

ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN
KHOA TOÁN - TIN HỌC



BÁO CÁO CUỐI KỲ

Đề tài: Project 2 - CDC Diabetes Health Indicators

MTH10513 - XỬ LÝ SỐ LIỆU THỐNG KÊ

Sinh viên thực hiện:

22110014 - Đậu Quang Anh

22110023 - Lâm Gia Bảo

22110035 - Trần Quốc Danh

22110033 - Lê Thị Hồng Đăng

22110008 - Trần Duy An

Giảng viên bộ môn: TS. Tô Đức Khánh

Ngày 23 tháng 1 năm 2025

I. CÁC YÊU CẦU

1. Bản đề xuất phân tích và xử lý số liệu dựa trên các phương pháp đã được học trong học phần.
2. Các mục tiêu phân tích cần đạt được.
3. Các phương pháp và chiến lược (các bước) phân tích cho mỗi mục tiêu đã đề ra.
4. Mô tả và biểu diễn tổng hợp dữ liệu (bảng tổng hợp, biểu đồ).
5. Phân tích để đạt được các mục tiêu đã đề ra, kết quả được biểu diễn dưới dạng bảng tổng hợp, biểu đồ.
6. Viết các nhận xét và kết luận về các kết quả đã thu được sau quá trình phân tích

II. TỔNG QUAN ĐỀ TÀI

Bệnh tiểu đường hiện nay đang trở thành một trong những căn bệnh mãn tính phổ biến và nghiêm trọng nhất trên toàn cầu. Căn bệnh này xảy ra khi cơ thể không thể điều chỉnh hiệu quả mức độ glucose trong máu, dẫn đến những biến chứng nghiêm trọng như bệnh tim mạch, mất thị lực, cắt cụt chi dưới, và bệnh thận. Mặc dù hiện tại chưa có phương pháp chữa khỏi bệnh tiểu đường, nhưng các chiến lược can thiệp như giảm cân, duy trì chế độ ăn uống lành mạnh, tăng cường vận động thể chất và điều trị y tế có thể giúp làm giảm các tác động tiêu cực của bệnh, đặc biệt là khi chẩn đoán sớm và thay đổi lối sống có thể cải thiện hiệu quả điều trị.

Dữ liệu `diabetes_012_health_indicators_BRFSS2015.csv` được thu thập từ khảo sát của 253680 người dân Hoa Kỳ vào năm 2015, cung cấp một tập hợp các chỉ số sức khỏe và các yếu tố nguy cơ liên quan đến bệnh tiểu đường. Dữ liệu này bao gồm 22 biến quan sát, bao gồm các biến sau:

- `Diabetes_012`: Tình trạng bệnh tiểu đường (0: không tiểu đường, 1: tiền tiểu đường, 2: tiểu đường).
- `HighBP`: Tình trạng cao huyết áp (0: Không cao huyết áp, 1: Cao huyết áp).
- `HighChol`: Tình trạng cholesterol cao (0: Không cholesterol cao, 1: Cholesterol cao).
- `CholCheck`: Kiểm tra cholesterol trong vòng 5 năm qua (0: Không, 1: Có).
- `BMI`: Chỉ số khối cơ thể (Body Mass Index).
- `Smoker`: Người đã hút ít nhất 100 điếu thuốc trong suốt cuộc đời mình? (0: Không, 1: Có). Lưu ý: 5 gói = 100 điếu thuốc.
- `Age`: Tuổi của người tham gia khảo sát.

- **Stroke:** Đã từng được chẩn đoán bị đột quỵ (0: Chưa từng bị đột quỵ, 1: Đã từng bị đột quỵ).
- **HeartDiseaseorAttack:** Đã từng được chẩn đoán mắc bệnh tim mạch vành (CHD) hoặc nhồi máu cơ tim (MI) (0: Chưa từng, 1: Đã từng).
- **PhysActivity:** Có hoạt động thể chất trong vòng 30 ngày, không tính hoạt động liên quan tới công việc? (0: Không, 1: Có).
- **Fruits:** Tiêu thụ trái cây ít nhất 1 lần mỗi ngày? (0: Không, 1: Có).
- **Veggies:** Tiêu thụ rau ít nhất 1 lần mỗi ngày? (0: Không, 1: Có).
- **HvyAlcoholConsump:** Người uống rượu nặng (nam giới uống hơn 14 ly mỗi tuần và nữ giới uống hơn 7 ly mỗi tuần) (0: Không, 1: Có).
- **AnyHealthcare:** Có bất kỳ hình thức bảo hiểm y tế nào, bao gồm bảo hiểm y tế, kế hoạch trả trước như HMOs hoặc các kế hoạch chính phủ như Medicare hoặc Dịch vụ Y tế Ấn Độ không? (0: Không, 1: Có).
- **NoDocbcCost:** Trong 12 tháng qua, có cần gặp bác sĩ nhưng không thể vì chi phí không? (0: Không, 1: Có).
- **GenHlth:** Bạn có thể đánh giá sức khỏe chung của mình như thế nào? (1: Excellent (Cực kỳ tốt), 2: Very good (Rất tốt), 3: Good (Tốt), 4: Fair (Trung bình), 5: Poor (Kém)).
- **MentHlth:** Nghĩ về sức khỏe tinh thần của bạn, bao gồm căng thẳng, trầm cảm và các vấn đề về cảm xúc, trong 30 ngày qua, bạn đã có bao nhiêu ngày không cảm thấy khỏe về mặt tinh thần? (1 ~ 30 ngày).
- **PhysHlth:** Nghĩ về sức khỏe thể chất của bạn, bao gồm bệnh tật và chấn thương, trong 30 ngày qua, bạn đã có bao nhiêu ngày không cảm thấy khỏe về mặt thể chất? (1 ~ 30 ngày).
- **DiffWalk:** Bạn có gặp khó khăn nghiêm trọng khi đi bộ hoặc leo cầu thang không? (0: Không, 1: Có).
- **Sex:** Giới tính (0: Nữ, 1: Nam).
- **Age:** Nhóm tuổi (1 ~ 13)
- **Education:** Trình độ học vấn đã hoàn thành (1 ~ 6).
- **Income:** Mức thu nhập hộ gia đình hàng năm (Nếu người tham gia từ chối bất kỳ mức thu nhập nào, mã hóa là **Refused**) (1 ~ 8)

Bài báo cáo này sẽ phân tích các yếu tố sức khỏe có liên quan đến bệnh tiểu đường và đánh giá sự tương quan giữa các yếu tố này để xác định những yếu tố nguy cơ cao, từ đó hỗ trợ việc phát triển các mô hình dự đoán hiệu quả cho bệnh tiểu đường.

III. BÀI LÀM

1. Bản đề xuất phân tích và xử lý số liệu dựa trên các phương pháp đã được học trong học phần

- Mô tả dự án:

Mục tiêu là phân tích các yếu tố ảnh hưởng đến tình trạng tiểu đường và xây dựng mô hình dự đoán nguy cơ mắc bệnh dựa trên các chỉ số sức khỏe và thói quen sinh hoạt.

- Dữ liệu đầu vào:

Dữ liệu `diabetes_012_health_indicators_BRFSS2015.csv` chứa thông tin khảo sát của 253,680 người dân Hoa Kỳ (năm 2015) với 22 biến được quan sát.

- Phân tích và xử lý số liệu:

- Mô tả dữ liệu và phân phối của tình trạng bệnh tiểu đường theo từng nhóm (0: Không bị tiểu đường, 1: Tiền tiểu đường, 2: Tiểu đường).
- Xác định các yếu tố về sức khỏe ảnh hưởng đến nguy cơ mắc bệnh tiểu đường.
- Xây dựng các mô hình dự đoán nguy cơ mắc bệnh tiểu đường.

2. Các mục tiêu phân tích cần đạt được

- Mục tiêu của bài báo cáo này là phân tích các yếu tố sức khỏe có liên quan đến bệnh tiểu đường, đánh giá mức độ tương quan giữa các yếu tố này và xác định các yếu tố nguy cơ chính. Từ đó, hỗ trợ việc xây dựng các mô hình dự đoán hiệu quả nhằm cải thiện khả năng nhận diện sớm và quản lý bệnh tiểu đường.

3. Các phương pháp và chiến lược (các bước) phân tích cho mỗi mục tiêu đã đề ra

Để đạt được các mục tiêu đã đề ra thì sử dụng các phương pháp sau:

- **A/B Testing:** So sánh giữa hai nhóm để kiểm tra các yếu tố sức khỏe như mức độ vận động, chế độ ăn uống, hoặc chỉ số BMI có ảnh hưởng đáng kể đến bệnh tiểu đường hay không.
- **Logistic regression:** Xác định mối quan hệ giữa các biến giải thích (các yếu tố sức khỏe) và biến phụ thuộc (tình trạng mắc bệnh tiểu đường) để xây dựng mô hình, sử dụng để dự đoán nguy cơ mắc bệnh tiểu đường. Qua đó, đánh giá mô hình xem có phù hợp để áp dụng vào việc dự đoán nguy cơ mắc bệnh tiểu đường hay không.
- **Chiến lược:** Ta xây dựng mô hình thông qua các bước sau:
 - a) Khai báo các thư viện cần thiết và đọc dữ liệu.

- b) Tiền xử lý dữ liệu: Kiểm tra giá trị khuyết, trùng lặp, chuyển đổi cũng như thống kê sơ lược tổng quát về bộ dữ liệu đã cho,...
- c) Khám phá phân tích dữ liệu: Chỉ ra sự liên hệ giữa biến phản hồi với tất cả các biến khác có mặt trong bộ dữ liệu, từ đó rút ra nhận xét về những biến có ảnh hưởng lớn, ảnh hưởng ít hoặc không ảnh hưởng đến quá trình xây dựng mô hình hồi quy.
- d) Đặt câu hỏi và mục tiêu: Đặt các câu hỏi và mục tiêu rõ ràng để từ đó ta xây dựng mô hình theo mục tiêu đã định sẵn.
- e) Kiểm định Chi-square và A/B Testing: Dùng các phương pháp đã được học trong học phần, ta thực hiện các bài toán kiểm định sự độc lập của các biến cũng như sự ý nghĩa về mặt thống kê của chúng trước khi bước vào xây dựng mô hình.
- f) Mô hình hóa dữ liệu: Tiến hành xây dựng mô hình hồi quy với các bước rõ ràng như sau: Xây dựng mô hình, Lựa chọn mô hình và Chuẩn đoán mô hình.
- g) Đánh giá mô hình: Xử lý một số bước cũng như đánh giá lại mô hình đã xây dựng.
- h) Kết luận.

4. Mô tả và biểu diễn tổng hợp dữ liệu (bảng tổng hợp, biểu đồ)

a) Tiền xử lý dữ liệu:

Bảng thống kê dữ liệu sau khi được chuyển đổi và làm sạch:

```
##          diabetes_012          high_bp          high_chol
## No diabetes :190055   No High Blood:125359   No High Cholesterol:128273
## Pre-diabetes: 4629   High Blood   :104422   High Cholesterol   :101508
## Diabetes     : 35097
##
##
##
##
## chol_check          bmi          smoker          stroke
## No : 9298   Min.   :12.00   No Smoker:122781   No Stroke:219497
## Yes:220483   1st Qu.:24.00   Smoker   :107000   Stroke    : 10284
##
##           Median :27.00
##           Mean   :28.69
##           3rd Qu.:32.00
##           Max.   :98.00
##
## heart_diseaseor_attack phys_activity fruits          veggies
## No :206064             No : 61270   No : 88933   No : 47148
## Yes: 23717             Yes:168511   Yes:140848   Yes:182633
##
## hvy_alcohol_consump any_healthcare no_docbc_cost          gen_hlth
## No :215831             No : 12391   No :208455   Excellent:34907
## Yes: 13950             Yes:217390   Yes: 21326   Very Good:77536
##
##                                     Good      :73714
##                                     Fair       :31546
##                                     Poor       :12078
##
##
##          ment_hlth          phys_hlth          diff_walk          sex          age
## Min.   : 0.000   Min.   : 0.000   No :187155   Female:128854   9          :29736
## 1st Qu.: 0.000   1st Qu.: 0.000   Yes: 42626   Male  :100927   10         :29168
## Median : 0.000   Median : 0.000
## Mean   : 3.505   Mean   : 4.675
## 3rd Qu.: 2.000   3rd Qu.: 4.000
## Max.   :30.000   Max.   :30.000
##
##                                     (Other):81096
##
## education          income
## 1: 174 8          :71818
## 2: 4040 7          :40189
## 3: 9467 6          :35001
## 4:61158 5          :25345
## 5:66499 4          :19957
## 6:88443 3          :15922
##
##          (Other):21549
```

Nhận xét: Dữ liệu gồm các biến với ý nghĩa như sau:

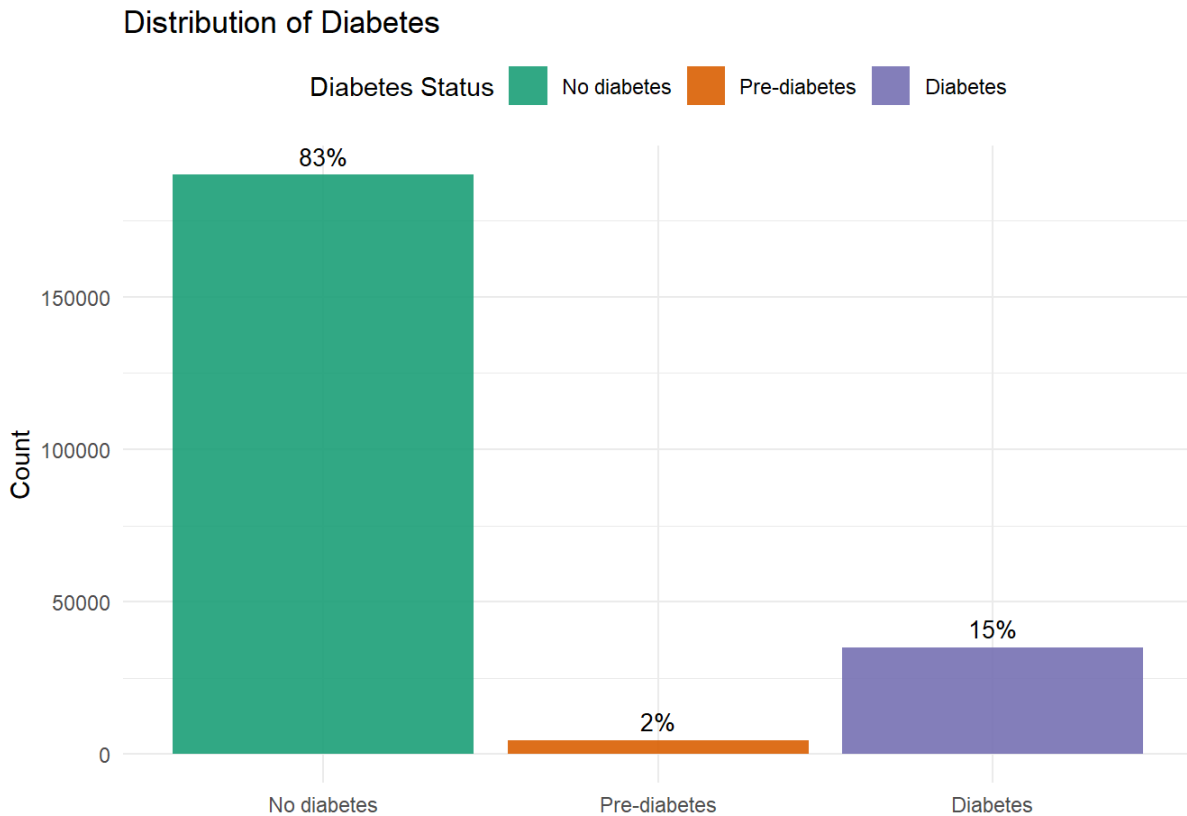
- **diabetes_012:** Chỉ tình trạng bệnh tiểu đường, gồm các giá trị **No diabetes** (Không tiểu đường) với 190055 quan sát; **Pre-diabetes** (Tiền tiểu đường) với 4629 quan sát; và **Diabetes** (Tiểu đường) với 35097 quan sát.
- **high_bp:** Chỉ tình trạng cao huyết áp, gồm hai giá trị **No High Blood** (Không bị cao huyết áp) với 125359 quan sát và **High Blood** (Cao huyết áp) với 104422 quan sát.
- **high_chol:** Chỉ tình trạng cholesterol cao, gồm hai giá trị **No High Cholesterol** (Không cholesterol cao) với 128273 quan sát và **High Cholesterol** (Cao cholesterol) với 101508 quan sát.
- **chol_check:** Chỉ việc đã kiểm tra cholesterol trong vòng 5 năm qua hay chưa, gồm các giá trị **No** với 9298 quan sát và **Yes** với 220483 quan sát.
- **bmi:** Chỉ số khối cơ thể, là một biến định lượng với phạm vi thuộc [12.00, 98.00] và có trung bình, trung vị lần lượt là 27.00 và 28.69.
- **smoker:** Chỉ tình trạng một người có hút trên 100 điếu thuốc trong cuộc đời mình hay không. Gồm các giá trị **No Smoker** (Không) với 122781 quan sát và **Smoker** (Có) với 107000 quan sát.
- **stroke:** Chỉ tình trạng một người đã từng được chẩn đoán bị đột quỵ hay chưa. Gồm các giá trị **No Stroke** (Chưa từng) với 219497 quan sát và **Stroke** (Đã từng) với 10284 quan sát.
- **heart_diseaseor_attack:** Chỉ tình trạng một người đã từng được chẩn đoán mắc bệnh tim mạch vành (CHD) hoặc nhồi máu cơ tim (MI) hay chưa. Gồm các giá trị **No** với 206064 quan sát và **Yes** với 13950 quan sát.
- **phys_activity:** Chỉ việc một người có hoạt động thể chất trong vòng 30 ngày (không tính hoạt động liên quan tới công việc) hay không. Gồm các giá trị **No** với 61270 quan sát và **Yes** với 168511 quan sát.
- **fruits:** Chỉ việc một người có tiêu thụ trái cây ít nhất 1 lần mỗi ngày hay không. Gồm các giá trị **No** với 88933 quan sát và **Yes** với 140848 quan sát.
- **veggies:** Chỉ việc một người có tiêu thụ rau củ ít nhất 1 lần mỗi ngày hay không. Gồm các giá trị **No** với 47148 quan sát và **Yes** với 182633 quan sát.
- **hvy_alcohol_consump:** Chỉ việc một người có uống rượu nặng (nam uống hơn 14 ly mỗi tuần, nữ uống hơn 7 ly mỗi tuần) hay không. Gồm các giá trị **No** với 215831 quan sát và **Yes** với 13950 quan sát.
- **any_healthcare:** Chỉ việc một người có bất kỳ hình thức bảo hiểm y tế nào (bảo hiểm y tế thông thường, HMOs, Medicare, Dịch vụ Y tế Ấn Độ,...) hay không. Gồm các giá trị **No** với 12931 quan sát và **Yes** với 217390 quan sát.
- **no_docbc_cost:** Chỉ việc một người trong 12 tháng qua có cần gặp bác sĩ nhưng không thể vì chi phí hay không. Gồm các giá trị **No** với 208455 quan

sát và Yes với 21326 quan sát.

- **gen_hlth**: Chỉ điểm đánh giá sức khỏe chung của bản thân. Gồm các giá trị **Excellent** (Xuất sắc), **Very Good** (Rất tốt), **Tốt**, **Fair** (Trung bình) và **Poor** (Kém) với số quan sát lần lượt là 34907, 77536, 73714, 31546 và 12078.
- **ment_hlth**: Chỉ số ngày (từ 1 đến 30) một người không cảm thấy khỏe về mặt tinh thần. Ở đây trung bình được thể hiện là khoảng 3.505 ngày.
- **phys_hlth**: Chỉ số ngày (từ 1 đến 30) một người cảm thấy không khỏe về mặt thể chất. Ở đây trung bình được thể hiện là khoảng 4.675 ngày.
- **diff_walk**: Chỉ việc một người có gặp khó khăn nghiêm trọng khi đi bộ hoặc leo cầu thang hay không. Gồm các giá trị **No** với 187155 quan sát và **Yes** với 42626 quan sát.
- **sex**: Chỉ giới tính, gồm 128854 quan sát là **Female** (Nữ) và 100927 quan sát là **Male** (Nam).
- **age**: Chỉ nhóm tuổi của một người, từ 1 (18-24 tuổi) đến 13 (trên 80 tuổi) với số quan sát như trên bảng.
- **education**: Chỉ trình độ học vấn đã hoàn thành của một người, từ 1 (Chưa học hoặc chỉ học mẫu giáo) đến 6 (Đại học năm 4 trở lên) với số quan sát như trên bảng.
- **income**: Chỉ mức thu nhập hộ gia đình hằng năm, từ 1 (Dưới 10,000\$) đến 8 (Trên 75000\$) với số quan sát như trên bảng.

b) Khám phá phân tích dữ liệu:

- Phân bố của bệnh tiểu đường:



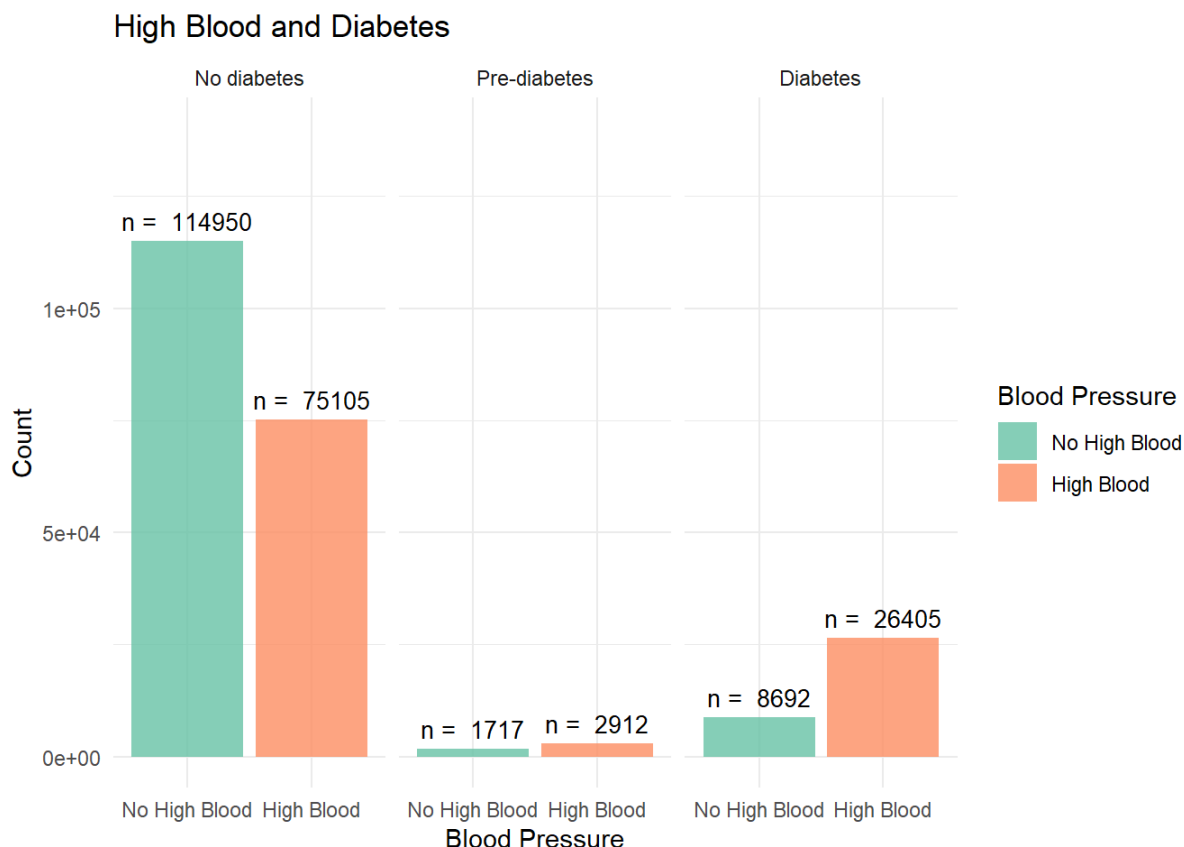
Biểu đồ trên cho ta thấy được:

- Dữ liệu không được cân bằng khi chỉ có khoảng 2% số người được hỏi là bị tiền tiểu đường và có khoảng 15% số người được hỏi bị mắc bệnh tiểu đường, còn lại là khoảng 83% số người được hỏi là không mắc bệnh.
- Sự mất cân bằng này có thể ảnh hưởng tiêu cực đến hiệu quả của mô hình dự đoán. Nhóm tiền tiểu đường, với số lượng nhỏ, sẽ khó được mô hình học và dự đoán chính xác, dẫn đến việc giảm độ chính xác trong phân loại và dự đoán.
- Để giải quyết vấn đề này, chúng ta có thể sử dụng kỹ thuật resampling như oversampling hoặc undersampling để cân bằng dữ liệu. Hoặc có thể điều chỉnh trọng số cho mô hình học máy để tăng độ quan trọng của nhóm thiểu số. Ngoài ra, ta cũng có thể sử dụng các phương pháp khác.

5. Phân tích để đạt được các mục tiêu đã đề ra, kết quả được biểu diễn dưới dạng bảng tổng hợp, biểu đồ

a) Kiểm tra mối liên hệ giữa Diabetes và các biến khác:

- High_BP và Diabetes:



diabetes_012 <fct>	No High Blood <chr>	High Blood <chr>
No diabetes	60.48%	39.52%
Pre-diabetes	37.09%	62.91%
Diabetes	24.77%	75.23%

3 rows

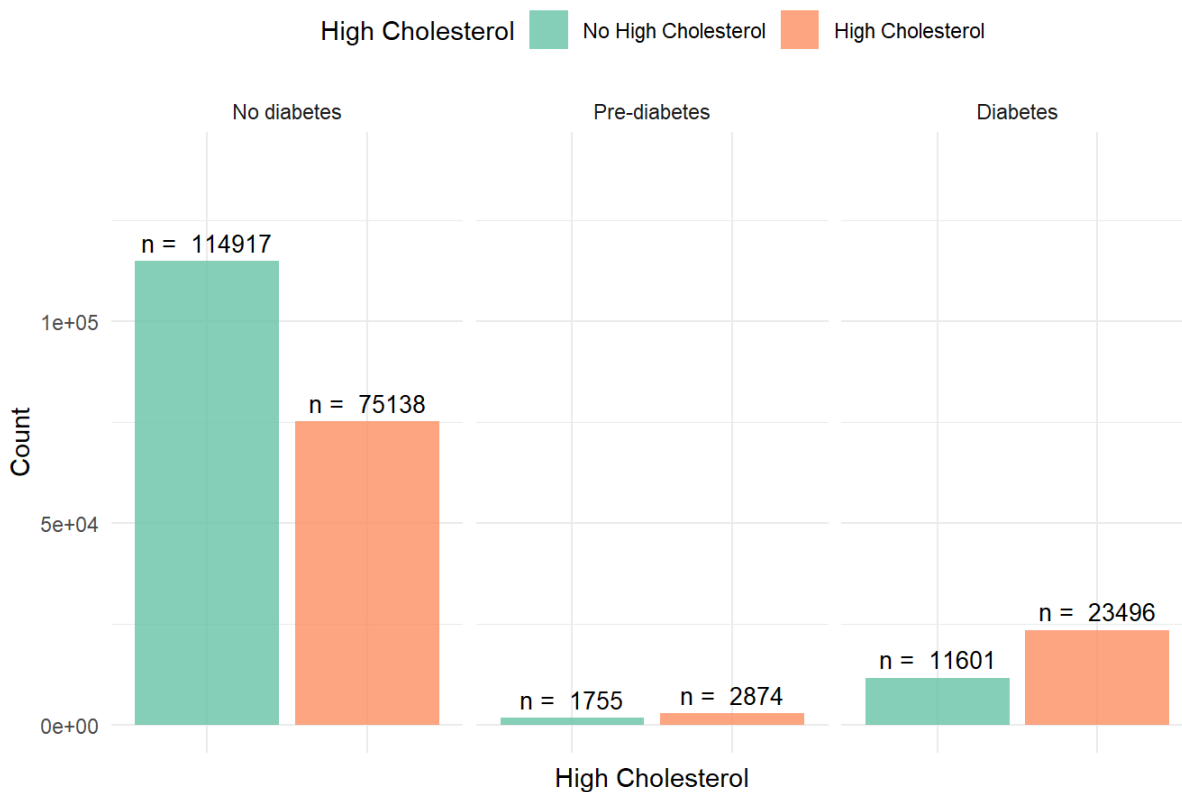
Qua biểu đồ và bảng tỷ số trên ta thấy được rằng:

- Biểu đồ cho thấy nhóm người bị tiền tiểu đường và tiểu đường có tỷ lệ bị cao huyết áp cao hơn đáng kể so với nhóm không bị tiểu đường. Điều này gợi ý một mối liên hệ tiềm năng giữa tình trạng tiểu đường và nguy cơ cao huyết áp.
- Ở nhóm không bị tiểu đường, tỷ lệ người bị cao huyết áp thấp hơn đáng kể so với các nhóm bị tiền tiểu đường và tiểu đường, cho thấy nguy cơ cao huyết áp tăng lên khi mức độ nghiêm trọng của tiểu đường gia tăng.

- Đáng chú ý, trong nhóm bị tiền tiểu đường và tiểu đường, tỷ lệ người bị cao huyết áp chiếm trên 60%, cho thấy tình trạng cao huyết áp rất phổ biến ở các nhóm này.

- High_Cholesterol và Diabetes:

High Cholesterol and Diabetes



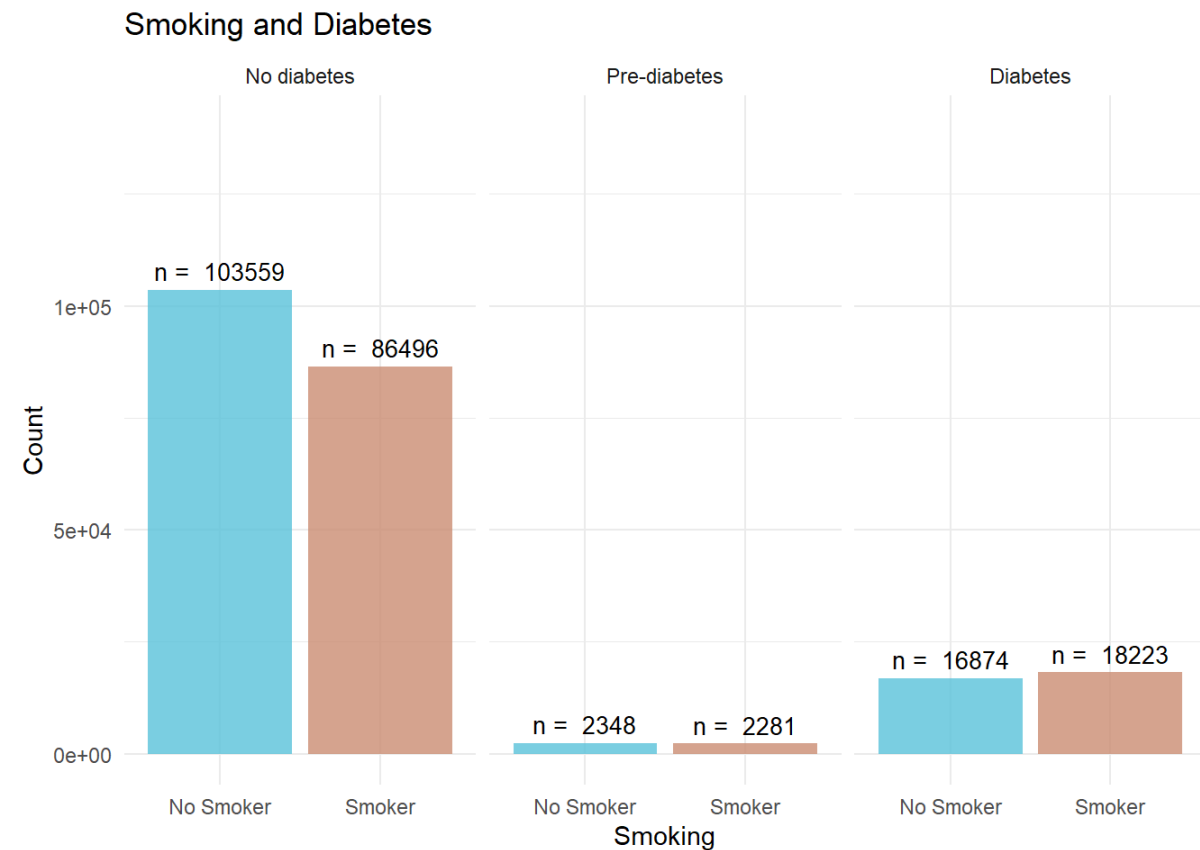
diabetes_012 <fct>	No High Cholesterol <chr>	High Cholesterol <chr>
No diabetes	60.47%	39.53%
Pre-diabetes	37.91%	62.09%
Diabetes	33.05%	66.95%

3 rows

Dựa vào biểu đồ và bảng tỷ số trên ta có được một số nhận xét sau:

- Tỷ lệ người bị cholesterol cao tăng dần khi mức độ bị bệnh tiểu đường tăng lên. Trong nhóm không bị tiểu đường, tỷ lệ người bị cholesterol cao chiếm 39.53%, trong khi đó tỷ lệ này ở nhóm tiền tiểu đường lại tăng lên tới 62.09% và đạt tới 66.95% ở nhóm bị tiểu đường.
- Có thể thấy rằng nguy cơ bị cholesterol cao tăng lên khi mức độ nghiêm trọng của bệnh tiểu đường tăng lên.
- Điều này cho thấy có một sự liên hệ giữa bệnh tiểu đường và cholesterol cao.

- Smoke và Diabetes:

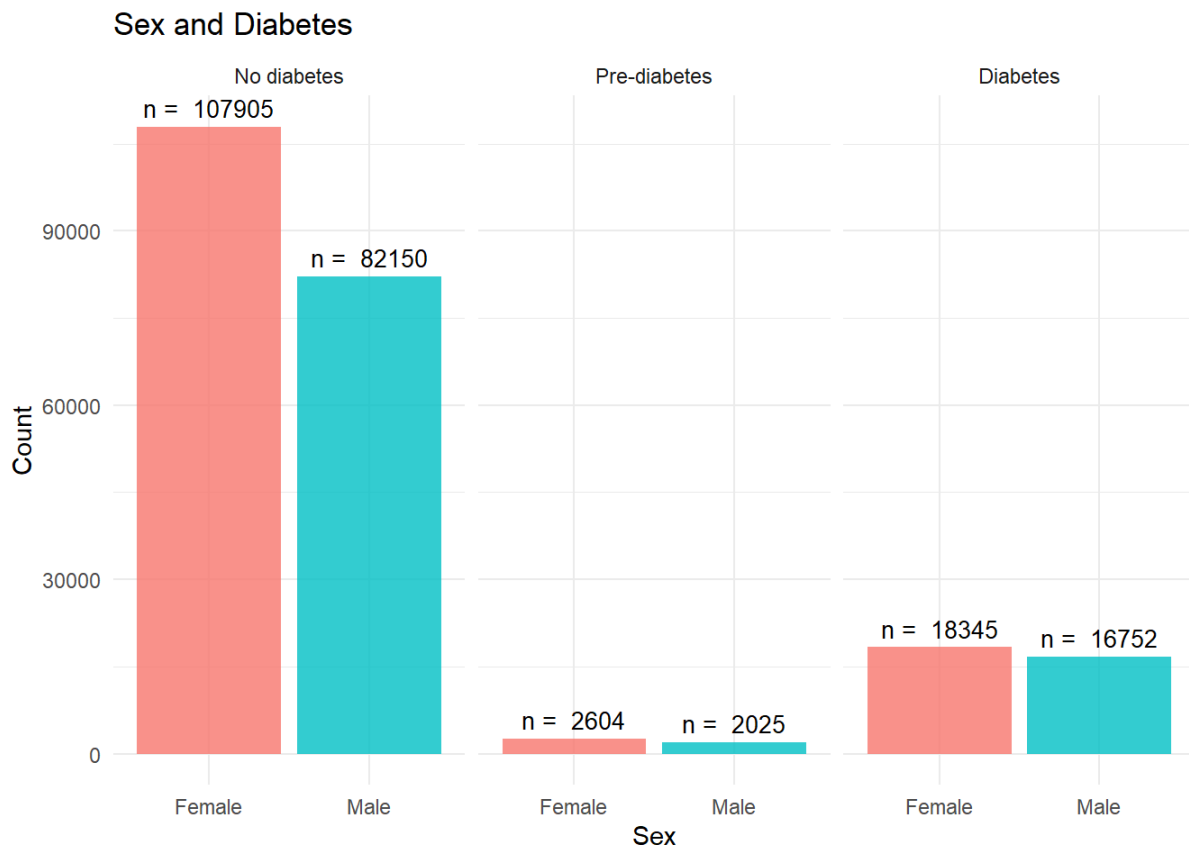


diabetes_012 <fct>	No Smoker <chr>	Smoker <chr>
No diabetes	54.49%	45.51%
Pre-diabetes	50.72%	49.28%
Diabetes	48.08%	51.92%

3 rows

Biểu đồ và bảng tỷ số cho ta thấy được là tỷ lệ thói quen hút thuốc đường như là tương đương bằng nhau ở cả 3 nhóm không bị tiểu đường, tiền tiểu đường và tiểu đường. Điều này cho thấy được là thói quen hút thuốc không ảnh hưởng nhiều đến nguy cơ mắc bệnh tiểu đường.

- Sex và Diabetes:

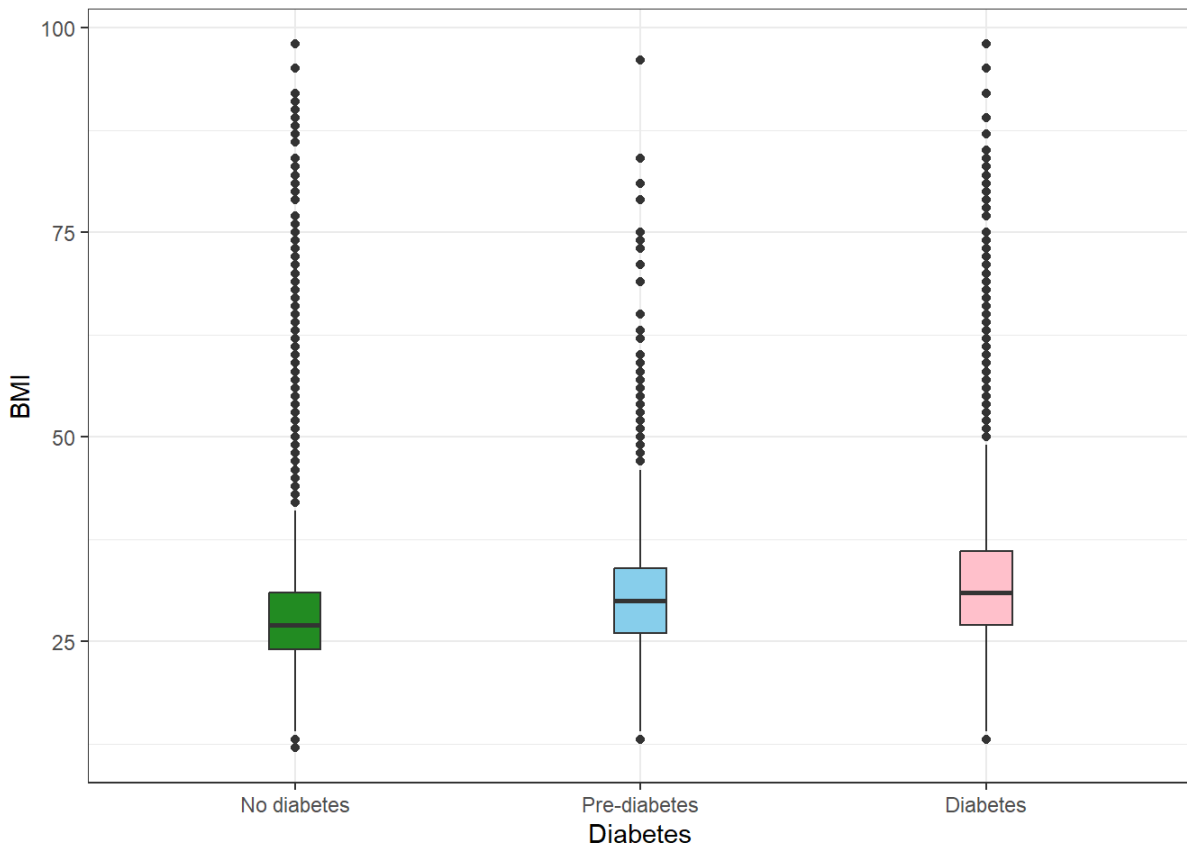


diabetes_012	Female	Male
<fct>	<chr>	<chr>
No diabetes	56.78%	43.22%
Pre-diabetes	56.25%	43.75%
Diabetes	52.27%	47.73%

3 rows

Biểu đồ và bảng tỷ số cho thấy được rằng không có sự khác biệt đáng kể giữa nam và nữ về tỷ lệ mắc bệnh tiểu đường. Qua đó, ta có thể kết luận rằng giới tính không ảnh hưởng nhiều đến nguy cơ mắc bệnh tiểu đường.

- BMI và Diabetes:



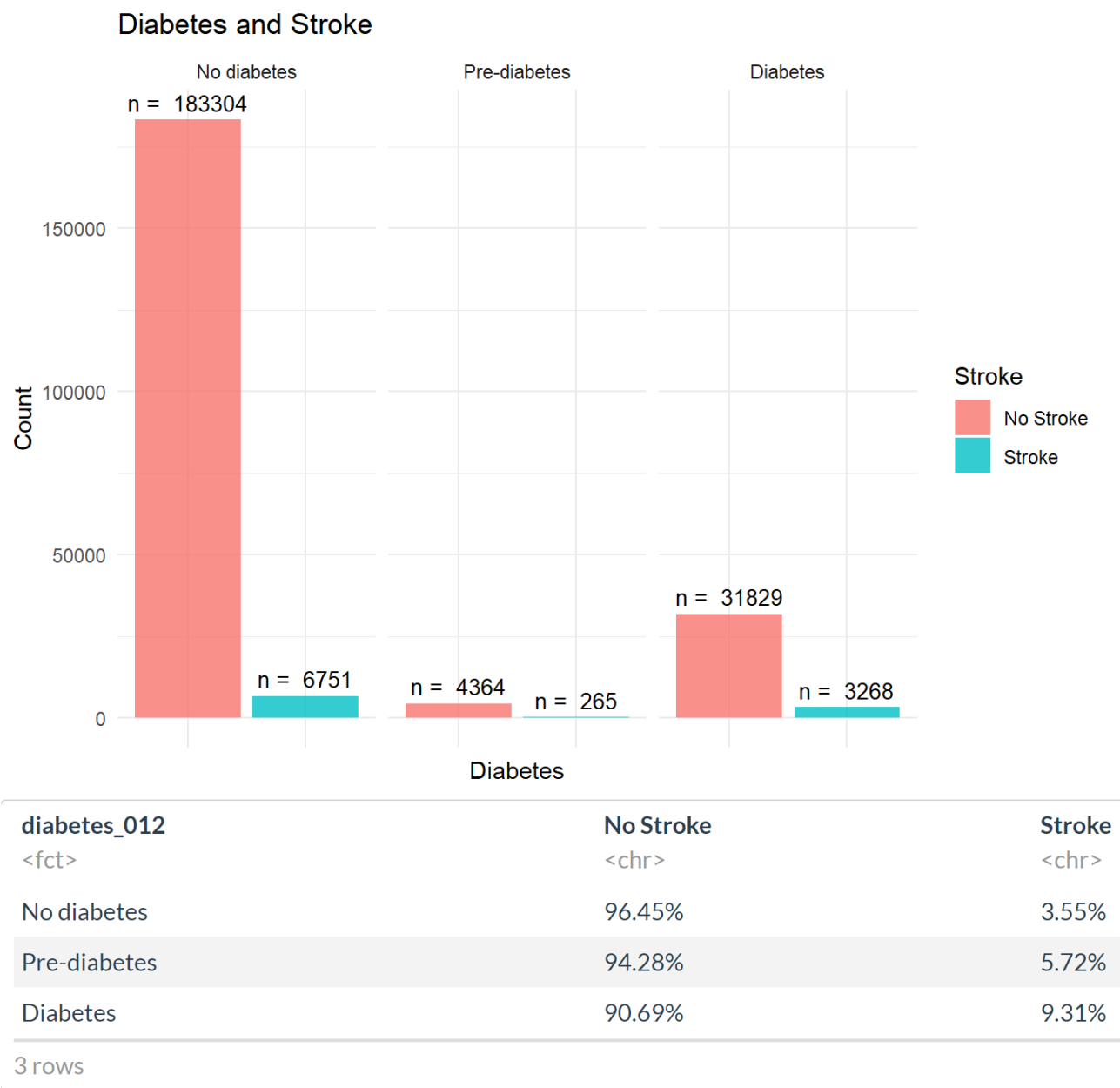
Biểu đồ boxplot cho ta thấy được sự phân bố các giá trị BMI giữa các tình trạng bệnh tiểu đường khác nhau. Xu hướng của mỗi nhóm được thể hiện như sau:

- **No Diabetes:** Những người nằm trong nhóm này có chỉ số BMI trung bình thấp hơn so với những người nằm trong nhóm tiền tiểu đường và tiểu đường. Phạm vi tứ phân vị chủ yếu nằm trong khoảng từ 20 đến 30.
- **Pre-Diabetes:** Chỉ số BMI trung bình của nhóm này cao hơn nhóm không bị tiểu đường nhưng thấp hơn nhóm bị tiểu đường. Điều này cho thấy mối liên quan giữa tăng BMI và nguy cơ tiến triển từ tình trạng tiền tiểu đường sang tiểu đường.
- **Diabetes:** Nhóm này có chỉ số BMI trung bình cao nhất so với 2 nhóm còn lại và khoảng tứ phân vị đã rộng ra hơn so với nhóm tiền tiểu đường, đồng thời cũng có rất nhiều giá trị ngoại lai (outliers) của BMI ở mức rất cao. Điều này cho thấy được rằng có một mối liên hệ mạnh giữa BMI và bệnh tiểu đường, chỉ số BMI cao có thể là một trong những yếu tố nguy cơ dẫn đến bệnh tiểu đường.

Hình ảnh trực quan chỉ ra rằng khi tình trạng tiểu đường tiến triển, chỉ số BMI cũng tăng theo. Điều này củng cố nhận định rằng có mối tương quan chặt giữa chỉ số BMI và nguy cơ của việc phát triển bệnh tiểu đường. Mặt khác, sự xuất hiện của các giá trị ngoại lai (outliers) ở cả 3 nhóm cũng chỉ ra rằng mặc dù yếu tố BMI cao là một yếu tố dự báo mạnh mẽ, nhưng không phải tất cả các trường

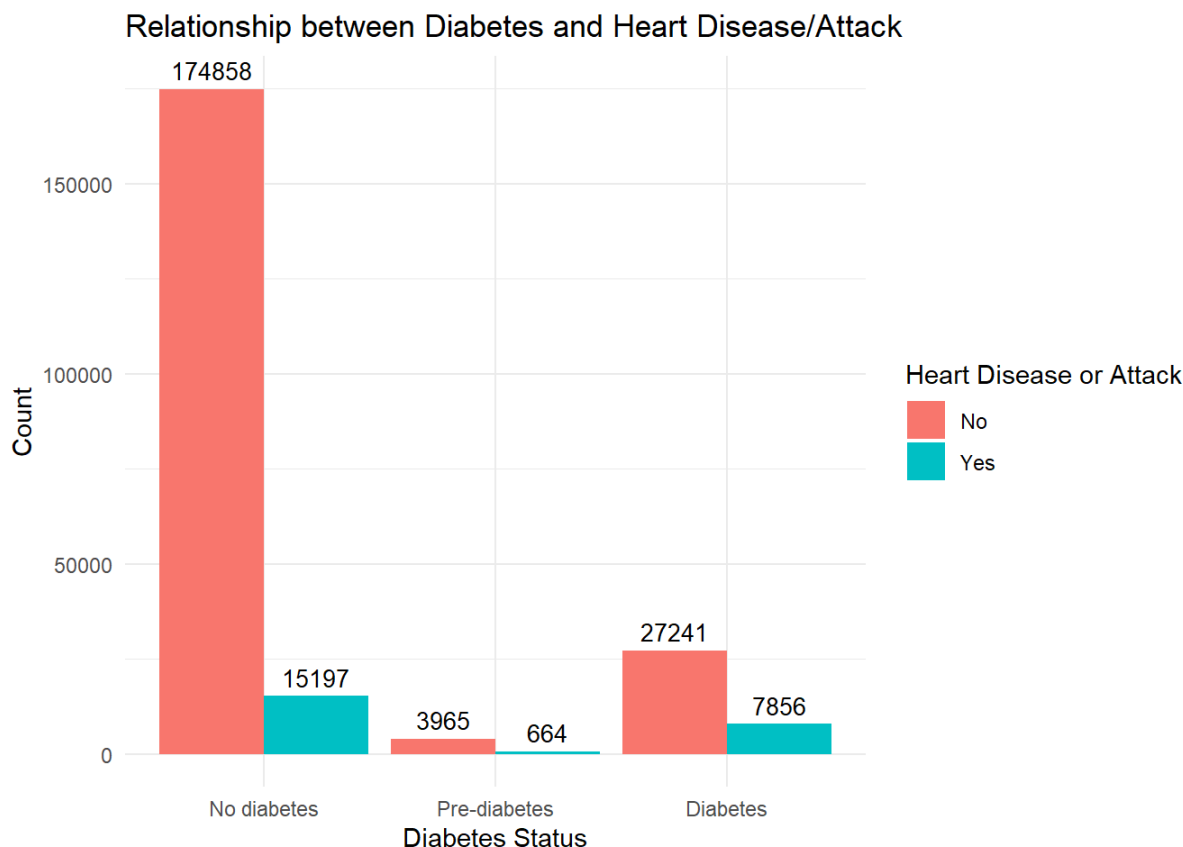
hợp đều đúng theo nhận định này.

- Stroke và Diabetes:



Theo như biểu đồ và bảng tỷ số phần trăm trên, ta có thể đưa ra được kết luận là bị đột quỵ trong quá khứ không phải là một yếu tố mạnh báo hiệu cho việc nguy cơ bị mắc bệnh tiểu đường, bởi vì theo như bảng tỷ số thì chưa đến 10% số người từng bị đột quỵ trong quá khứ mắc bệnh tiểu đường. Tuy nhiên khi so sánh nhóm người bị tiểu đường so với nhóm tiền tiểu đường thì tỷ lệ bị đột quỵ trong quá khứ lại cao hơn. Điều này cho thấy rằng bệnh tiểu đường có liên quan đến nguy cơ bị đột quỵ cao hơn.

- HeartDiseaseorAttack và Diabetes:



diabetes_012 <fct>	No <chr>	Yes <chr>
No diabetes	92.00%	8.00%
Pre-diabetes	85.66%	14.34%
Diabetes	77.62%	22.38%

3 rows

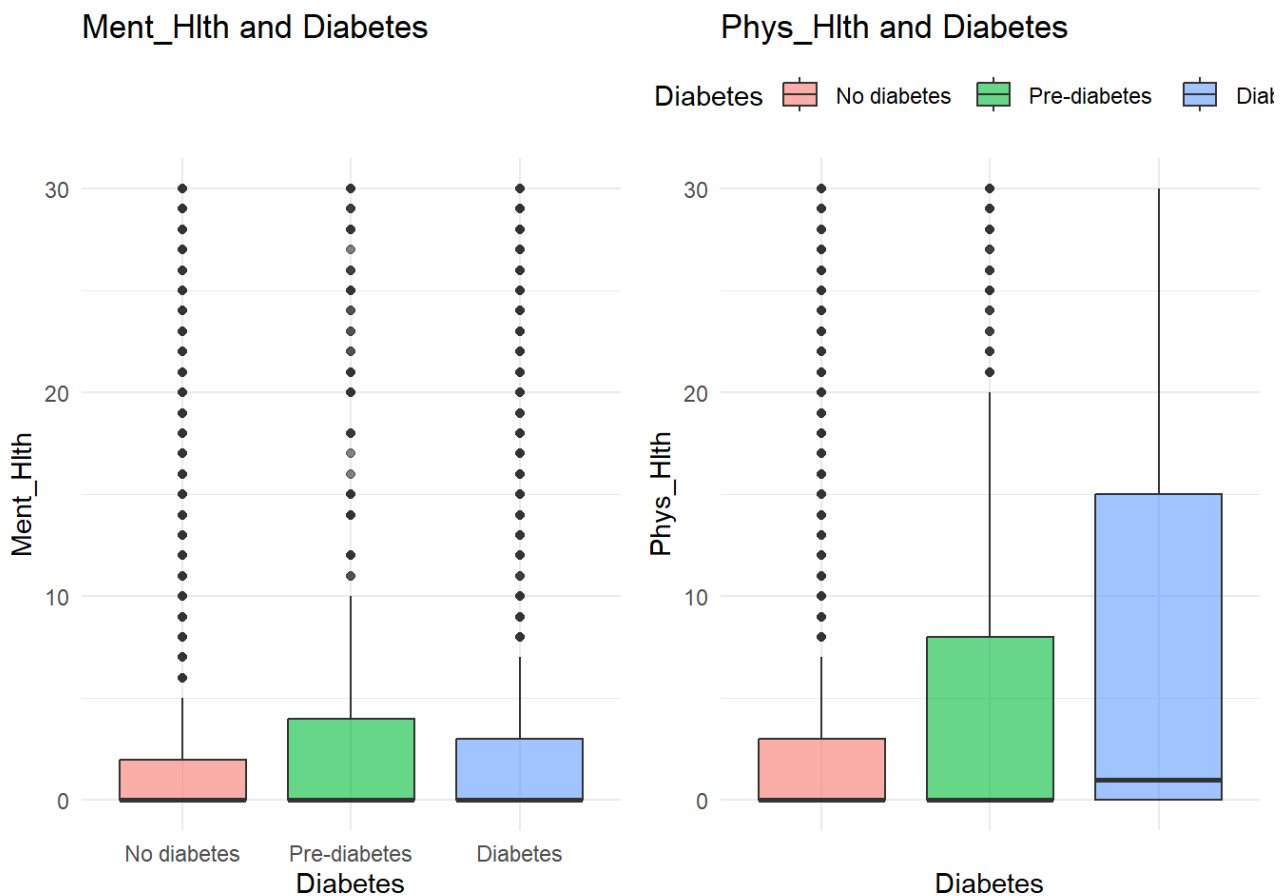
- Tỷ lệ người đã từng được chẩn đoán mắc bệnh tim mạch vành (CHD) hoặc nhồi máu cơ tim (MI) trong quá khứ tăng dần khi chuyển từ nhóm Không tiểu đường (8%) sang Tiền tiểu đường (14.34%) và cao nhất ở nhóm Tiểu đường (22.38%).
- Điều này cho thấy một mối liên hệ tiềm năng giữa tình trạng bệnh tiểu đường `diabetes_012` và khả năng từng được chẩn đoán mắc bệnh tim mạch vành (CHD) hoặc nhồi máu cơ tim (MI) trong `heart_diseaseor_attack`.
- Những người bị **Diabetes** có nguy cơ cao hơn đáng kể trong việc từng được chẩn đoán mắc bệnh tim mạch hoặc nhồi máu cơ tim so với nhóm **No diabetes**. Tương tự, nhóm **Pre-diabetes** cũng có nguy cơ cao hơn nhóm **No diabetes**, nhưng mức độ thấp hơn so với nhóm **Diabetes**.
- Dù vậy, có 8% người thuộc nhóm **No diabetes** vẫn từng được chẩn đoán mắc bệnh tim mạch hoặc nhồi máu cơ tim. Điều này cho thấy bệnh tiểu đường

không phải là yếu tố duy nhất dẫn đến tình trạng bệnh lý tim mạch này, mà có thể còn phụ thuộc vào các yếu tố nguy cơ khác.

- diabetes_012 với age, income, education, gen_hlth:

Thông qua việc dùng phép Kiểm định Chi-square cho tính độc lập (Chi-square Test for Independence) với mức ý nghĩa $\alpha = 0.05$ để kiểm định giả thuyết biến diabetes_012 độc lập với các biến age, income, education, gen_hlth, ta ra được các giá trị p_{value} đều nhỏ hơn 0.05, từ đó kết luận có sự liên hệ (phụ thuộc) giữa diabetes_012 và các biến age, income, education, gen_hlth.

- diabetes_012 với ment_hlth, phys_hlth:



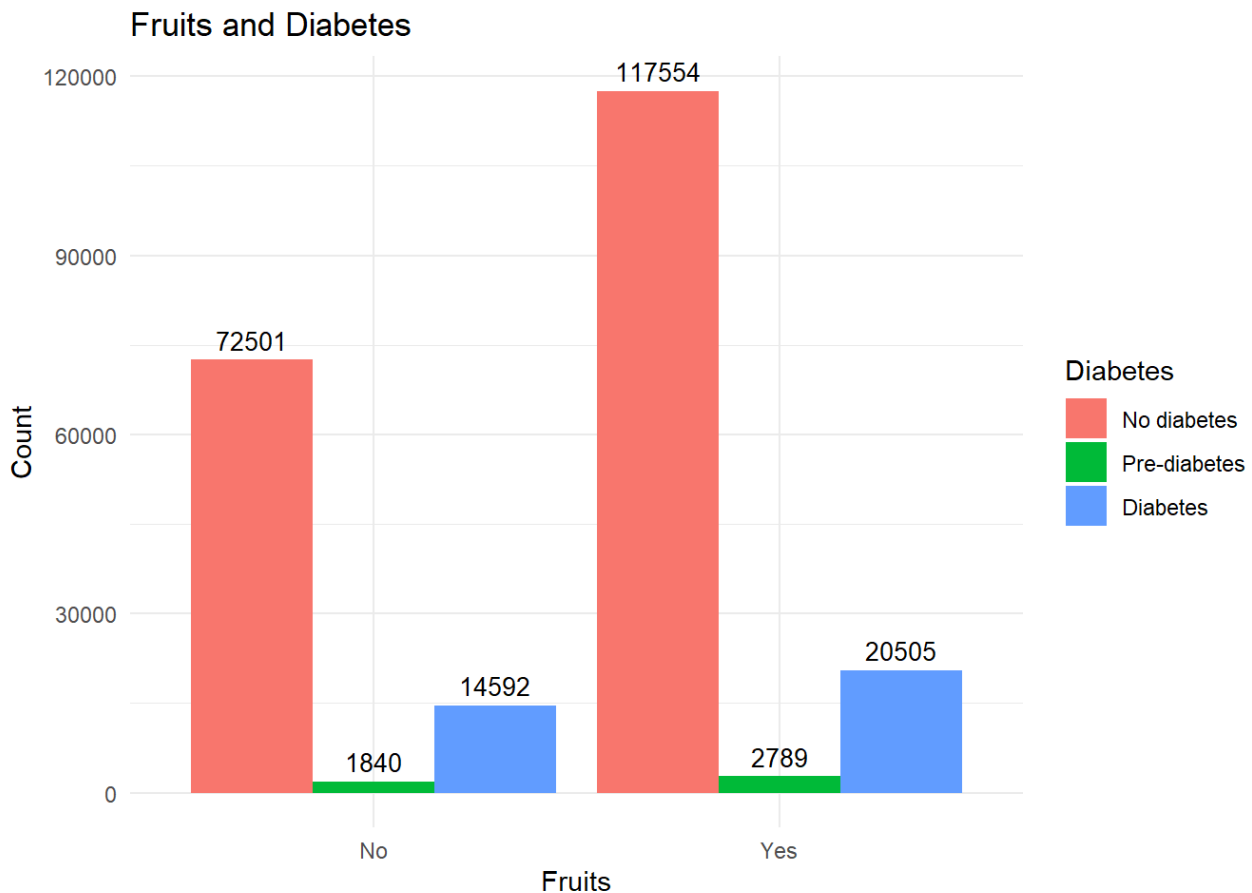
Từ biểu đồ trên, ta nhận xét:

- Cặp biến ment_hlth và diabetes_012: Phân phối các giá trị của ment_hlth tương đối giống nhau ở cả 3 nhóm: không bị tiểu đường (No Diabetes), tiền tiểu đường (Pre-Diabetes) và bị tiểu đường (Diabetes). Điều này cho thấy rằng không có sự khác biệt đáng kể về sức khỏe tinh thần giữa 3 nhóm này.
- Cặp biến phys_hlth và diabetes_012: Nhóm bị tiểu đường (Diabetes) có sự phân tán lớn hơn và có giá trị trung vị cao hơn so với 2 nhóm còn lại. Điều này cho thấy nhóm này có số ngày không khỏe về mặt thể chất nhiều hơn so với 2 nhóm còn lại. Nhóm tiền tiểu đường (Pre-Diabetes) có khoảng tứ phân vị lớn hơn nhóm không bị tiểu đường nhưng lại nhỏ hơn nhóm bị tiểu

đường, mặt khác nhóm này còn có các giá trị ngoại lai (outliers).

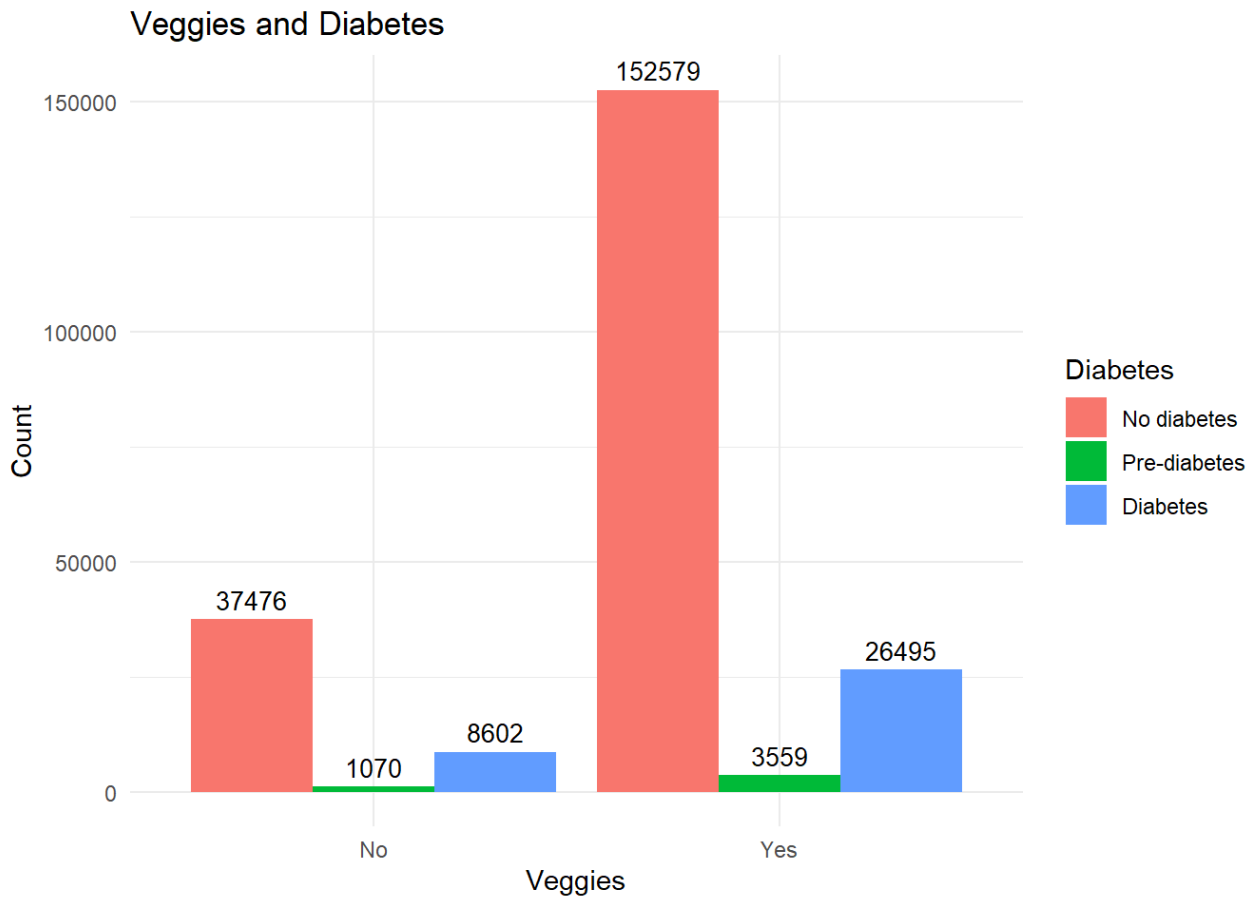
Hình ảnh trực quan chỉ ra rằng khi tình trạng tiểu đường tiến triển, chỉ số về số ngày không khoẻ về mặt thể chất cũng tăng theo. Điều này làm củng cố thêm nhận định rằng có mối tương quan giữa chỉ số về số ngày không khoẻ về mặt thể chất `phys_hlth`. Mặt khác, sự xuất hiện các giá trị ngoại lai ở nhóm không tiểu đường và tiền tiểu đường cũng cho thấy rằng không thể khẳng định rõ được là biến `phys_hlth` có ảnh hưởng lớn đến việc xác định nguy cơ bệnh tiểu đường hay không. Điều này cũng tương tự với biến `ment_hlth` trong việc xác định nguy cơ bệnh tiểu đường.

- Diabetes và Fruits:



Biểu đồ trên cho thấy được rằng việc tiêu thụ trái cây ít nhất 1 lần mỗi ngày sẽ có thể làm giảm nguy cơ mắc bệnh tiểu đường.

- Diabetes và Veggies:



Tương tự với việc tiêu thụ trái cây ít nhất 1 lần mỗi ngày thì việc tiêu thụ rau củ ít nhất 1 lần mỗi ngày cũng giúp giảm nguy cơ mắc bệnh tiểu đường.

- **Nhận xét:** Từ quá trình khám phá phân tích dữ liệu trên, ta có được các biến quan trọng có thể ảnh hưởng đến nguy cơ mắc bệnh tiểu đường như `high_bp`, `high_chol`, `bmi`, `phys_hlth`, `ment_hlth`, `gen_hlth`, `age`, `income`, `education`, `heart_diseaseor_attack`, `veggies` và `fruits`.

b) Kiểm định giả thuyết

- Kiểm tra sự độc lập giữa biến `diabetes_012` và các biến nhị thức khác:

```
## Is diabetes_012 and high_bp independent of each other? FALSE
## Is diabetes_012 and high_chol independent of each other? FALSE
## Is diabetes_012 and chol_check independent of each other? FALSE
## Is diabetes_012 and smoker independent of each other? FALSE
## Is diabetes_012 and stroke independent of each other? FALSE
## Is diabetes_012 and heart_diseaseor_attack independent of each other? FALSE
## Is diabetes_012 and phys_activity independent of each other? FALSE
## Is diabetes_012 and fruits independent of each other? FALSE
## Is diabetes_012 and veggies independent of each other? FALSE
## Is diabetes_012 and hvy_alcohol_consump independent of each other? FALSE
## Is diabetes_012 and any_healthcare independent of each other? FALSE
## Is diabetes_012 and no_docbc_cost independent of each other? FALSE
## Is diabetes_012 and diff_walk independent of each other? FALSE
## Is diabetes_012 and sex independent of each other? FALSE
```

Kết quả cho thấy, có sự liên hệ (phụ thuộc) với nhau giữa biến `diabetes_012` và các biến nhị thức khác có trong bộ dữ liệu.

- ANOVA cho trung bình `bmi` theo từng nhóm `diabetes_012`:

Ta có bảng tổng hợp trung bình và phương sai của `bmi` theo từng nhóm `diabetes_012`:

<code>diabetes_012</code>	<code>n</code>	<code>mean_bmi</code>	<code>sd_bmi</code>
<fct>	<int>	<dbl>	<dbl>
No diabetes	190055	28.03053	6.474981
Pre-diabetes	4629	30.72607	6.965973
Diabetes	35097	31.96424	7.380385

3 rows

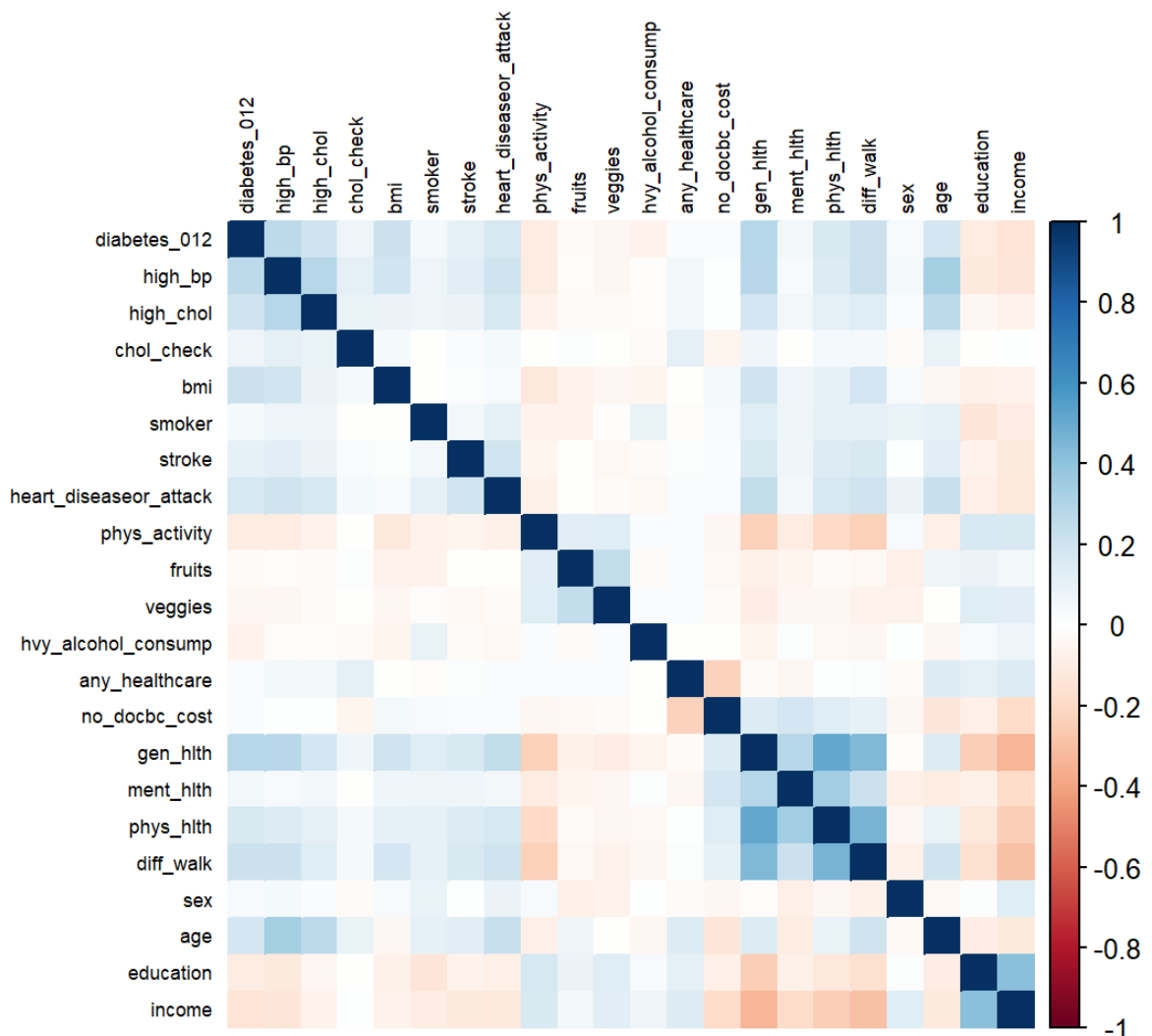
Do $p_{value} < 0.05$ nên sự khác biệt trung bình `bmi` giữa các nhóm `diabetes_012` là có ý nghĩa thống kê.

Sau khi đã hoàn tất việc khám phá và phân tích dữ liệu, cũng như kiểm định các giả thuyết liên quan, ta sang thực hiện quá trình Mô hình hóa dữ liệu.

c) Đánh giá sự tương quan

Nhận thấy rằng Tiền tiểu đường cũng tương tự như Tiểu đường, nên ta sẽ gộp nhóm Tiền tiểu đường vào nhóm Tiểu đường. Bây giờ, thay vì tạo ra một mô hình để xác định xem một người nào đó không có nguy cơ mắc bệnh tiểu đường, có nguy cơ bị tiền tiểu đường hay có nguy cơ mắc bệnh tiểu đường hay không thì giờ đây, mô hình sẽ chỉ xác định xem một người có nguy cơ hay không có nguy cơ mắc bệnh tiểu đường.

Trước khi xây dựng mô hình, ta sẽ vẽ biểu đồ tương quan giữa các biến định lượng và biến phản hồi (`diabetes_012`) trong dữ liệu gốc để ban đầu để xác định mức độ tương quan giữa chúng.



Dựa vào biểu đồ tương quan trên, ta có thể nhận thấy được mối quan hệ giữa biến `diabetes_012` và các biến khác trong dữ liệu như sau:

- Biến `fruits`, `any_healthcare`, `no_docbc_cost`, `veggies`, `sex` và `hvy_alcohol_consump` có mối tương quan yếu với biến `diabetes_012`.

- Các biến `high_bp`, `high_chol`, `chol_check`, `bmi`, `smoker`, `stroke`, `heart_diseaseor_attack`, `phys_activity`, `gen_hlth`, `ment_hlth`, `phys_hlth`, `diff_walk`, `age`, `education` và `income` có mối tương quan đáng kể với biến `diabetes_012`.

d) Xây dựng mô hình

Vì biến phản hồi (Y) là biến nhị phân nên mô hình phù hợp nhất là mô hình logistic regression.

Từ các quá trình khám phá phân tích dữ liệu và biểu đồ tương quan đã phân tích trước đó, ta sẽ chọn ra các biến sau: `high_bp`, `high_chol`, `bmi`, `phys_hlth`, `ment_hlth`, `gen_hlth`, `age`, `income`, `education`, `heart_diseaseor_attack`, `stroke` để xây dựng mô hình. Khi đó ta có được mô hình:

$$\log \frac{p(x)}{1-p(x)} = \beta_0 + \beta_1 \text{high_bp} + \beta_2 \text{high_chol} + \beta_3 \text{bmi} + \beta_4 \text{phys_hlth} \\ + \beta_5 \text{ment_hlth} + \beta_6 \text{gen_hlth} + \beta_7 \text{age} + \beta_8 \text{income} \\ + \beta_9 \text{education} + \beta_{10} \text{heart_diseaseor_attack} + \beta_{11} \text{stroke}$$

##	GVIF	Df	GVIF^(1/(2*Df))
## high_bp	1.127256	1	1.061723
## high_chol	1.069771	1	1.034297
## bmi	1.095414	1	1.046620
## phys_hlth	1.689652	1	1.299866
## ment_hlth	1.251609	1	1.118753
## gen_hlth	1.870750	4	1.081439
## age	1.301735	12	1.011048
## income	1.476939	7	1.028247
## education	1.305536	5	1.027020
## heart_diseaseor_attack	1.130010	1	1.063019
## stroke	1.067232	1	1.033069

Thông qua quá trình kiểm tra đa cộng tuyến, vì các VIF_j , $j = \overline{1, 11}$ đều nhỏ hơn 5 nên ta kết luận rằng không có hiện tượng đa cộng tuyến trong mô hình.

Kết quả ước lượng hệ số như sau:

```
##
## Call:
## glm(formula = diabetes_012 ~ high_bp + high_chol + bmi + phys_hlth +
##      ment_hlth + gen_hlth + age + income + education + heart_diseaseor_attack +
##      stroke, family = "binomial", data = df_clean)
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)      -6.4641673    0.2193616 -29.468  < 2e-16 ***
## high_bpHigh Blood    0.6791728    0.0138380  49.080  < 2e-16 ***
## high_cholHigh Cholesterol 0.5538696    0.0129371  42.813  < 2e-16 ***
## bmi                0.0588584    0.0008659  67.975  < 2e-16 ***
## phys_hlth          -0.0021946    0.0007482  -2.933  0.003356 **
## ment_hlth          -0.0030894    0.0008010  -3.857  0.000115 ***
## gen_hlthVery Good    0.6147576    0.0303114  20.281  < 2e-16 ***
## gen_hlthGood         1.2234423    0.0295070  41.463  < 2e-16 ***
## gen_hlthFair         1.6812483    0.0320150  52.514  < 2e-16 ***
## gen_hlthPoor         1.8642341    0.0389416  47.873  < 2e-16 ***
## age2                0.2380207    0.1281222   1.858  0.063203 .
## age3                0.4264564    0.1170058   3.645  0.000268 ***
## age4                0.8762618    0.1109239   7.900  2.80e-15 ***
## age5                1.0815244    0.1086848   9.951  < 2e-16 ***
## age6                1.3340694    0.1069542  12.473  < 2e-16 ***
## age7                1.5227579    0.1058449  14.387  < 2e-16 ***
## age8                1.6181657    0.1054446  15.346  < 2e-16 ***
## age9                1.8433531    0.1052061  17.521  < 2e-16 ***
## age10              1.9998605    0.1051953  19.011  < 2e-16 ***
```

```
## age11          2.0543422  0.1055531  19.463 < 2e-16 ***
## age12          1.9790696  0.1061892  18.637 < 2e-16 ***
## age13          1.8186682  0.1062751  17.113 < 2e-16 ***
## income2       -0.0284621  0.0340915  -0.835  0.403790
## income3       -0.0724238  0.0328155  -2.207  0.027314 *
## income4       -0.1074446  0.0321312  -3.344  0.000826 ***
## income5       -0.1683231  0.0316138  -5.324  1.01e-07 ***
## income6       -0.2367878  0.0310100  -7.636  2.24e-14 ***
## income7       -0.2594555  0.0312150  -8.312 < 2e-16 ***
## income8       -0.3456297  0.0306664 -11.271 < 2e-16 ***
## education2      0.1580255  0.1922643   0.822  0.411124
## education3      0.0101114  0.1903925   0.053  0.957646
## education4     -0.1017247  0.1890889  -0.538  0.590596
## education5     -0.0729971  0.1891572  -0.386  0.699566
## education6     -0.1193565  0.1892674  -0.631  0.528287
## heart_diseaseor_attackYes 0.2699509  0.0170740  15.811 < 2e-16 ***
## strokeStroke    0.1451967  0.0242197   5.995  2.03e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##      Null deviance: 211598  on 229780  degrees of freedom
## Residual deviance: 172085  on 229745  degrees of freedom
## AIC: 172157
##
## Number of Fisher Scoring iterations: 6
```

Từ kết quả trên, ta có thể rút ra một số kết luận:

- Hệ số dương của biến `high_bp` High Blood cho thấy những người bị huyết áp cao có khả năng mắc bệnh tiểu đường cao hơn so với những người không bị huyết áp cao. Việc tăng một đơn vị của biến `high_bp` sẽ làm tăng tỷ lệ mắc bệnh tiểu đường lên $e^{0.679} \approx 1.972$ lần.
- Hệ số dương của biến `high_chol` High Cholesterol cho thấy những người bị cholesterol cao có khả năng mắc bệnh tiểu đường cao hơn so với những người không bị cholesterol cao.
- Hệ số dương đối với `bmi` cho thấy rằng việc tăng chỉ số BMI có liên quan đến việc tăng nguy cơ mắc bệnh tiểu đường.
- Việc tăng một đơn vị BMI có liên quan đến việc tăng tỷ lệ mắc bệnh tiểu đường là $e^{0.0589} \approx 1.06$ lần.
- Hệ số âm đối với `phys_hlth` cho thấy rằng người cảm thấy không khỏe về mặt thể chất với số ngày càng nhiều thì càng dễ có nguy cơ mắc bệnh tiểu đường hơn là những người không khỏe về mặt thể chất với số ngày càng ít.
- Hệ số âm cho `ment_hlth` cho thấy người nào cảm thấy không khỏe về mặt

tinh thần với số ngày càng ít thì càng ít có khả năng bị mắc bệnh tiểu đường hơn là những người không khỏe về mặt tinh thần với số ngày càng nhiều.

- Từ hệ số dương cho `gen_hlth` Very Good, `gen_hlth` Good, `gen_hlth` Fair, `gen_hlth` Poor càng tăng dần khi chỉ số từ Very Good cho tới Poor ta thấy rằng người có sức khỏe tốt, khá, trung bình hoặc kém hơn so với người có sức khỏe xuất sắc có khả năng mắc bệnh tiểu đường cao hơn.
- Với các biến `education2`, `education3`, `education4`, `education5` và `education6` có p_{value} đều lớn hơn $\alpha = 0.05$ nên chúng không có ý nghĩa thống kê đối với mô hình hồi quy logistic.
- Với hệ số âm cho các biến từ `income2` cho tới `income6` càng giảm dần thì cho ta thấy được là một người có thu nhập cao thì càng ít có khả năng mắc bệnh tiểu đường hơn vì có các dịch vụ y tế, chăm sóc sức khỏe tốt hơn.
- Với hệ số dương cho biến `age` càng tăng dần theo độ tuổi cho ta biết được là người càng cao tuổi càng có nguy cơ bị mắc bệnh tiểu đường.
- Hệ số dương của biến `heart_diseaseor_attack` cho thấy rằng những người đã từng mắc bệnh tim hoặc đau tim thì sẽ dễ mắc bệnh tiểu đường hơn là những người không mắc bệnh tim hoặc đau tim. Nói đúng hơn là, nếu tăng một đơn vị về biến `heart_diseaseor_attack` thì sẽ làm tăng tỉ lệ mắc bệnh tiểu đường là $e^{0.27} \approx 1.31$ lần.
- Hệ số dương của biến `stroke` cũng cho ta thấy được rằng những người đã từng mắc đột quỵ thì sẽ dễ dàng mắc bệnh tiểu đường hơn là người chưa từng bị mắc bệnh đột quỵ.

Tiếp theo, ta áp dụng phương pháp bootstrap để ước lượng khoảng tin cậy và kiểm định giả thuyết $\beta_j = 0$. Tức là với mức ý nghĩa $\alpha = 0.05$, ta đi kiểm tra cặp Giả thuyết và Đối thuyết sau:

$$H_0 : \beta_j = 0$$

$$H_1 : \beta_j \neq 0$$

Kiểm định Giả thuyết này tương đương với việc trả lời câu hỏi: "Có sự tồn tại mối liên hệ giữa biến X_j và Y hay không?".

```
## [1] "(Intercept)" "high_bpHigh Blood"
## [3] "high_cholHigh Cholesterol" "bmi"
## [5] "phys_hlth" "ment_hlth"
## [7] "gen_hlthVery Good" "gen_hlthGood"
## [9] "gen_hlthFair" "gen_hlthPoor"
## [11] "age3" "age4"
## [13] "age5" "age6"
## [15] "age7" "age8"
## [17] "age9" "age10"
## [19] "age11" "age12"
## [21] "age13" "income3"
## [23] "income4" "income5"
## [25] "income6" "income7"
## [27] "income8" "heart_diseaseor_attackYes"
## [29] "strokeStroke"
```

Sau khi tiến hành kiểm định, ta có thể thấy rằng có 10 biến có ý nghĩa thống kê là `high_bp`, `high_chol`, `bmi`, `phys_hlth`, `ment_hlth`, `gen_hlth`, `age`, `income`, `heart_diseaseor_attack`, `stroke`. Do đó ta có thể loại bỏ biến `education` ra khỏi mô hình.

Bây giờ, giả sử ta có thông tin về 5 người và chúng ta muốn dự đoán xác suất họ mắc bệnh tiểu đường. Thông tin được đưa ra như sau:

id	high_bp	high_chol	bmi	phys_hlth	ment_hlth	gen_hlth	age	income	education	heart_diseaseor_attack	stroke
1	High Blood	High Cholesterol	27	0	0	Good	10	6	6	No	No Stroke
2	High Blood	High Cholesterol	40	0	0	Fair	9	6	6	No	No Stroke
3	No High Blood	No High Cholesterol	32	0	0	Good	4	8	6	No	No Stroke
4	High Blood	High Cholesterol	24	0	0	Good	13	3	5	No	No Stroke
5	No High Blood	No High Cholesterol	24	0	0	Excellent	6	8	5	No	No Stroke

Ta sẽ phân loại 5 người trên thành hai nhóm là: **Không mắc bệnh tiểu đường** và **Mắc bệnh tiểu đường** bằng cách so sánh xác suất ở ngưỡng $c = 0.5$.

Kết quả thu được như sau:

```
##           1           2           3           4           5
## "No Diabetes" "Diabetes" "No Diabetes" "No Diabetes" "No Diabetes"
```

e) Lựa chọn mô hình

Từ những mô hình trên, ta vẫn chưa thể kết luận được là những mô hình trên đâu là mô hình tốt nhất. Do đó ta sẽ thực hiện các phương pháp để tìm ra mô hình hồi quy tốt nhất.

Ta có thể sử dụng 1 trong 3 phương pháp sau:

- Hồi quy từng bước (Stepwise Regression).
- Hồi quy từng bước và cross - validation.
- Phương pháp co hệ số (Shrinkage method).

Vì số lượng biến khá nhiều nên ta sẽ sử dụng phương pháp co hệ số, trong đó ta thực hiện ridge regression và lasso regression (cụ thể ta dùng 5-fold cross-validation).

Measure: Mean-Squared Error

##

##	Lambda	Index	Measure	SE	Nonzero
## min	6.95e-05	79	0.2347	0.0009652	44
## 1se	2.87e-03	39	0.2357	0.0009679	31

Kết quả cho thấy λ tối ưu là $\lambda = 6.95 \times 10^{-5}$, ứng với bình phương sai số (mean squared error) là 0.2347. Đồng thời, ta thu được 44 hệ số khác 0.

Sau khi ước lượng các hệ số trong mô hình sử dụng toàn bộ dữ liệu bằng cách sử dụng λ tối ưu vừa thu được, ta được kết quả:

```
## 46 x 1 sparse Matrix of class "dgCMatrix"
##                                     s1
## (Intercept)                      -4.258237445
## high_bpHigh Blood                 0.768892522
## high_cholHigh Cholesterol         0.528393908
## chol_checkYes                     0.297756144
## bmi                               0.043143883
## smokerSmoker                      .
## strokeStroke                      0.023217162
## heart_diseaseor_attackYes        0.296352592
## phys_activityYes                  .
## fruitsYes                         .
## veggiesYes                       .
## hvy_alcohol_consumpYes            -0.284022300
## any_healthcareYes                 .
## no_docbc_costYes                  .
## gen_hlthVery Good                 .
## gen_hlthGood                      0.384840034
## gen_hlthFair                      0.795754454
## gen_hlthPoor                      0.850356634
## ment_hlth                         .
## phys_hlth                         .
## diff_walkYes                      0.246242021
## sexMale                           0.005204262
```

## age2	-0.117935177
## age3	-0.239945418
## age4	-0.109768485
## age5	-0.049323312
## age6	.
## age7	.
## age8	.
## age9	.
## age10	0.151184288
## age11	0.173798936
## age12	0.056402728
## age13	.
## education2	.
## education3	.
## education4	.
## education5	.
## education6	.
## income2	.
## income3	.
## income4	.
## income5	.
## income6	.
## income7	.
## income8	-0.097666076

Kết luận: Kết quả cho thấy các biến smokerSmoker, phys_activityYes, fruitsYes, veggiesYes, any_healthcareYes, no_docbc_costYes, gen_hlthVery Good, ment_hlth, age6, age7, age8, age9, age13, education2, education3, education4, education5, education6, income2, income3, income4, income5, income6, income7 có hệ số ước lượng bằng không.

Như vậy, ta có thể sử dụng mô hình logistic regression với các biến còn lại để dự đoán biến phụ thuộc `diabetes_012`. Khi đó mô hình được chọn sẽ có 12 biến bao gồm các biến `high_bp`, `high_chol`, `chol_check`, `bmi`, `stroke`, `heart_diseaseor_attack`, `hvy_alcohol_consump`, `gen_hlth`, `diff_walk`, `sex`, `age`, `income`.

Vậy ta có được mô hình hồi quy logistic khác như sau:

$$\log \frac{p(x)}{1-p(x)} = \beta_0 + \beta_1 \text{high_bp} + \beta_2 \text{high_chol} + \beta_3 \text{chol_check} + \beta_4 \text{bmi} \\ + \beta_5 \text{stroke} + \beta_6 \text{hvy_alcohol_consump} + \beta_7 \text{gen_hlth} + \beta_8 \text{age} \\ + \beta_9 \text{income} + \beta_{10} \text{sex} + \beta_{11} \text{diff_walk} + \beta_{12} \text{heart_diseaseor_attack}$$

##	GVIF	Df	GVIF^(1/(2*Df))
## high_bp	1.127393	1	1.061788
## high_chol	1.066582	1	1.032755
## chol_check	1.006030	1	1.003011
## bmi	1.120119	1	1.058357
## stroke	1.069720	1	1.034273
## hvy_alcohol_consump	1.006626	1	1.003307
## gen_hlth	1.431794	4	1.045888
## age	1.272165	12	1.010080
## income	1.272739	7	1.017376
## sex	1.067925	1	1.033405
## diff_walk	1.371124	1	1.170950
## heart_diseaseor_attack	1.148512	1	1.071686

Thông qua quá trình kiểm tra đa cộng tuyến, vì các VIF_j , $j = \overline{1, 12}$ đều nhỏ hơn 5 nên ta kết luận rằng không có hiện tượng đa cộng tuyến trong mô hình.

Kết quả ước lượng hệ số như sau:

```
##
## Call:
## glm(formula = diabetes_012 ~ high_bp + high_chol + chol_check +
##     bmi + stroke + hvy_alcohol_consump + gen_hlth + age + income +
##     sex + diff_walk + heart_diseaseor_attack, family = "binomial",
##     data = df_clean)
##
## Coefficients:
##
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -7.7138489    0.1275845  -60.461  < 2e-16 ***
## high_bpHigh Blood    0.6593079    0.0138824   47.492  < 2e-16 ***
## high_cholHigh Cholesterol  0.5413982    0.0129597   41.775  < 2e-16 ***
## chol_checkYes    1.2183692    0.0612906   19.879  < 2e-16 ***
## bmi    0.0570291    0.0008798   64.818  < 2e-16 ***
## strokeStroke    0.1226989    0.0242974    5.050 4.42e-07 ***
## hvy_alcohol_consumpYes  -0.7328574    0.0348094  -21.053  < 2e-16 ***
## gen_hlthVery Good    0.6047187    0.0303597   19.918  < 2e-16 ***
## gen_hlthGood    1.1946047    0.0295302   40.454  < 2e-16 ***
## gen_hlthFair    1.6063816    0.0316047   50.827  < 2e-16 ***
## gen_hlthPoor    1.7223768    0.0365326   47.146  < 2e-16 ***
## age2    0.2509953    0.1280930    1.959  0.05006 .
## age3    0.4574934    0.1169260    3.913 9.13e-05 ***
## age4    0.9004995    0.1109017    8.120 4.67e-16 ***
## age5    1.0996239    0.1086542   10.120  < 2e-16 ***
## age6    1.3446650    0.1069236   12.576  < 2e-16 ***
## age7    1.5315927    0.1058207   14.473  < 2e-16 ***
```

```
## age8          1.6226474  0.1054261  15.391 < 2e-16 ***
## age9          1.8378613  0.1051727  17.475 < 2e-16 ***
## age10         1.9909285  0.1051375  18.936 < 2e-16 ***
## age11         2.0457903  0.1054756  19.396 < 2e-16 ***
## age12         1.9736194  0.1061200  18.598 < 2e-16 ***
## age13         1.8044836  0.1062175  16.989 < 2e-16 ***
## income2       -0.0431277  0.0340713  -1.266  0.20558
## income3       -0.0873858  0.0327431  -2.669  0.00761 **
## income4       -0.1327741  0.0319551  -4.155 3.25e-05 ***
## income5       -0.2046164  0.0313176  -6.534 6.42e-11 ***
## income6       -0.2855286  0.0305667  -9.341 < 2e-16 ***
## income7       -0.3190698  0.0306259 -10.418 < 2e-16 ***
## income8       -0.4195071  0.0297387 -14.106 < 2e-16 ***
## sexMale        0.2457466  0.0126612  19.409 < 2e-16 ***
## diff_walkYes   0.1124369  0.0156074   7.204 5.84e-13 ***
## heart_diseaseor_attackYes 0.2147496  0.0172447  12.453 < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##      Null deviance: 211598  on 229780  degrees of freedom
## Residual deviance: 170734  on 229748  degrees of freedom
## AIC: 170800
##
## Number of Fisher Scoring iterations: 6
```

Từ kết quả trên, ta có thể rút ra một số kết luận:

- Hệ số dương của biến **high_bpHigh Blood** cho thấy những người bị huyết áp cao có khả năng mắc bệnh tiểu đường cao hơn so với những người không bị huyết áp cao. Việc tăng một đơn vị của biến **high_bp** sẽ làm tăng tỷ lệ mắc bệnh tiểu đường lên $e^{0.659} \approx 1.933$ lần.
- Hệ số dương của biến **high_cholHigh Cholesterol** cho thấy những người bị cholesterol cao có khả năng mắc bệnh tiểu đường cao hơn so với những người không bị cholesterol cao.
- Hệ số dương đối với **bmi** cho thấy rằng việc tăng chỉ số BMI có liên quan đến việc tăng nguy cơ mắc bệnh tiểu đường. Việc tăng đơn vị BMI có liên quan đến việc tăng tỷ lệ mắc bệnh tiểu đường là $e^{0.057} \approx 1.06$ lần.
- Hệ số âm đối với **strokeStroke** cho thấy rằng người từng bị đột quỵ trong quá khứ có khả năng mắc bệnh tiểu đường ít hơn so với những người chưa từng bị đột quỵ.
- Hệ số âm đối với biến **hvy_alcohol_consump** cho thấy rằng người tiêu thụ rượu nặng có nguy cơ mắc bệnh tiểu đường thấp hơn so với người không tiêu thụ rượu nặng. Tuy nhiên, điều này không phải là một kết luận chắc chắn vì

tiêu thụ rượu nặng quá nhiều cũng không tốt cho sức khỏe.

- Từ hệ số dương cho `gen_hlth` Very Good, `gen_hlth` Good, `gen_hlth` Fair, `gen_hlth` Poor càng tăng dần khi chỉ số từ Very Good cho tới Poor ta thấy rằng người có sức khỏe tốt, khá, trung bình hoặc kém hơn so với người có sức khỏe xuất sắc có khả năng mắc bệnh tiểu đường cao hơn.
- Với hệ số âm cho các biến từ `income2` cho tới `income6` càng giảm dần thì cho ta thấy được là một người có thu nhập cao thì càng ít có khả năng mắc bệnh tiểu đường hơn vì có các dịch vụ y tế, chăm sóc sức khỏe tốt hơn.
- Với hệ số dương cho biến `age` càng tăng dần theo độ tuổi cho ta biết được là người càng cao tuổi càng có nguy cơ bị mắc bệnh tiểu đường.
- Hệ số dương của biến `heart_diseaseor_attack` Yes cho thấy rằng những người đã từng mắc bệnh tim hoặc đau tim thì sẽ dễ mắc bệnh tiểu đường hơn là những người không mắc bệnh tim hoặc đau tim. Nói đúng hơn là, nếu tăng một đơn vị về biến `heart_diseaseor_attack` thì sẽ làm tăng tỉ lệ mắc bệnh tiểu đường là $e^{0.214} \approx 1.24$ lần.
- Việc tăng một đơn vị của biến `sex` sẽ làm tăng tỷ lệ mắc bệnh tiểu đường lên $e^{0.112} \approx 1.12$ lần.
- Hệ số dương của biến `diff_walk` cho thấy rằng người gặp khó khăn khi leo cầu thang có khả năng mắc bệnh tiểu đường cao hơn so với người không gặp khó khăn khi leo cầu thang.
- Tất cả các biến trong mô hình đều có $p_{value} < 0.05$ nên chúng đều có ý nghĩa thống kê với mô hình.

Bây giờ, giả sử ta có thông tin về 5 người và chúng ta muốn dự đoán xác suất họ mắc bệnh tiểu đường. Thông tin được đưa ra như sau:

id	high_bp	high_chol	chol_check	bmi	stroke	hvy_alcohol_consump	gen_hlth	age	income	sex	diff_walk	heart_diseaseor_attack
1	High Blood	High Cholesterol	Yes	41	No Stroke	No	Fair	9	8	Male	No	No
2	No High Blood	No High Cholesterol	Yes	24	No Stroke	No	Very Good	3	7	Female	No	No
3	High Blood	High Cholesterol	Yes	26	No Stroke	No	Good	13	6	Male	No	No
4	No High Blood	High Cholesterol	Yes	30	No Stroke	No	Very Good	7	7	Male	No	No
5	High Blood	High Cholesterol	Yes	22	No Stroke	No	Poor	13	2	Female	Yes	Yes

Ta sẽ phân loại 5 người trên thành hai nhóm là: **Không mắc bệnh tiểu đường** và **Mắc bệnh tiểu đường** bằng cách so sánh xác suất ở ngưỡng $c = 0.5$.

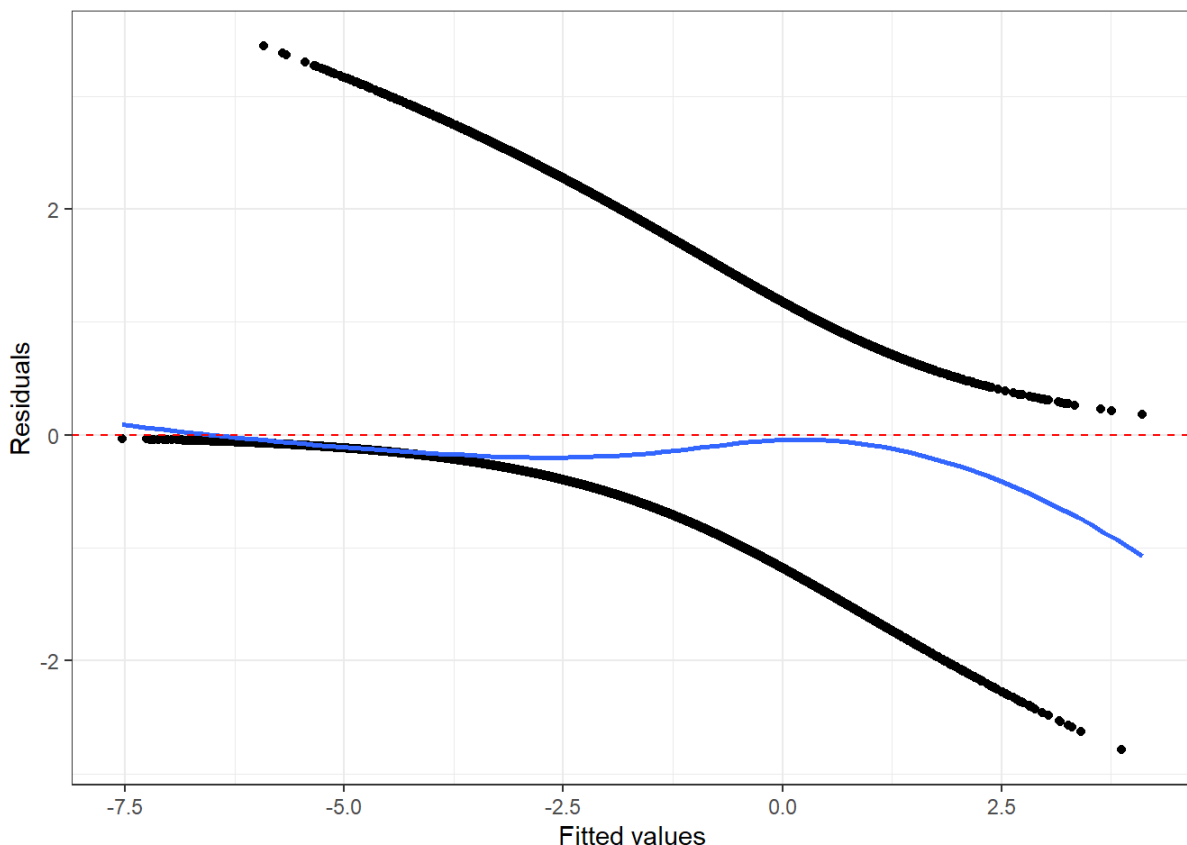
Kết quả thu được như sau:

```
##          1          2          3          4          5
##  "Diabetes" "No Diabetes" "No Diabetes" "No Diabetes" "No Diabetes"
```

f) Chuẩn đoán mô hình:

Tiếp theo, ta sẽ thực hiện chuẩn đoán mô hình đã lựa chọn các cặp biến tốt nhất ở trên (mô hình `logistic_md_select`).

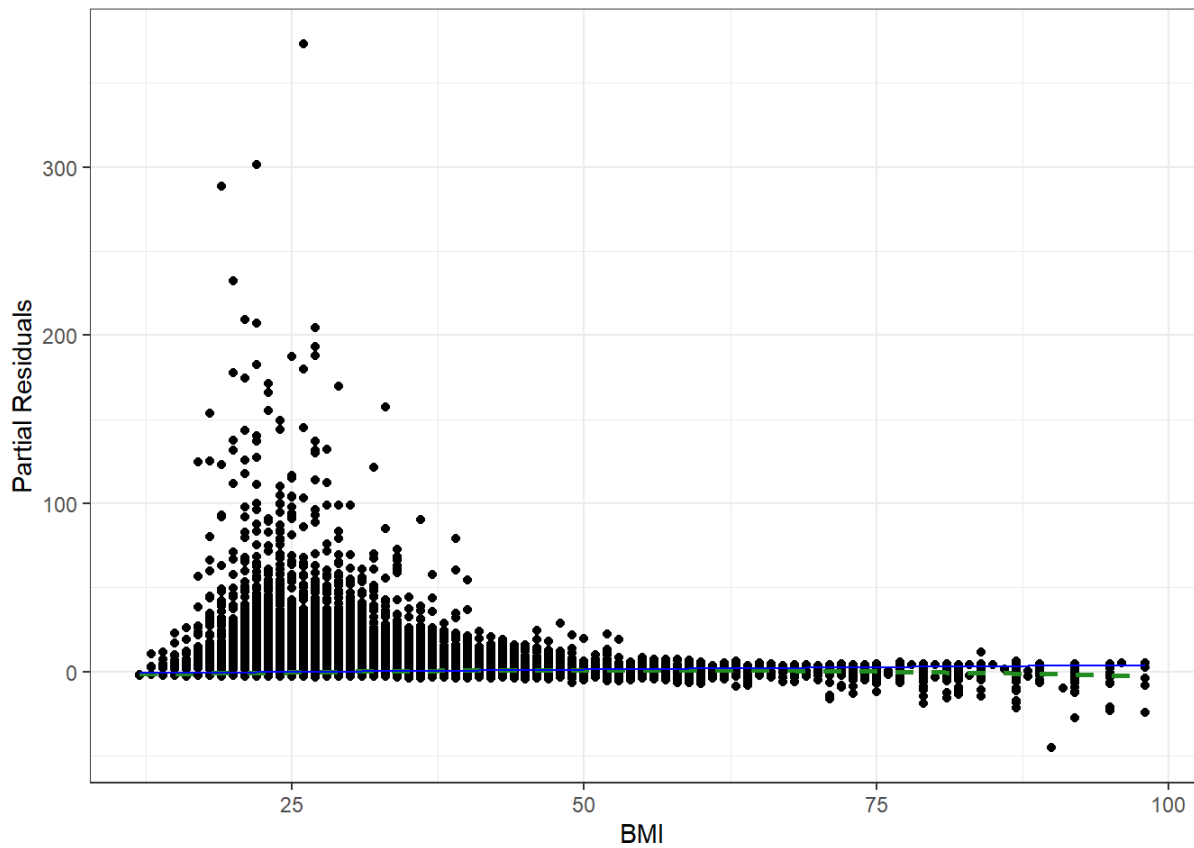
- Kiểm tra tính tuyến tính của mô hình:



Nhận xét: Hình vẽ cho thấy xu hướng đường cong là tương đối. Ta kết luận giả định về tính tuyến tính của mô hình là không phù hợp.

- Kiểm tra tính tuyến tính từng phần:

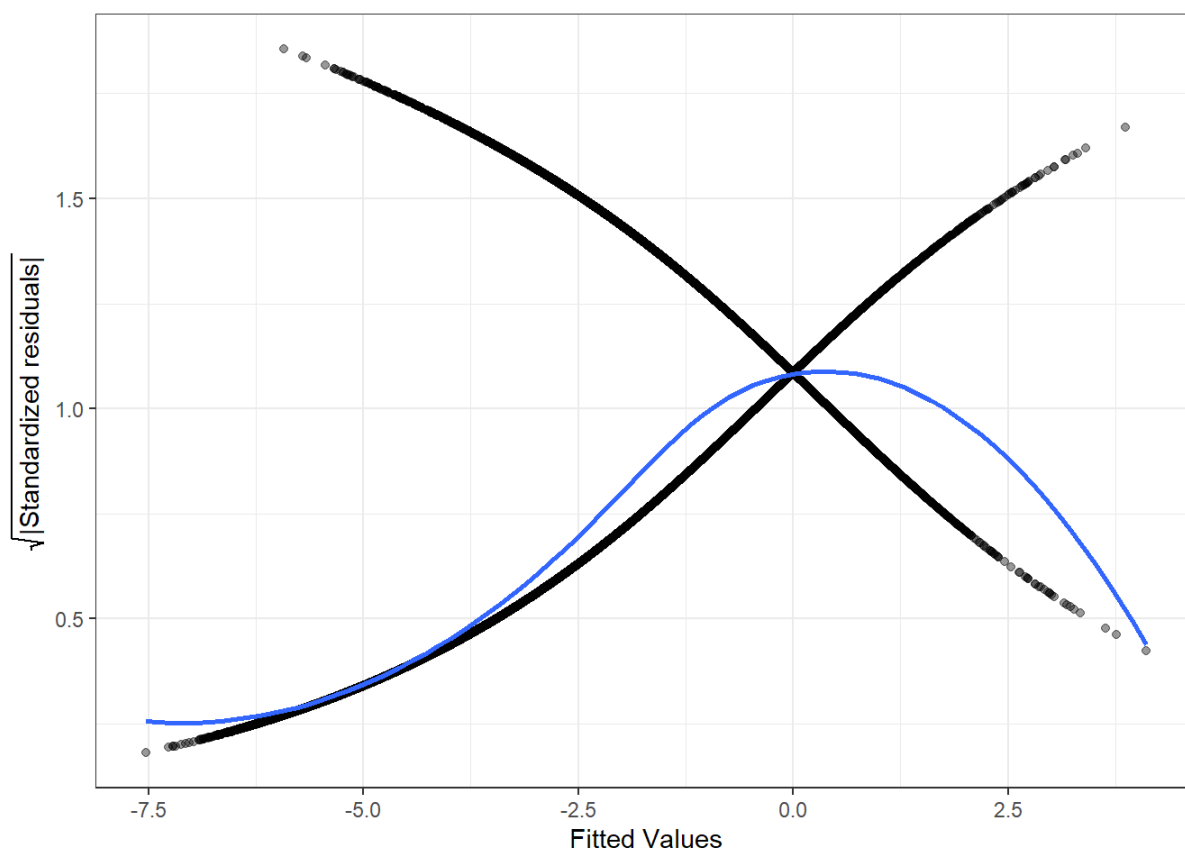
Vì chỉ có biến `bmi` là biến định lượng trong khi các biến còn lại của mô hình là định tính nên ta chỉ kiểm tra tính tuyến tính từng phần cho biến `bmi`.



Nhận xét: Kết quả cho thấy đường thẳng tuyến tính (màu xanh dương) ước lượng tương đối khớp với dữ liệu. Ta kết luận có sự tuyến tính từng phần giữa biến `bmi` và biến phụ thuộc.

Vậy thì có thể là do có nhiều biến định tính nên mô hình bị ảnh hưởng dẫn tới sự không tuyến tính trong mô hình.

- Kiểm tra tính đồng nhất phương sai:

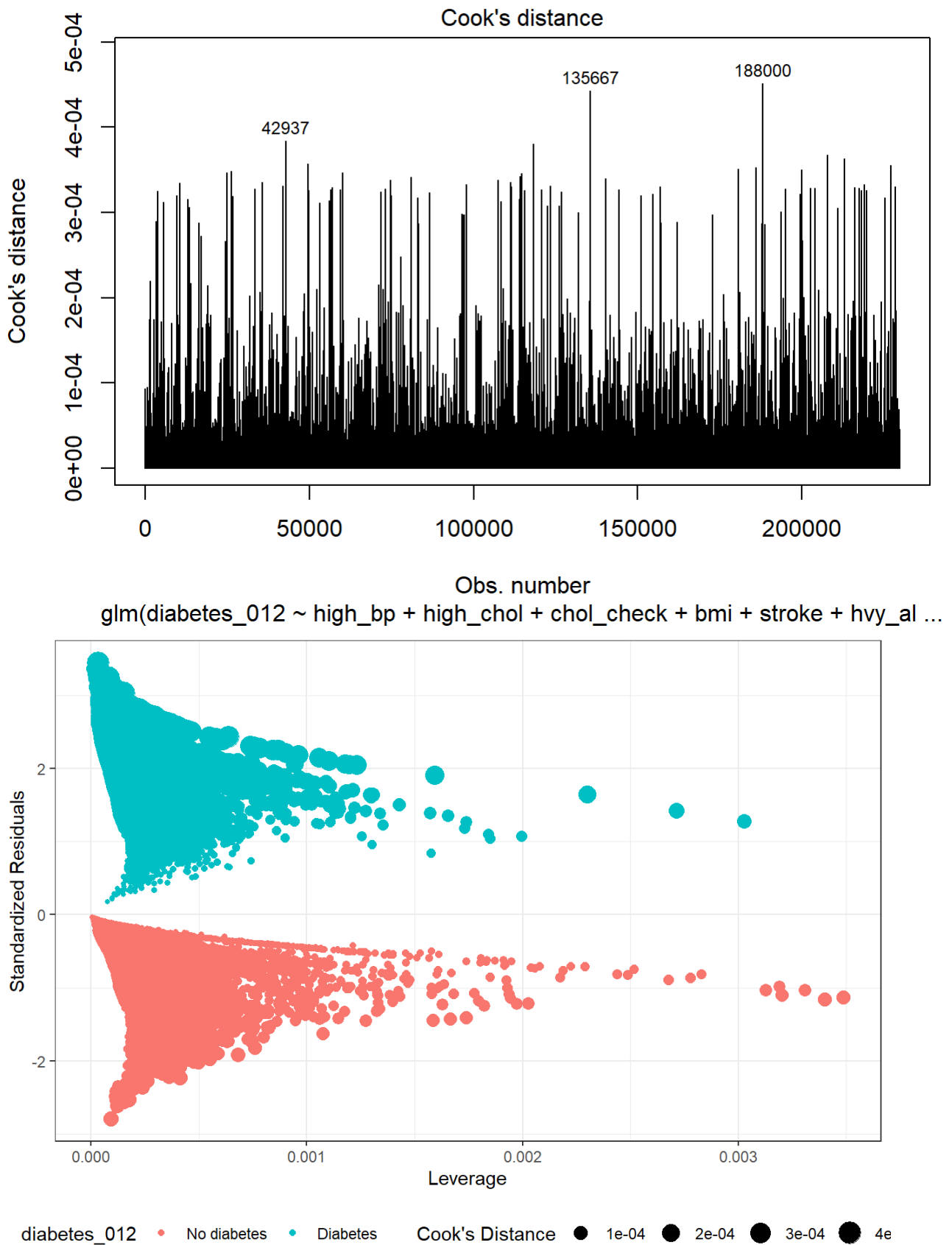


Nhận xét:

- Ở hai đầu của trục fitted values (xấp xỉ -7.5 và +2.5), residuals có giá trị lớn hơn đáng kể so với phần giữa. Điều này chỉ ra rằng phương sai không đồng nhất, đặc biệt khi fitted values xa trung tâm.
- Đường xu hướng dữ liệu (màu xanh dương) không phải là một đường ngang và cho thấy sự thay đổi theo giá trị fitted values.

Ta kết luận mô hình không đáp ứng được giả định về tính đồng nhất phương sai. Có thể là do có nhiều biến định tính nên mô hình bị ảnh hưởng dẫn tới sự không đồng nhất phương sai trong mô hình.

- Kiểm tra điểm ngoại lai trong mô hình:



Nhận xét: Vì các Cook's Distance của từng điểm dữ liệu trên đều nhỏ hơn 0.5 nên không có giá trị ngoại lai nào cần loại bỏ.

g) Mở rộng mô hình:

Vì không thể kiểm tra tính tuyến tính cho các biến định tính trong mô hình được lựa chọn ở trên nên ta sử dụng mô hình Cộng tính tổng quát (Generalize Additive Moldels) hay còn gọi là GAM.

Kết quả tổng hợp như sau:

```
##
## Family: binomial
## Link function: logit
##
## Formula:
## diabetes_012 ~ high_bp + high_chol + chol_check + s(bmi) + stroke +
##      hvy_alcohol_consump + gen_hlth + age + income + sex + diff_walk +
##      heart_diseaseor_attack
##
## Parametric coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -5.92946    0.12439  -47.670  < 2e-16 ***
## high_bpHigh Blood      0.61189    0.01396   43.842  < 2e-16 ***
## high_cholHigh Cholesterol 0.52951    0.01301   40.693  < 2e-16 ***
## chol_checkYes        1.19908    0.06124   19.581  < 2e-16 ***
## strokeStroke         0.13935    0.02450    5.688 1.29e-08 ***
## hvy_alcohol_consumpYes -0.70375    0.03489  -20.169  < 2e-16 ***
## gen_hlthVery Good      0.56013    0.03038   18.439  < 2e-16 ***
## gen_hlthGood          1.12922    0.02957   38.187  < 2e-16 ***
## gen_hlthFair          1.55594    0.03165   49.162  < 2e-16 ***
## gen_hlthPoor          1.70198    0.03665   46.444  < 2e-16 ***
## age2                 0.22322    0.12720    1.755 0.079290 .
## age3                 0.42366    0.11622    3.645 0.000267 ***
## age4                 0.81610    0.11034    7.396 1.40e-13 ***
## age5                 1.02128    0.10807    9.450  < 2e-16 ***
## age6                 1.26444    0.10636   11.889  < 2e-16 ***
## age7                 1.45374    0.10526   13.811  < 2e-16 ***
```

```
## age8          1.55269      0.10486    14.807 < 2e-16 ***
## age9          1.77089      0.10461    16.928 < 2e-16 ***
## age10         1.92899      0.10457    18.446 < 2e-16 ***
## age11         1.99557      0.10492    19.020 < 2e-16 ***
## age12         1.94519      0.10557    18.425 < 2e-16 ***
## age13         1.82349      0.10566    17.258 < 2e-16 ***
## income2       -0.05018      0.03425    -1.465 0.142910
## income3       -0.10364      0.03293    -3.148 0.001647 **
## income4       -0.15056      0.03212    -4.688 2.76e-06 ***
## income5       -0.22731      0.03148    -7.220 5.19e-13 ***
## income6       -0.31350      0.03072   -10.204 < 2e-16 ***
## income7       -0.34586      0.03077   -11.239 < 2e-16 ***
## income8       -0.44280      0.02989   -14.816 < 2e-16 ***
## sexMale        0.22666      0.01280    17.709 < 2e-16 ***
## diff_walkYes   0.08699      0.01572     5.535 3.12e-08 ***
## heart_diseaseor_attackYes 0.21459      0.01737    12.352 < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##           edf Ref.df Chi.sq p-value
## s(bmi) 6.913  7.727   5427  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.187   Deviance explained =   20%
## UBRE = -0.26308   Scale est. = 1           n = 229781
```

Một số nhận xét:

- Hệ số dương của biến `high_bp` High Blood cho thấy những người bị huyết áp cao có khả năng mắc bệnh tiểu đường cao hơn so với những người không bị huyết áp cao. Việc tăng một đơn vị của biến `high_bp` sẽ làm tăng tỷ lệ mắc bệnh tiểu đường lên $e^{0.61189} \approx 1.843$ lần.
- Hệ số dương của biến `high_chol` High Cholesterol cho thấy những người bị cholesterol cao có khả năng mắc bệnh tiểu đường cao hơn so với những người không bị cholesterol cao.
- Hệ số âm của biến `hvy_alcohol_consump` Yes cho thấy rằng người tiêu thụ rượu nặng có nguy cơ mắc bệnh tiểu đường thấp hơn so với người không tiêu thụ rượu nặng. Tuy nhiên, đây không phải là kết luận chắc chắn, vì tiêu thụ rượu nặng quá mức cũng có thể gây ảnh hưởng tiêu cực đến sức khỏe tổng

thể.

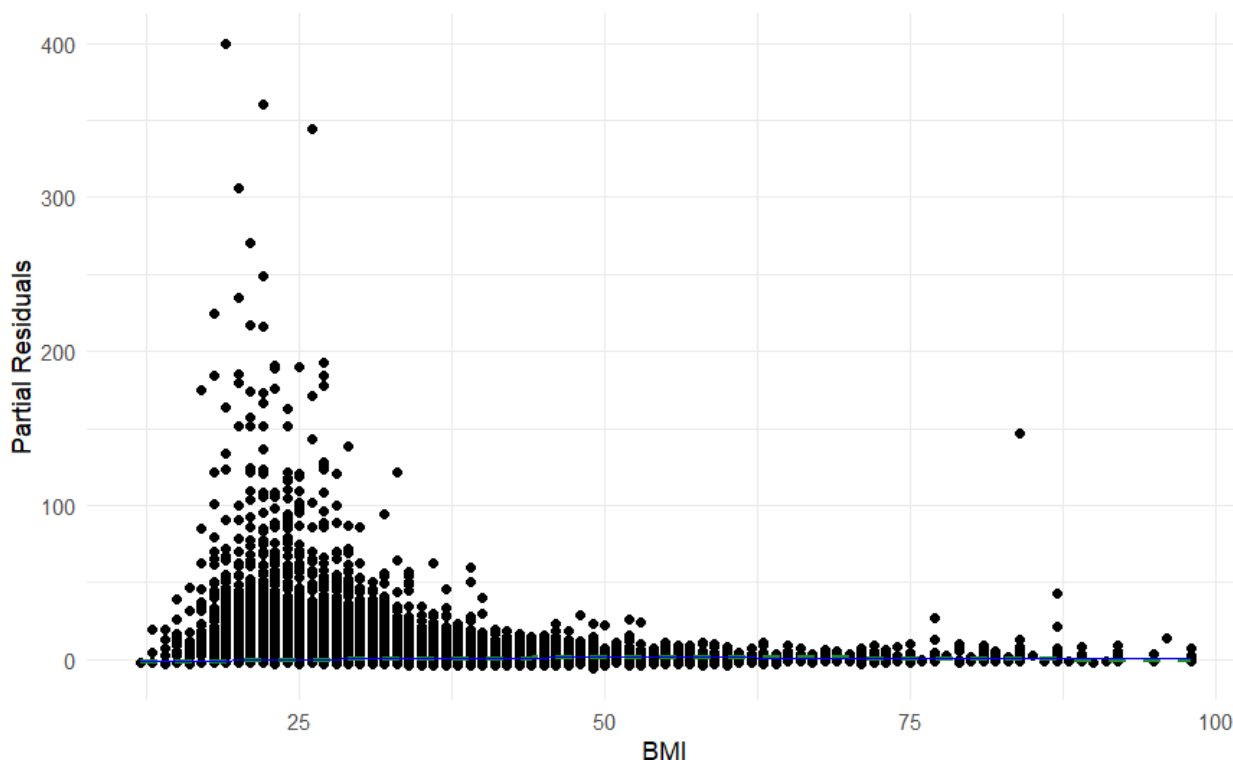
- Hệ số dương của các mức biến `gen_hlth` (Very Good, Good, Fair, Poor) cho thấy rằng nguy cơ mắc bệnh tiểu đường tăng dần theo sự giảm sút của sức khỏe. Những người có sức khỏe Fair hoặc Poor có nguy cơ mắc bệnh tiểu đường cao hơn so với những người có sức khỏe Excellent.
- Với hệ số âm của các mức biến `income` (từ `income2` đến `income8`), ta nhận thấy rằng người có thu nhập càng cao thì nguy cơ mắc bệnh tiểu đường càng thấp. Điều này có thể do khả năng tiếp cận dịch vụ y tế tốt hơn hoặc duy trì lối sống lành mạnh hơn.
- Hệ số dương của các mức biến `age` (từ `age2` đến `age13`) cho thấy rằng nguy cơ mắc bệnh tiểu đường tăng dần theo độ tuổi. Cụ thể, người ở nhóm tuổi cao nhất (`age13`) có nguy cơ mắc bệnh tiểu đường cao hơn đáng kể so với nhóm tuổi trẻ nhất.
- Hệ số dương của biến `sexMale` cho thấy rằng nam giới có nguy cơ mắc bệnh tiểu đường cao hơn so với nữ giới. Cụ thể, việc tăng một đơn vị của biến này sẽ làm tăng tỷ lệ mắc bệnh tiểu đường lên $e^{0.22666} \approx 1.254$ lần.
- Hệ số dương của biến `diff_walkYes` chỉ ra rằng người gặp khó khăn khi leo cầu thang có nguy cơ mắc bệnh tiểu đường cao hơn người không gặp khó khăn.
- Hệ số dương của biến `heart_diseaseor_attackYes` cho thấy rằng người từng mắc bệnh tim hoặc đau tim có nguy cơ mắc bệnh tiểu đường cao hơn. Cụ thể, việc tăng một đơn vị của biến này sẽ làm tăng tỷ lệ mắc bệnh tiểu đường lên $e^{0.21459} \approx 1.239$ lần.

Ta sẽ sử dụng mô hình này để phân loại dữ liệu về 5 người ở bước **Lựa chọn mô hình** phía trên thành hai nhóm là: Không mắc bệnh tiểu đường và Mắc bệnh tiểu đường bằng cách so sánh xác suất ở ngưỡng $c = 0.5$.

Kết quả thu được như sau:

##	1	2	3	4	5
##	"Diabetes"	"No Diabetes"	"No Diabetes"	"No Diabetes"	"No Diabetes"

Sau đó ta sẽ xác định thặng dư từng phần vẽ biểu đồ để kiểm tra tính tuyến tính của `bmi` và biến phụ thuộc.



Nhận xét: Đường thẳng tuyến tính (màu xanh dương) trùng với đường thẳng xu hướng của dữ liệu (màu xanh lá cây). Ta kết luận giả định về tính tuyến tính của bmi là phù hợp.

h) Đánh giá mô hình:

Đầu tiên ta sẽ chia dữ liệu thành 2 tập huấn luyện (train: 70% dữ liệu) và kiểm tra (test: 30% dữ liệu). Sau đó ta thực hiện xử lý mất cân bằng dữ liệu trên tập huấn luyện.

##

No diabetes Diabetes

132995 27851

Under-sampling

SMOTE

##

##

No diabetes Diabetes

No diabetes Diabetes

27851 27851

132995 132995

Ta bước vào quá trình đánh giá mô hình, với các bước sau:

- Xây dựng mô hình với tập huấn luyện sau khi thực hiện 2 phương pháp Under-sampling và SMOTE.

Under-sampling:

```
##
## Call:
## glm(formula = diabetes_012 ~ high_bp + high_chol + chol_check +
##      bmi + stroke + hvy_alcohol_consump + gen_hlth + age + income +
##      sex + diff_walk + heart_diseaseor_attack, family = "binomial",
##      data = train_df_under)
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)      -6.32364    0.16871  -37.481 < 2e-16 ***
## high_bpHigh Blood    0.61856    0.02164   28.588 < 2e-16 ***
## high_cholHigh Cholesterol 0.52857    0.02072   25.513 < 2e-16 ***
## chol_checkYes       1.30357    0.08580   15.193 < 2e-16 ***
## bmi                0.06645    0.00168   39.553 < 2e-16 ***
## strokeStroke        0.12549    0.04350    2.885 0.003918 **
## hvy_alcohol_consumpYes -0.73093    0.05094  -14.349 < 2e-16 ***
## gen_hlthVery Good    0.58382    0.04132   14.129 < 2e-16 ***
## gen_hlthGood         1.17637    0.04060   28.975 < 2e-16 ***
## gen_hlthFair         1.59484    0.04545   35.091 < 2e-16 ***
## gen_hlthPoor         1.73633    0.05738   30.260 < 2e-16 ***
## age2                0.24088    0.16108    1.495 0.134812
## age3                0.27558    0.14795    1.863 0.062510 .
## age4                0.76942    0.13991    5.499 3.81e-08 ***
## age5                0.92107    0.13685    6.731 1.69e-11 ***
## age6                1.16673    0.13428    8.689 < 2e-16 ***
## age7                1.40932    0.13229   10.653 < 2e-16 ***
## age8                1.50844    0.13161   11.462 < 2e-16 ***
## age9                1.68500    0.13124   12.839 < 2e-16 ***
## age10               1.85622    0.13120   14.148 < 2e-16 ***
## age11               1.95196    0.13196   14.792 < 2e-16 ***
## age12               1.84592    0.13324   13.854 < 2e-16 ***
## age13               1.73581    0.13324   13.028 < 2e-16 ***
## income2             -0.03275    0.06014   -0.545 0.586011
## income3             -0.13728    0.05691   -2.412 0.015866 *
## income4             -0.18429    0.05525   -3.335 0.000852 ***
## income5             -0.22429    0.05385   -4.165 3.11e-05 ***
## income6             -0.33091    0.05220   -6.339 2.31e-10 ***
## income7             -0.38076    0.05195   -7.330 2.31e-13 ***
## income8             -0.49329    0.05035   -9.796 < 2e-16 ***
## sexMale              0.28549    0.02063   13.841 < 2e-16 ***
## diff_walkYes         0.12856    0.02694    4.772 1.82e-06 ***
## heart_diseaseor_attackYes 0.24288    0.03085    7.874 3.45e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##      Null deviance: 77219  on 55701  degrees of freedom
## Residual deviance: 59840  on 55669  degrees of freedom
## AIC: 59906
##
## Number of Fisher Scoring iterations: 5
```

SMOTE:

```
##
## Call:
## glm(formula = diabetes_012 ~ high_bp + high_chol + chol_check +
##      bmi + stroke + hvy_alcohol_consump + gen_hlth + age + income +
##      sex + diff_walk + heart_diseaseor_attack, family = "binomial",
##      data = train_df_smote)
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)      -9.0237494   0.1152282  -78.312 < 2e-16 ***
## high_bpHigh Blood    0.6494501   0.0102162   63.571 < 2e-16 ***
## high_cholHigh Cholesterol 0.5311677   0.0097169   54.664 < 2e-16 ***
## chol_checkYes       2.4563582   0.0652755   37.631 < 2e-16 ***
## bmi                0.0716242   0.0008329   85.990 < 2e-16 ***
## strokeStroke       -0.2614840   0.0215612  -12.128 < 2e-16 ***
## hvy_alcohol_consumpYes -1.6306329   0.0318433  -51.208 < 2e-16 ***
## gen_hlthVery Good    0.9253951   0.0213900   43.263 < 2e-16 ***
## gen_hlthGood        1.5886476   0.0211236   75.207 < 2e-16 ***
## gen_hlthFair        2.0164356   0.0234584   85.958 < 2e-16 ***
## gen_hlthPoor        2.1398415   0.0288640   74.135 < 2e-16 ***
## age2                0.4293251   0.1058546    4.056 5.00e-05 ***
## age3                0.5889175   0.0977711    6.023 1.71e-09 ***
## age4                1.3044329   0.0924751   14.106 < 2e-16 ***
## age5                1.5028866   0.0910329   16.509 < 2e-16 ***
## age6                1.8642200   0.0898001   20.760 < 2e-16 ***
## age7                2.2666766   0.0889785   25.474 < 2e-16 ***
```

```
## age8          2.4178030  0.0887256  27.250 < 2e-16 ***
## age9          2.6840585  0.0885980  30.295 < 2e-16 ***
## age10         2.8954527  0.0885988  32.681 < 2e-16 ***
## age11         2.9446420  0.0888523  33.141 < 2e-16 ***
## age12         2.8589226  0.0893180  32.008 < 2e-16 ***
## age13         2.7295987  0.0893119  30.563 < 2e-16 ***
## income2       -0.0373888  0.0288812  -1.295 0.195468
## income3       -0.0689578  0.0275368  -2.504 0.012273 *
## income4       -0.0695078  0.0266611  -2.607 0.009132 **
## income5       -0.1189334  0.0260086  -4.573 4.81e-06 ***
## income6       -0.1650268  0.0253131  -6.519 7.06e-11 ***
## income7       -0.1622009  0.0252460  -6.425 1.32e-10 ***
## income8       -0.1796382  0.0245478  -7.318 2.52e-13 ***
## sexMale        0.2639824  0.0097167  27.168 < 2e-16 ***
## diff_walkYes   0.0071611  0.0129396   0.553 0.579974
## heart_diseaseor_attackYes 0.0560027  0.0146545   3.822 0.000133 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##    Null deviance: 368740  on 265989  degrees of freedom
## Residual deviance: 274397  on 265957  degrees of freedom
## AIC: 274463
##
## Number of Fisher Scoring iterations: 6
```

- Dự đoán trên tập kiểm tra:

Under-sampling:

```
##          logit_2_under_pred_class
##          No diabetes Diabetes
## No diabetes          40284      16776
## Diabetes             2916       8959
```

```
## $Accuaracy
## [1] 0.7143396
##
## $Precision
## No diabetes      Diabetes
##    0.9325000      0.3481251
##
## $Recall
## No diabetes      Diabetes
##    0.7059937      0.7544421
##
## $Macro_F1
## [1] 0.3190291
##
## $Kappa
## [1] 0.314908
```

Như vậy, ta có được các nhận định sau:

1. **Accuracy:** 71.4%

Mặc dù đây là một chỉ số phổ biến, nhưng khi dữ liệu đã được cân bằng, độ chính xác không còn là tiêu chí đáng tin cậy để đánh giá mô hình. Mức 71.4% cho thấy mô hình đạt độ chính xác khá ổn, nhưng cần kết hợp với Precision, Recall, và F1-Score để đánh giá toàn diện.

2. **Precision**

- No diabetes: 93.25%

Mô hình dự đoán lớp No diabetes khá chính xác. Trong số các mẫu được dự đoán là No diabetes có đến 93.25% là đúng.

- Diabetes: 34.81%

Với lớp Diabetes Precision thấp cho thấy mô hình có nhiều dự đoán sai dương tính (false positives) tức là nhiều mẫu bị dự đoán nhầm thành Diabetes.

3. **Recall (Sensitivity)**

- No diabetes: 70.6%

Mô hình nhận diện được phần lớn các mẫu thuộc lớp No diabetes nhưng vẫn bỏ sót khoảng 29.4% (False Negative).

- Diabetes: 75.44%
Mô hình nhận diện tốt hơn các mẫu thuộc lớp Diabetes với tỷ lệ bỏ sót là 24.56%
- 4. Macro F1-Score: 31.90%
Giá trị thấp (31.9%) cho thấy hiệu suất của mô hình với lớp Diabetes chưa đạt được kỳ vọng.
- 5. Kappa: 31.49%
Hệ số Kappa thấp phản ánh rằng mô hình chỉ hiệu quả hơn một chút so với việc đoán ngẫu nhiên.

SMOTE:

```
##                logit_2_smote_pred_class
##                No diabetes Diabetes
## No diabetes      40116      16944
## Diabetes         2954       8921
## $Accuracy
## [1] 0.7113513
##
## $Precision
## No diabetes      Diabetes
## 0.9314140      0.3449062
##
## $Recall
## No diabetes      Diabetes
## 0.7030494      0.7512421
##
## $Macro_F1
## [1] 0.3154295
##
## $Kappa
## [1] 0.3097874
```

Như vậy, ta có được các nhận định sau:

1. **Accuracy:** 71.14%
Mức độ chính xác tổng thể của mô hình, nhưng không phản ánh toàn diện

hiệu suất trong bài toán mất cân bằng.

2. Precision

- No diabetes: 93.14%
Dự đoán rất chính xác.
- Diabetes: 34.49%
Dự đoán còn nhiều nhầm lẫn.

3. Recall

- No diabetes: 70.30%
Phát hiện được phần lớn các trường hợp “No diabetes”.
- Diabetes: 75.12%
Ít bỏ sót các trường hợp mắc bệnh.

4. Macro F1-Score: 31.54%
Hiệu suất tổng thể của mô hình còn thấp.

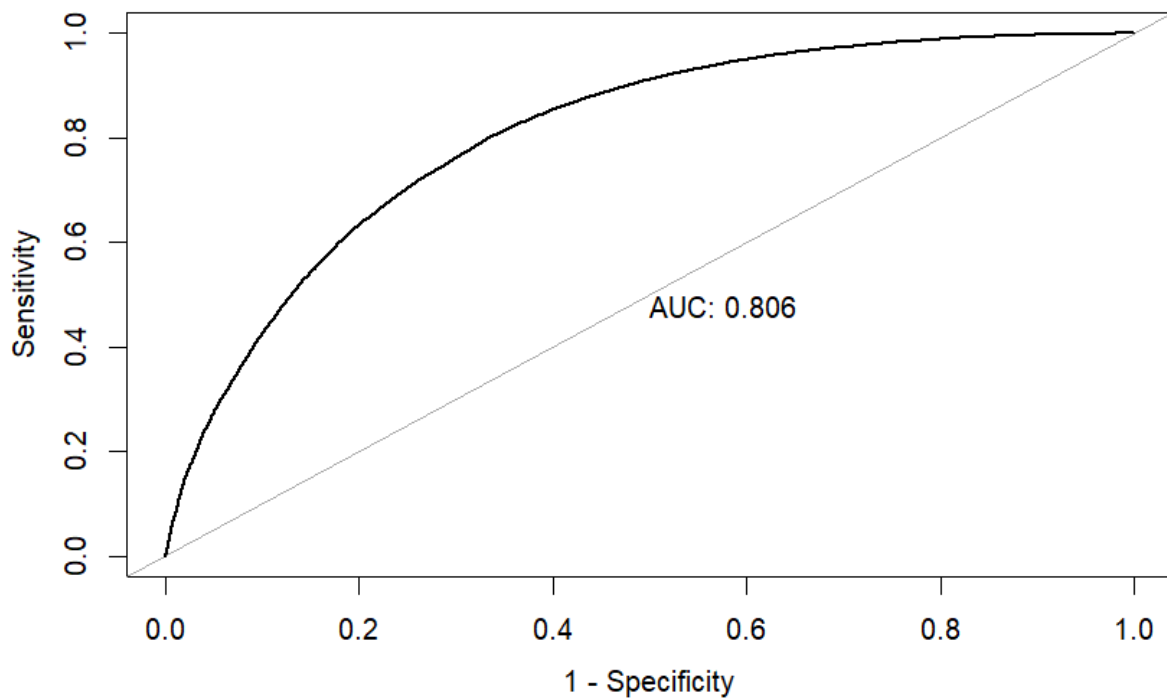
5. Kappa: 30.98%
Dự đoán của mô hình chỉ tốt hơn ngẫu nhiên một chút, mặc dù đã áp dụng cân bằng dữ liệu bằng SMOTE.

- Đánh giá dựa trên ROC và AUC:

Under-sampling:

Thông qua một số phân tích, ta nhận định:

- Giá trị AUC đạt 80.58% cho thấy mô hình Logistic Regression sau khi áp dụng under-sampling có khả năng phân biệt khá tốt giữa hai nhóm Diabetes và No diabetes.
- AUC nằm trong khoảng từ 0.8 đến 0.9, được xem là tốt. Điều này phản ánh rằng mô hình có khả năng phân loại tốt trong việc dự đoán xác suất mắc bệnh.



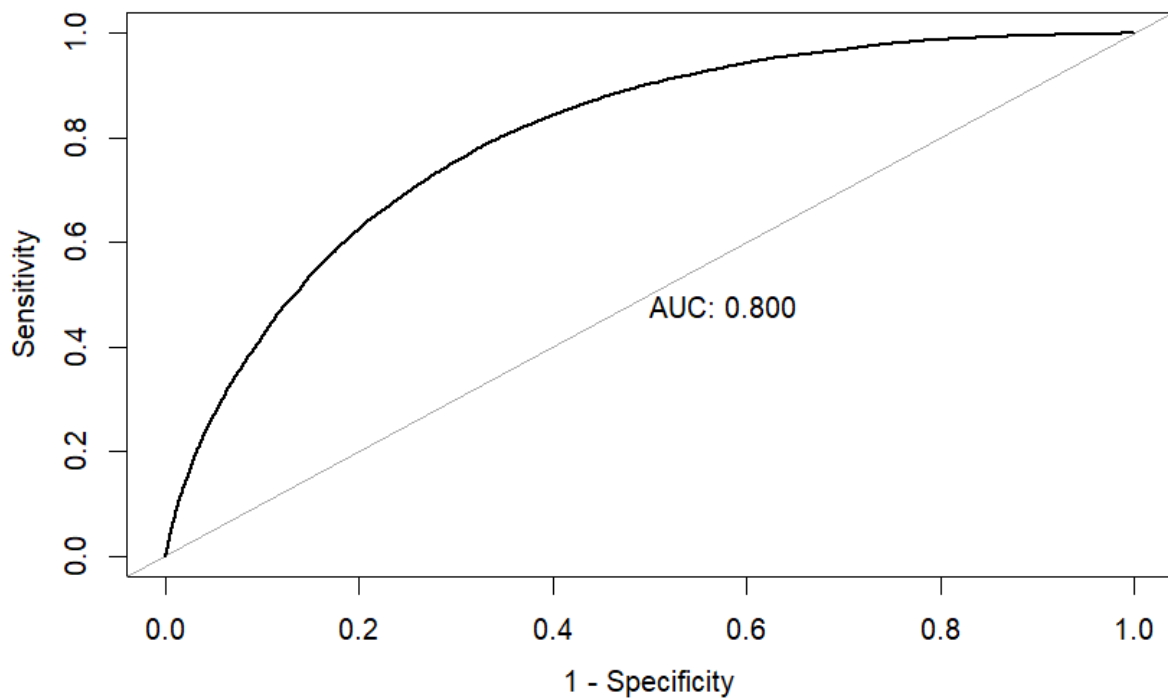
Biểu đồ trên cho ta thấy rằng:

- Một AUC cao như 0.8058 cho thấy mô hình có thể cân bằng tốt giữa Sensitivity và Specificity ở một số ngưỡng tối ưu.
- Với $AUC = 0.8058$, mô hình có khả năng dự đoán tốt hơn ngẫu nhiên ($AUC = 0.5$) một cách đáng kể.

SMOTE:

Thông qua một số phân tích, ta nhận định:

- Giá trị AUC đạt 80.03% cho thấy mô hình Logistic Regression sau khi áp dụng SMOTE có khả năng phân biệt khá tốt giữa hai nhóm Diabetes và No diabetes.
- AUC nằm trong khoảng từ 0.8 đến 0.9, được xem là tốt. Điều này phản ánh rằng mô hình có khả năng phân loại tốt trong việc dự đoán xác suất mắc bệnh.



Qua biểu đồ ta thấy được:

- Một AUC cao như 0.8003 cho thấy mô hình có thể cân bằng tốt giữa Sensitivity và Specificity ở một số ngưỡng tối ưu.
- Với $AUC = 0.8003$, mô hình có khả năng dự đoán tốt hơn ngẫu nhiên ($AUC = 0.5$) một cách đáng kể.

6. Viết các nhận xét và kết luận về các kết quả đã thu được sau quá trình phân tích

Dựa trên phân tích và kết quả, một số kết luận quan trọng có thể rút ra như sau:

1. **Hiệu suất mô hình:** Trong bài toán dự đoán bệnh tiểu đường, mô hình Logistic Regression kết hợp kỹ thuật cân bằng dữ liệu SMOTE và mô hình Logistic Regression sử dụng kỹ thuật Under-sampling đều đạt hiệu suất dự đoán tương đương.
2. **Các yếu tố làm tăng nguy cơ tiểu đường:** Nguy cơ mắc bệnh tiểu đường gia tăng cùng với các yếu tố như huyết áp cao, cholesterol cao, tuổi tác và chỉ số BMI cao.
3. **Các yếu tố giảm nguy cơ tiểu đường:**
 - Những người có thu nhập cao hơn thường có nguy cơ mắc bệnh tiểu đường thấp hơn nhờ điều kiện tiếp cận dịch vụ y tế và chế độ dinh dưỡng tốt hơn.
 - Sức khỏe tổng thể, thể chất và tinh thần tốt cũng là yếu tố quan trọng giúp giảm nguy cơ mắc bệnh.
4. **Giải pháp giảm nguy cơ tiểu đường:**
 - Duy trì cân nặng ổn định, kiểm soát huyết áp và cholesterol.
 - Tăng cường vận động thể chất và giữ tinh thần thoải mái.
 - Hạn chế tiêu thụ thức ăn nhanh, thực phẩm chứa nhiều đường, chất béo và natri.
 - Kiểm tra sức khỏe định kỳ và nâng cao kiến thức về bệnh tiểu đường để nhận thức được các nguy cơ và cách phòng ngừa.