặp nghiêm trọng, và cỡ dữ liệu sẽ quá lớn hoặc quá bé so với dữ liệu gốc.

3. MÔ HÌNH BÀI TOÁN PHÂN LOẠI

Bài toán phân loại đa lớp với bộ dữ liệu tương quan cao:

Mô hình	**Accuracy**	**Kappa**	**Macro-F1**
Multinominal Logistic Regression	0.5147588	0.2721382	0.1343537
Random Forest	**0.7739381**	**0.6609071**	**0.466901**
NaiveBayes	0.5007199	0.2510799	0.1217087

Nhận xét:

- Hai mô hình Multinominal Logistic Regression và Naive Bayes có kết quả dự đoán khá kém với các thông số đánh giá khá tương tự nhau, ở mức khá thấp, đặc biệt là Macro-F1 (chỉ dưới 0.15).
- Tuy nhiên, Mô hình *Random Forest* có các chỉ số đánh giá khả quan hơn:
    + Accuracy (0.7739): với gần 77,4% tỷ lệ dự đoán đúng, ta thấy mô hình đạt được độ chính xác khá cao.
    + Kappa (0.6609): Kết quả chỉ số cho thấy dự đoán của mô hình có phụ thuộc vào phân loại thực tế nhưng do có sự chênh lệch giữa các nhãn nên con số này không thể cải thiện lên quá cao.
    + Macro-F1 (0.4669): Giá trị này vẫn còn khá thấp cho thấy mô hình vẫn chưa có sự cân bằng tốt giữa các nhãn.

Kết luận: Mô hình Random Forest là phù hợp nhất để đưa ra dự đoán chính xác về tình trạng bệnh tiểu đường dựa vào các chỉ số đầu vào của một người, từ đó có thể phát hiện bệnh sớm và giúp người mắc bệnh có phương án điều trị hợp lý với tình trạng bệnh hiện tại.

Bài toán phân loại đa lớp với bộ dữ liệu liên quan đến chỉ số sức khoẻ:

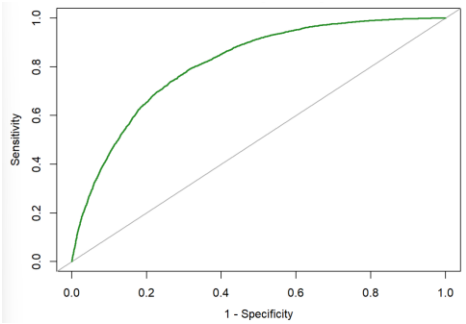
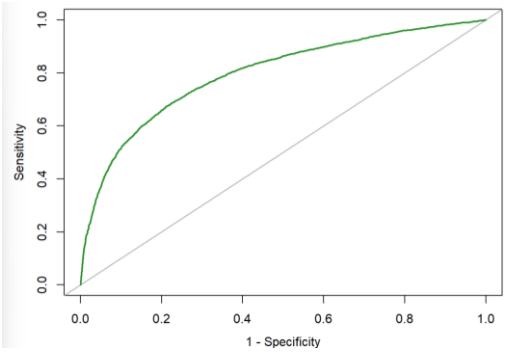
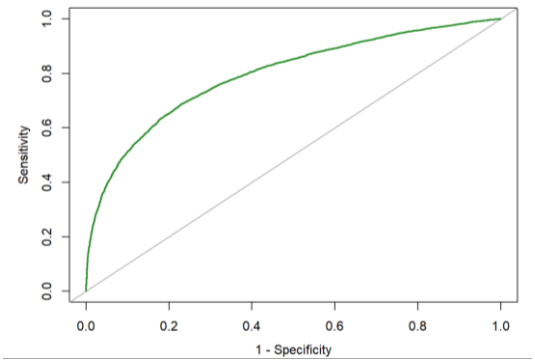
Mô hình	**Accuracy**	**Kappa**	**Macro-F1**
Multinominal Logistic Regression	0.5057595	0.2586393	0.1252183
Random Forest	**0.7760979**	**0.6641469**	**0.470444**

Nhận xét:

- Mô hình Random Forest có chỉ số Macro-F1 cao gần gấp 4 lần chỉ số Macro-F1 của mô hình Multinominal Logistic Regression. Điều này chứng tỏ RF xử lý và đạt được sự cân bằng các nhãn tốt hơn.
- Các chỉ số Accuracy và Kappa của Random Forest cũng cao hơn của Multinominal Logistic Regression, lần lượt là 0.7760979 và 0.6641469, cho thấy độ chính xác vượt trội và hiệu suất cao.

Tổng quát: Với bài toán phân loại đa lớp, tập dữ liệu tốt nhất là tập chỉ gồm các biến liên quan tới chỉ số sức khỏe, được huấn luyện trên mô hình Random Forest. Như vậy, một người có thể được chẩn đoán sớm có đang trong giai đoạn tiền tiểu đường hoặc mắc phải căn bệnh này hay chưa dựa vào các chỉ số sức khỏe, tình trạng cơ thể của họ. Điều này có thể giúp cho họ cân nhắc điều chỉnh lối sống phù hợp cũng như là một công cụ hỗ trợ đội ngũ y tế trong việc điều trị.

1. Bài toán phân loại nhị phân:

Model	AUC	ROC Curve
Logistic Regression	0.8113	
Random Forest	0.7969	
Naive Bayes	0.7947	

Nhận xét:

Mô hình Logistic Regression có chỉ số AUC = 0.81, là con số cho thấy khả năng phân biệt giữa các lớp của mô hình khá tốt, đáng tin cậy.

C. TỔNG KẾT

Thông qua bộ dữ liệu BRFSS 2015 (Hệ thống giám sát các yếu tố nguy cơ và hành vi), bằng cách tích hợp các công cụ thống kê và các mô hình phân loại, phân tích của nhóm khám phá ra những hiểu biết về nguy cơ mắc bệnh tiểu đường. Chúng ta có thể thấy được rằng các yếu tố xã hội, sức khỏe và thói quen sinh hoạt tác động đáng kể đến nguy cơ mắc bệnh tiểu đường.

Qua bài báo cáo này chúng ta thấy được rằng việc kiểm soát tiểu đường không chỉ dựa vào các biện pháp y học mà còn đòi hỏi có sự thay đổi toàn diện về lối sống, cải thiện điều các điều kiện xã hội và các chính sách về y tế phù hợp. Điều này có thể giúp giảm thiểu tác động của bệnh tiểu đường đối với cộng đồng.