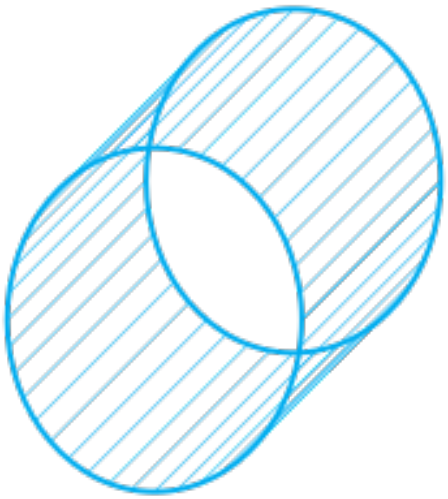
ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH
Trường Đại học Khoa học Tự nhiên
Khoa Toán - Tin học

Môn học

XỬ LÝ SỐ LIỆU THỐNG KÊ

- Đồ án cuối kì -



Project 2

CDC Diabetes Health Indicators

DANH SÁCH SINH VIÊN

1. Trần Gia Huy	22280040
2. Lương Thanh Nam	22280056
3. Mai Thị Kim Ngân	22280058
4. Huỳnh Hà Anh Thư	22280089
5. Lê Thanh Thùy	22280094

Mục lục

1 Bản đề xuất phân tích và xử lý số liệu

1.1 Đề xuất phân tích

- Với mục đích chính là dự đoán nguy cơ mắc bệnh tiểu đường dựa trên tập dữ liệu có sẵn, đề xuất phân tích mối liên hệ giữa biến mục tiêu là biến tình trạng bệnh tiểu đường (Diabetes_012) với các biến còn lại và ảnh hưởng nếu có của chúng đến biến mục tiêu.
- Qua mô tả dữ liệu còn cho thấy dữ liệu bị mất cân bằng nên đề xuất phân tích tiền xử lý và phân tích dữ liệu mất cân bằng liên quan đến bệnh tiểu đường dựa trên tập dữ liệu.
- Bên cạnh đó, qua quan sát ta thấy BMI có ảnh hưởng cao đến bệnh tiểu đường. Do đó, biến BMI khá quan trọng trong việc dự đoán bệnh tiểu đường, nên ta sẽ đặc biệt chú ý đến biến này và xem xét các ảnh hưởng của các biến khác lên BMI như thế nào.

1.2 Đề xuất các phương pháp xử lý số liệu

- Do biến mục tiêu là biến định tính (phân loại) nên xây dựng mô hình dự đoán bằng các mô hình phân loại sử dụng các phương pháp như Naive Bayes, hồi quy logistic, LDA và QDA.
- Sử dụng các phương pháp liên quan đến A/B testing như Permutation test, Chi-bình phương... để tìm ra mối liên hệ giữa các yếu tố khác đến biến mục tiêu, giúp tăng tính thuyết phục của việc sử dụng biến trong mô hình, đưa ra khuyến nghị thay đổi hành vi hoặc chính sách sức khỏe dựa trên kết quả kiểm định.
- Sử dụng mô hình hồi quy tuyến tính với việc lựa các biến thích hợp để xây dựng mô hình dự đoán chỉ số BMI có tính chính xác cao. Qua đó hiểu rõ các yếu tố ảnh hưởng đến BMI - một tác nhân quan trọng ảnh hưởng đến nguy cơ mắc bệnh tiểu đường.

2 Các mục tiêu phân tích

Mục tiêu 1. Kiểm tra tình trạng bệnh tiểu đường có bị ảnh hưởng bởi các nhóm đối tượng khác nhau hay không.

Mục tiêu 2. Xây dựng mô hình dự đoán bệnh tiểu đường với các bộ dữ liệu đã xử lý.

Mục tiêu 3. Tìm thấy các yếu tố mà ảnh hưởng của chúng lên mô hình hồi quy chỉ số BMI có ý nghĩa thống kê cao.

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

Mục tiêu 1: Sử dụng A/B testing.

- Sử dụng phương pháp permutation cho 2 nhóm đối với các biến nhị phân tác động đến biến mục tiêu
- Sử dụng phương pháp chi-bình phương đối với các biến phân loại tác động đến biến mục tiểu

Mục tiêu 2: Sử dụng nhiều mô hình phân loại

- Tiền xử lý dữ liệu:
  - Lược bỏ các biến không liên quan đến biến mục tiêu dựa trên kiểm định độc lập.
  - Xử lý giá trị ngoại lai cho các biến định lượng.
  - Xử lý mất cân bằng dữ liệu: thực hiện under-sampling, over-sampling và SMOTE để cân bằng giữa số lượng các nhóm.
- Xây dựng mô hình:
  - Áp dụng các thuật toán như Naive Bayes, LDA, QDA, và Logistic Regression.
- Đánh giá các mô hình với tập dữ liệu huấn luyện và kiểm tra, biểu diễn kết quả:
  - Vẽ biểu đồ ROC, Confusion matrix, tính toán chỉ số AUC, Kappa, F1-score, Precision, Recall và Accuracy.
- Sử dụng Lasso Regression để kiểm tra và xem xét lựa chọn bỏ biến từ kiểm định Chi bình phương đã tối ưu chưa.

Mục tiêu 3: Sử dụng hồi quy tuyến tính:

- Kiểm tra dữ liệu:
  - Tải dữ liệu lên và bỏ biến diabetes
  - Kiểm tra dữ liệu khuyết và xử lý nếu có
  - Kiểm tra sự tương quan giữa các biến và giữa các biến với BMI bằng ma trận tương quan
  - Kiểm tra đa cộng tuyến bằng VIF
  - Chuyển các biến phân loại sang dạng factor
- Xây dựng mô hình
  - Xây dựng mô hình ban đầu với tất cả các biến (trừ diabetes)
  - Đánh giá khái quát mô hình ban đầu (R^2, các biến có ý nghĩa thống kê và không có ý nghĩa thống kê qua p-value của các biến rồi xem xét xem có loại được biến nào không
  - Tìm kiếm các biến tương tác tiềm năng thông qua mô hình tuyến tính bậc 2, kiểm tra xem các biến tương tác nào có ý nghĩa thống kê, rồi đem từng biến tương tác tiềm năng vừa tìm được bỏ vào mô hình hồi quy ban đầu xem nó có thật sự có ý nghĩa thống kê với mô hình không.
  - Xây dựng được mô hình mới với các biến tương tác vừa chọn và kiểm tra xem hiệu suất mô hình có được cải thiện không

- Lựa chọn mô hình:
  - Tìm ra mô hình tối ưu với số lượng biến phù hợp và tang hiệu suất mô hình bằng phương pháp hồi quy từng bước dựa trên tiêu chí đánh giá Mallows' Cp.
  - Mallows' Cp đo lường sự cân bằng giữa độ chính xác của mô hình (bias) và độ phức tạp (variance). Mô hình tốt nhất có Mallows' Cp gần bằng số lượng biến trong mô hình.
  - Xây dựng mô hình tối ưu với dựa trên các biến được chọn bởi regsubsets
  - Đánh giá mô hình và xem có cải thiện hiệu suất không
  - Kiểm định hiệu suất mô hình bằng Cross-Validation
- Chẩn đoán mô hình:
  - Kiểm tra sự tuyến tính của mô hình bằng phương pháp sử dụng biểu đồ Residuals vs Fitted.
  - Kiểm tra đồng nhất phương sai bằng biểu đồ Scale-Location
  - Kiểm tra phân phối của phần dư:
    - Quan sát biểu đồ Histogram xem phân phối phần dư có ở dạng chuẩn không
    - Dùng Q-Q plot để kiểm tra sự phù hợp giữa phần dư và phân phối chuẩn lý thuyết
    - Nhận xét biểu đồ và xử lí nếu phần dư không theo phân phối chuẩn
    - Xử lí bằng cách loại bỏ Outliers dựa trên phương pháp dung Cook's Distance
  - Xây dựng mô hình tối ưu cuối cùng
    - Chạy lại mô hình cuối cùng sau khi chẩn đoán để được mô hình tối ưu nhất
    - Đánh giá hiệu suất và các chỉ số của mô hình
    - Kiểm tra xem các biến trong mô hình có thật sự ảnh hưởng lên BMI hay không và đưa ra kết luận

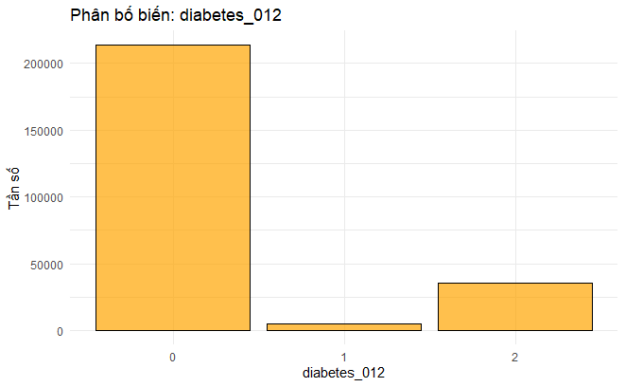
4 Mô tả và biểu diễn tổng hợp dữ liệu

4.1 Tổng quan

- Tên biến chưa được chuẩn hóa
- Dữ liệu không bị khuyết nhưng bị mất cân bằng
- Có tổng cộng 22 biến gồm cả biến định tính (phân loại, nhị phân) và biến định lượng)

4.2 Biến mục tiêu

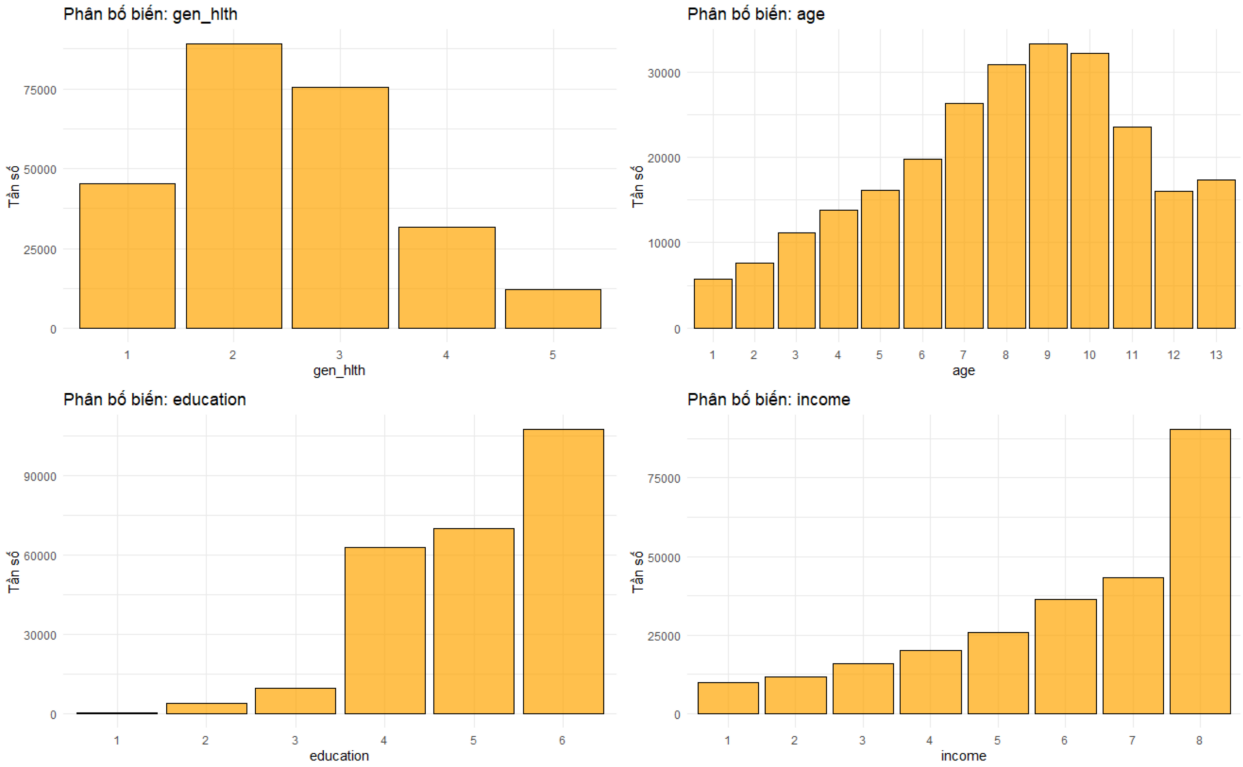
Biến mục tiêu có dạng định tính - phân loại phân bố không đều ở các lớp



Hình 1: Tần số biến mục tiêu

4.3 Biến định tính

- Biến phân loại: Gồm gen_hlth, age, education, income cũng phân bố không đều



Hình 2: Tần số các biến phân loại

- Biến nhị phân:

– Gồm các biến: high_bp, high_chol, chol_check, smoker, stroke, heart_diseaseor_attack, phys_activity, fruits, veggies, hvy_alcohol_consump, any_healthcare, no_docbc_cost, diff_walk, sex

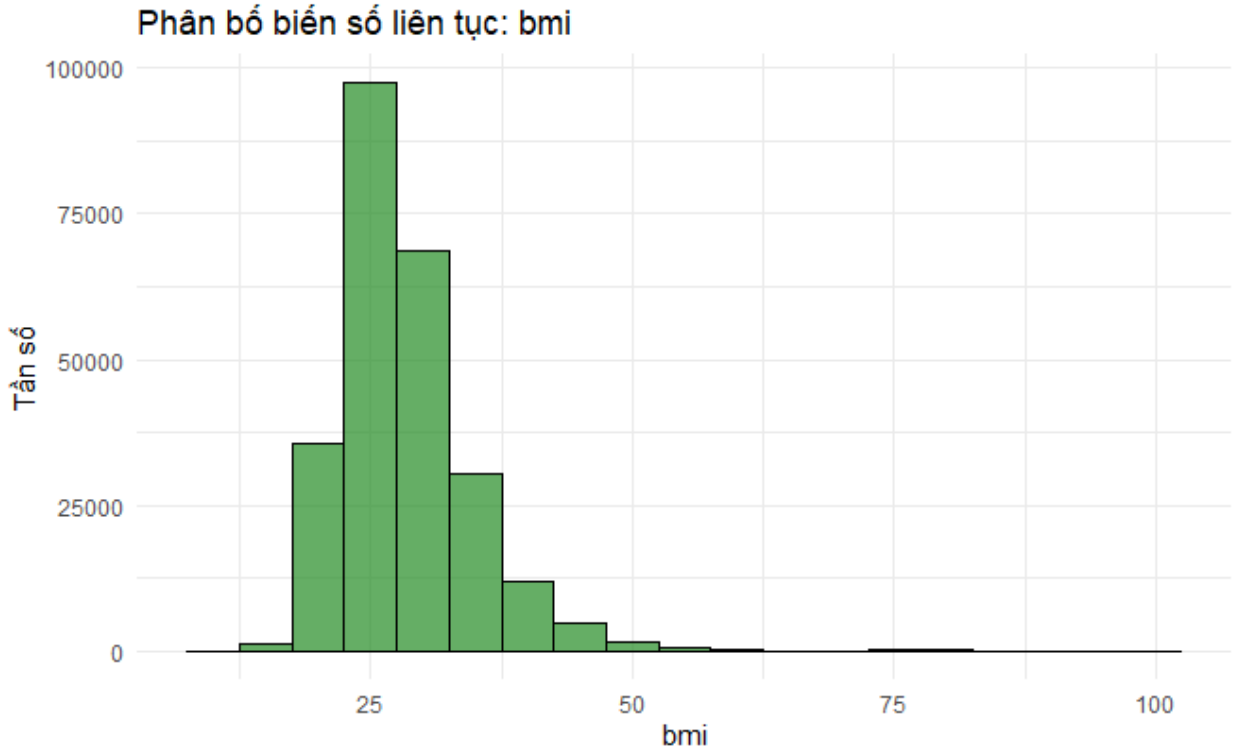
– Đa số các biến phân bố không đồng đều, chủ yếu lệch phải ở các biến thể hiện tính tiêu cực, lệch trái ở các biến tích cực cho thấy xu hướng chung của tập dữ liệu có lối sống khá lành mạnh.

Hình 3: Tần số các biến nhị phân

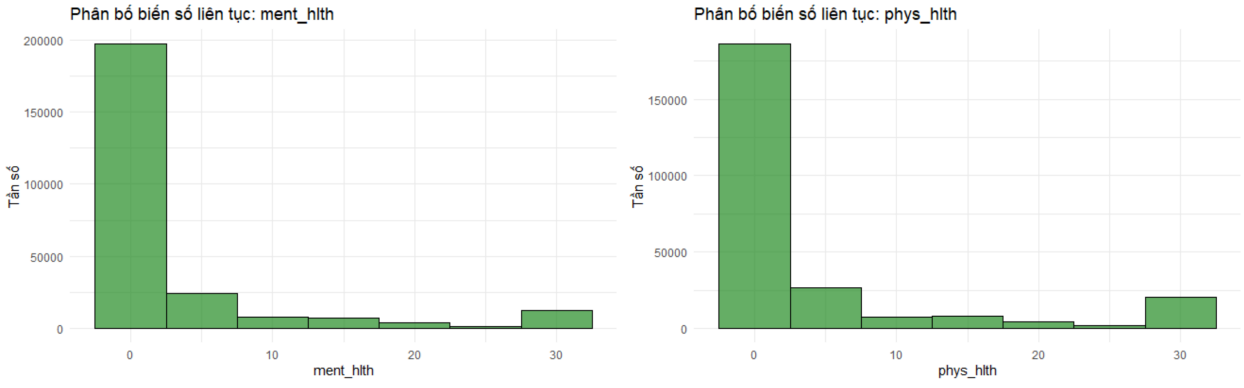
4.4 Biến liên tục

Gồm các biến bmi, ment_hlth, phys_hlth.

- Biến bmi phân bố có dạng chuông nhưng lệch phải.



- Hai biến ment_hlth, phys_hlth có dạng phân bố lệch phải hoàn toàn thể hiện đa số người tham gia khảo sát là khỏe mạnh

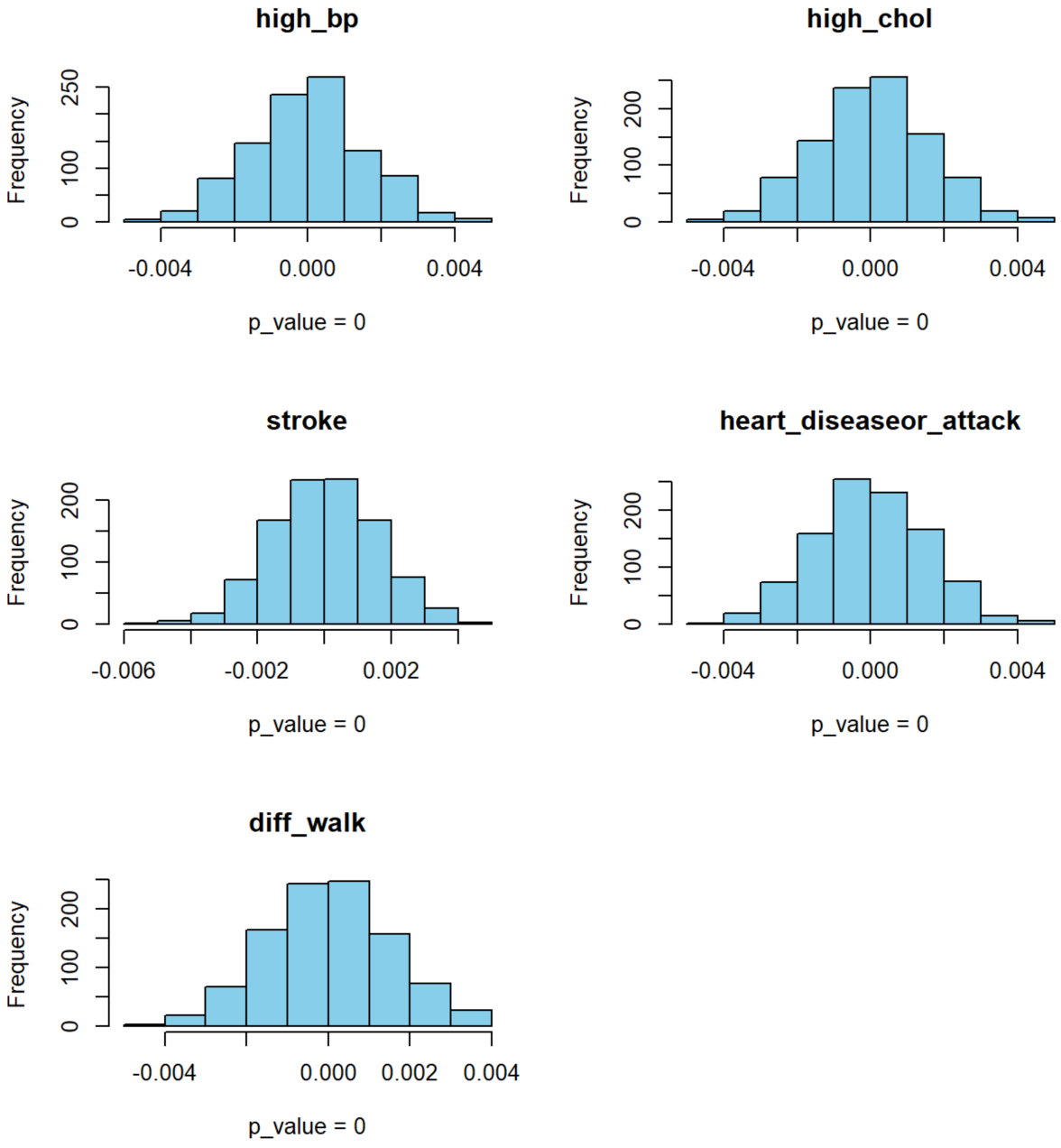


Hình 4: Biểu đồ phân bố Mental Health và Physic Health

5 Phân tích kết quả đạt được các mục tiêu đã đề ra

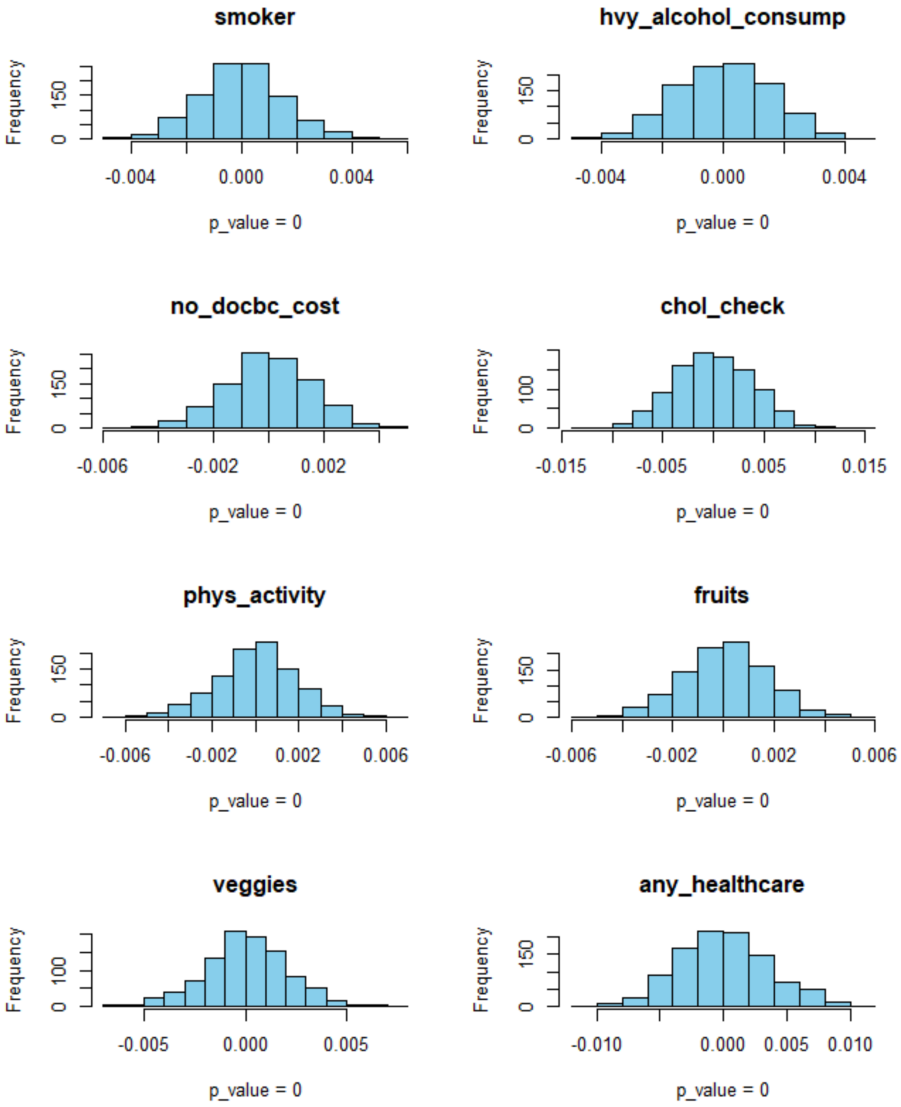
Mục tiêu 1. Chia các biến giải thích thành các nhóm có cùng chủ đề như sau:

(a) **Vấn đề sức khỏe và tình trạng bệnh lý:** Với mức ý nghĩa 0.05, bác bỏ được H_0 hay các bệnh lý như cao huyết áp, cao cholesterol, đột quỵ, đau tim, khó khăn về vận động đều ảnh hưởng đến tình trạng bệnh tiểu đường.



Hình 5: Permutation Test (1)

(b) **Liên quan đến chất lượng cuộc sống và sức khỏe tâm lý:** Với mức ý nghĩa 0.05, cả lối sống tiêu cực, lành mạnh đều ảnh hưởng đáng kể đến tình trạng bệnh tiểu đường và có thể tình trạng tiểu đường không phụ thuộc vào điểm tự đánh giá sức khỏe.

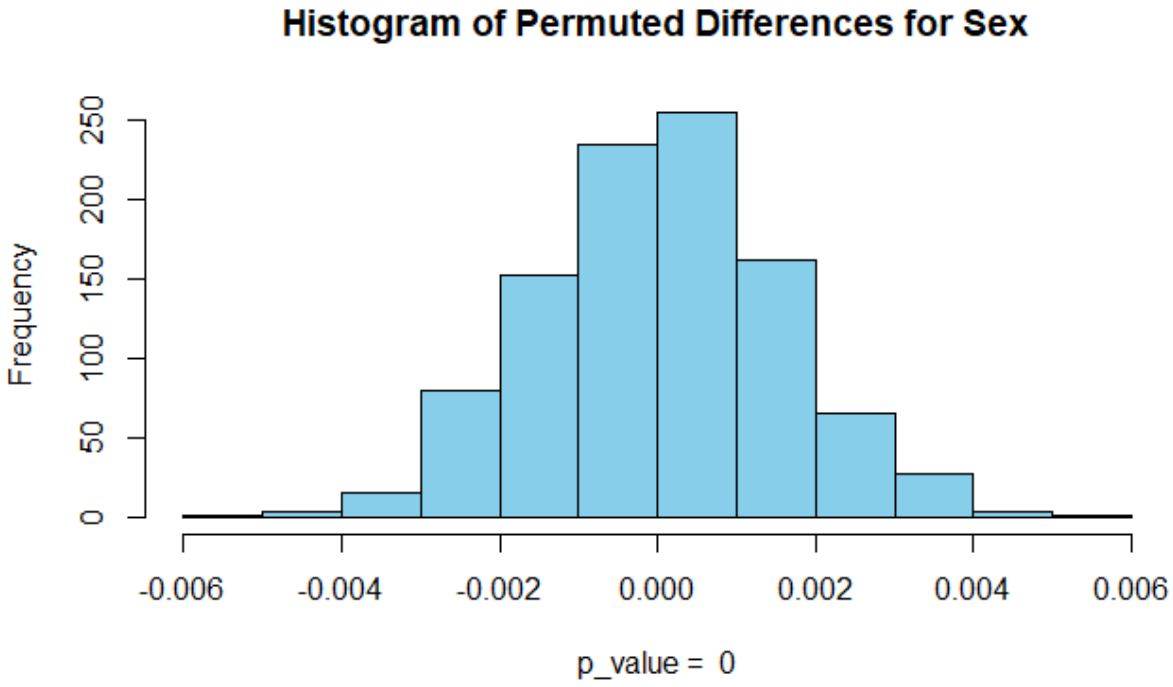


Hình 6: Permutation Test (2)

```
$P_value
 [1] 0.2131673 0.2373865 0.5017810 0.9986350 0.5218465 0.8741609 0.3733506
0.7128793 0.3204125 0.5075115 0.7822930 0.9443713 0.3144614 0.1012737
[15] 0.2504138 0.6468666 0.1035610 0.1253995 0.2291763 0.5569559
```

Hình 7: Giá trị P-value của kiểm định Chi-square cho Điểm tự đánh giá sức khỏe

(c) **Vấn đề về nhân khẩu học:** với mức ý nghĩa 0.05, giới tính có ảnh hưởng đến tình trạng bệnh tiểu đường và tình trạng bệnh tiểu đường không phụ thuộc vào tuổi, trình độ học vấn hay thu nhập.




```
$P_value
 [1] 0.606065840 0.506291035 0.449791466 0.418216474 0.644191593 0.565712421
0.384873904 0.271312167 0.240271467 0.523487208 0.001643089 0.896660093
[13] 0.118530402 0.653342836 0.666391333 0.467014574 0.541181781 0.068956786
0.955358778 0.858397682
```

Hình 8: Giá trị P-value của kiểm định Chi-square cho Tuổi

```
$P_value
 [1] 0.77424715 0.92881976 0.73934944 0.70578848 0.04775420 0.57126547 0.37674557
0.48449547 0.97997672 0.42863546 0.67889391 0.02327591 0.80726851
[14] 0.89233983 0.47372270 0.32896411 0.55960166 0.81487809 0.35374430 0.98786213
```

Hình 9: Giá trị P-value của kiểm định Chi-square cho Trình độ học vấn

```
$P_value
 [1] 0.99465569 0.58047108 0.22605276 0.11431796 0.24365497 0.19820716 0.18860396
0.08437699 0.57595593 0.45305230 0.25695045 0.16978134 0.35957295
[14] 0.86851297 0.97833681 0.84741077 0.12574669 0.32396770 0.82452631 0.85967909
```

Hình 10: Giá trị P-value của kiểm định Chi-square cho Thu nhập

Mục tiêu 2. Sau các bước phân tích và áp dụng các phương pháp phân loại, nhóm em nhận thấy dùng phương pháp phân loại bằng hồi quy Logistic và xử lý mất cân bằng dữ liệu bằng phương pháp SMOTE đưa ra mô hình có kết quả khả quan nhất.

- Các hệ số đánh giá cho mô hình trên tập test:

Nhóm	**0**	**1**	**2**
Precision	0.5669326	0.4463172	0.5376123

- Khoảng 56.7% các mẫu được dự đoán thuộc nhóm 0 là chính xác:

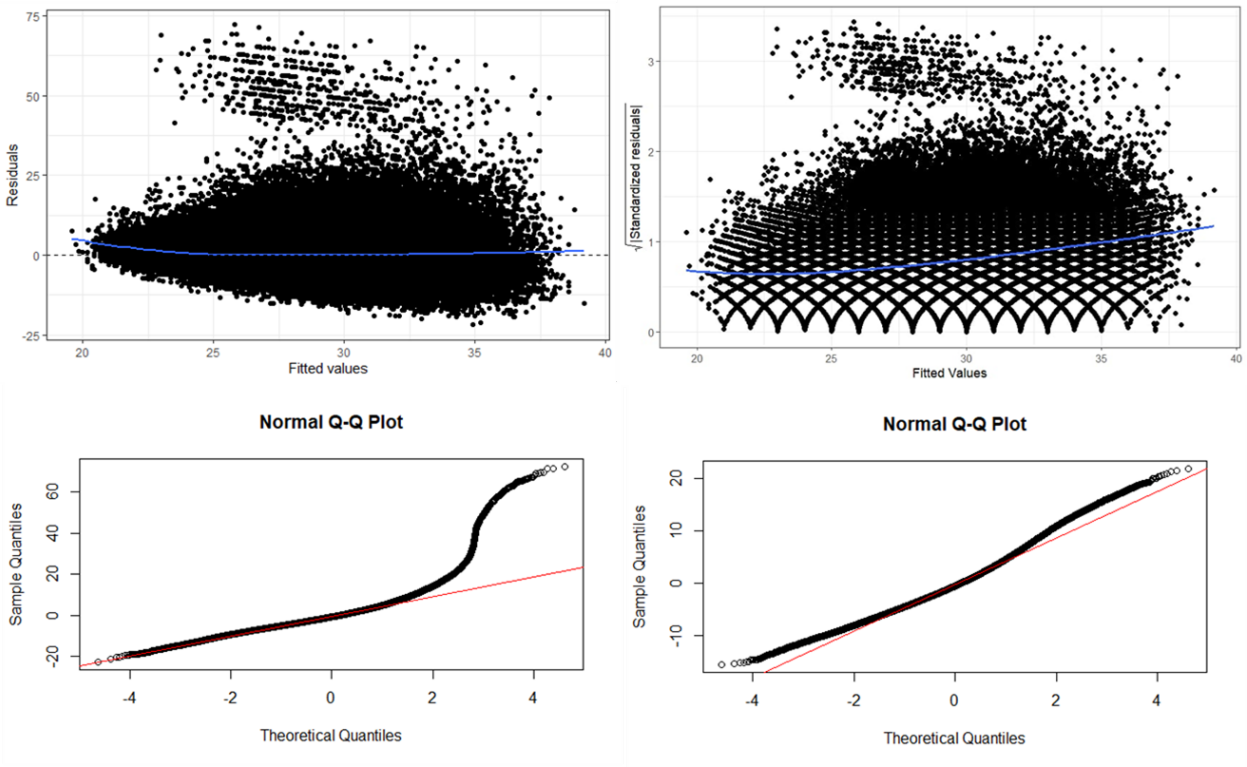
Nhóm	**0**	**1**	**2**
Recall	0.6884105	0. 3593511	0.5271720

- Trong những mẫu có giá trị 0 thực sự, mô hình chỉ tìm được 68.84%.
- Tổng thể, mô hình dự đoán chính xác 52.5% số mẫu trên tập kiểm tra.
- Trung bình các nhóm, mô hình đạt hiệu suất ở mức khoảng 52.1

Mục tiêu 3. Các yếu tố mà ảnh hưởng của chúng lên mô hình hồi quy chỉ số BMI có ý nghĩa thống kê cao được thể hiện trong mô hình cuối cùng có kết quả chuẩn đoán khả quan như sau:

```
final_md <- lm(bmi ~ high_bp*high_chol + high_bp*phys_activity +
                     high_bp*hvy_alcohol_consump + high_bp*gen_hlth +
                     high_bp*diff_walk + high_chol*smoker + smoker*fruits +
                     smoker*veggies + smoker*gen_hlth + smoker*diff_walk +
                     chol_check*age + stroke*gen_hlth + stroke*phys_hlth +
                     heart_diseaseor_attack + education + income + sex,
                     data = data_no_outliers)
```

Hình 11: Mô hình hồi quy chỉ số BMI



Hình 12: Chẩn đoán mô hình

6 Nhận xét và kết luận.

6.1 Nhận xét

6.1.1 Đối với tình trạng bệnh tiểu đường

- Các tình trạng bệnh lý khác, chất lượng cuộc sống và cả giới tính đều tác động đến tình trạng bệnh tiểu đường.
- Tuy nhiên, các yếu tố như điểm tự đánh giá sức khỏe, hay các vấn đề về nhân khẩu học không thấy sự ảnh hưởng đến tình trạng bệnh tiểu đường.

6.1.2 Đối với việc dự đoán tình trạng bệnh tiểu đường

- Tiêu thụ rượu nặng (hvy_alcohol_consump1): Hệ số hồi quy âm mạnh (-0.977148 cho lớp 1 và -1.209787 cho lớp 2) cho thấy việc tiêu thụ rượu làm giảm đáng kể nguy cơ mắc bệnh tiểu đường.
- Kiểm tra cholesterol (chol_check1): Hệ số hồi quy dương lớn (3.461872 cho lớp 1 và 2.573086 cho lớp 2) chỉ ra rằng việc kiểm tra cholesterol dương tính làm tăng khả năng mắc bệnh tiểu đường.
- BMI: Hệ số dương cao (2.611641 cho lớp 1 và 3.683613 cho lớp 2) chứng minh mối liên hệ mạnh giữa chỉ số BMI cao và nguy cơ tiểu đường.

- Hoạt động thể chất (phys_activity1): Hệ số âm (-0.0964624 cho lớp 1 và -0.10064663 cho lớp 2) cho thấy hoạt động thể chất thường xuyên làm giảm nguy cơ tiểu đường.
- Sức khỏe thể chất (phys_hlth): Hệ số dương đáng kể (3.003415 cho lớp 1 và 4.275152 cho lớp 2) cho thấy sức khỏe thể chất kém làm tăng nguy cơ tiểu đường.
- Mô hình dự đoán có thể hỗ trợ hiệu quả cho việc sàng lọc và tư vấn sức khỏe cộng đồng với tỷ lệ chính xác tương đối

```
$Precision
        0         1         2
0.5655291 0.4466695 0.5386441

$Recall
        0         1         2
0.6890969 0.3583528 0.5274528

$Accuracy
[1] 0.5249675

$Macro_F1
[1] 0.5209267

$Kappa
[1] 0.2874513
```

Hình 13: Chỉ số đánh giá mô hình

6.1.3 Đối Với BMI

- High_bp (Huyết áp cao) tác động đến BMI thông qua các tương tác:
  - high_bp * high_chol: Huyết áp cao và cholesterol cao có thể cùng tăng nguy cơ ảnh hưởng đến BMI.
  - high_bp * phys_activity: Hoạt động thể chất giảm hoặc không đủ ở người huyết áp cao có thể làm tăng BMI.
  - high_bp * hvy_alcohol_consump: Tiêu thụ rượu nặng và huyết áp cao góp phần làm tăng BMI.
  - high_bp * gen_hlth: Sự kết hợp giữa huyết áp cao và sức khỏe tổng quát kém làm tăng BMI.
  - high_bp * diff_walk: Khó khăn trong đi lại kết hợp với huyết áp cao cũng là yếu tố đáng kể.
- High_chol (Cholesterol cao) kết hợp với hành vi hút thuốc (high_chol* smoker) ảnh hưởng đến BMI.
- Smoker (Hút thuốc lá) hút thuốc tác động đến BMI qua nhiều yếu tố khác:
  - smoker * fruits: Việc hút thuốc và thiếu trái cây trong chế độ ăn góp phần làm thay đổi BMI.

  - smoker * veggies: Tương tự, sự thiếu rau củ cùng với hút thuốc làm tăng BMI.
  - smoker * gen_hlth: Sức khỏe tổng quát kém của người hút thuốc ảnh hưởng lớn đến BMI.
  - smoker * diff_walk: Khó khăn đi lại cũng có tác động tiêu cực khi kết hợp với hút thuốc.
  - Chol_check (Kiểm tra cholesterol) kết hợp với tuổi (chol_check *age), cho thấy tuổi tác cùng tầm soát cholesterol có ảnh hưởng đến BMI.

- Stroke và Heart_diseaseor_attack (Đột quỵ và bệnh tim) những người từng bị đột quỵ hoặc bệnh tim có thể có BMI khác biệt do ảnh hưởng đến sức khỏe tổng quát và hoạt động thể chất.

- Education (Trình độ học vấn) trình độ học vấn có thể phản ánh thói quen sống lành mạnh hơn (ăn uống, hoạt động thể chất) và từ đó ảnh hưởng đến BMI.

- Income (Thu nhập) thu nhập cao có thể đi kèm với khả năng tiếp cận thực phẩm chất lượng hoặc chế độ sống lành mạnh hơn, ảnh hưởng đến BMI.

- Sex (Giới tính) ảnh hưởng đến BMI qua sự khác biệt về sinh học, hành vi, và các yếu tố văn hóa.

6.2 Kết luận

Sau quá trình xử lý dữ liệu, ta có các kết luận sau:

Đối với tình trạng bệnh tiểu đường, những người khỏe mạnh, ít bệnh và có lối sống lành mạnh ít có nguy cơ mắc bệnh hơn mà không quan trọng các vấn đề tuổi tác, thu nhập hay trình độ học vấn. Các khuyến nghị có thể đưa ra để giảm nguy cơ mắc bệnh tiểu đường là triển khai các chiến lược giáo dục sức khỏe, thúc đẩy lối sống lành mạnh và can thiệp sớm cần được ưu tiên.

Riêng chỉ số BMI - biến quan trọng trong việc đánh giá sức khỏe khách quan của một liên quan trực tiếp đến tình trạng bệnh tiểu đường, ta cũng thấy rằng nó có liên hệ mật thiết đến các yếu tố liên quan đến sức khỏe và hành vi cá nhân cũng như các đặc điểm nhân khẩu. Trong đó, nhiều yếu tố không chỉ ảnh hưởng BMI mà còn có sự tương tác phức tạp, ví dụ như sự kết hợp giữa huyết áp cao, cholesterol, và các hành vi lối sống. Các biến như sức khỏe tổng quát (gen_hlth), khó khăn đi lại (diff_walk) phản ánh rõ rệt tác động của tình trạng sức khỏe chung lên BMI. Để chỉ số BMI ở tình trạng tốt, cần kiểm soát các bệnh lý như cao huyết áp, cao cholesterol và sống lành mạnh hơn.