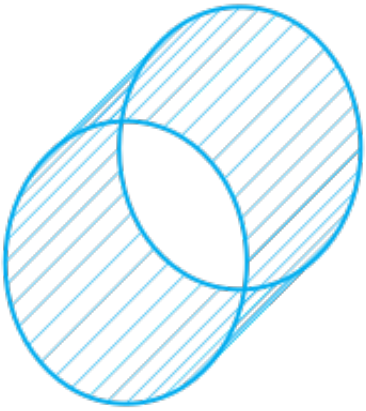
TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN, ĐHQG-HCM
KHOA TOÁN - TIN HỌC



BÁO CÁO ĐỒ ÁN CUỐI KỲ
MÔN HỌC: XỬ LÍ SỐ LIỆU THỐNG KÊ

PROJECT 2
CDC DIABETES HEALTH INDICATORS

Giảng viên môn học: Tô Đức Khánh

Nhóm: 2

STT	Họ và tên	MSSV
1	Võ Nguyễn Phúc	21110022
2	Đặng Ngọc Trúc Quỳnh	21110164
3	Trần Nguyễn Thanh Phong	22110155
4	Hồ Minh Quân	22110170
5	Lê Nguyễn Thanh Tân	22110196
6	Trần Thị Thanh Trúc	22110240

Thành phố Hồ Chí Minh, ngày 14 tháng 1 năm 2025

Mục lục

1 Phân công

BẢNG PHÂN CÔNG ĐỒ ÁN CUỐI KỲ

STT	MSSV	HỌ VÀ TÊN	PHÂN CÔNG
1	21110022	Võ Nguyễn Phúc	Phân loại LDA, QDA
2	21110164	Đặng Ngọc Trúc Quỳnh	Mô hình logistic và multinominal logistic
3	22110155	Trần Nguyễn Thanh Phong	Phân loại Naive Bayes- Tổng hợp file R
4	22110170	Hồ Minh Quân	A/B testing - Resampling method
5	22110196	Lê Nguyễn Thanh Tân	A/B testing - Kiểm định phi tham số
6	22110240	Trần Thị Thanh Trúc	A/B testing - Kiểm định phi tham số

Tất cả thành viên đều tự tìm hiểu và xử lí dữ liệu theo các phương pháp được phân công, tự lên ý tưởng và họp bàn về ý tưởng, thảo luận cách làm bài qua 3 buổi họp nhóm và tự thực hiện phần code, phần báo cáo cá nhân trong bài báo cáo tổng hợp.

2 A/B testing và Kiểm định phi tham số

Ở phần này chúng ta sẽ kiểm định sự ảnh hưởng của các biến độc lập đến biến mục tiêu ta cần xét là "diabetes_012 bằng cách tính theo phương pháp kiểm định Chi bình phương, đồng thời phân loại những biền này theo 3 nhóm bao gồm : biến nhị phân,biến phân bậc và biến định lượng

Đối biến nhị phân ta sử dụng kiểm định bằng phương pháp sử dụng OR để xét xem có sự khác nhau giữa tỉ lệ các trạng thái như hút thuốc/không hút thuốc đối với bệnh tiểu đường hay không, đồng thời xét xem biến đó ảnh hưởng đến bệnh theo cách tích cực (giảm tỉ lệ mắc bệnh), hay tiêu cực(tăng tỉ lệ mắc bệnh).

Đối với biến phân loại và định lượng ta sử dụng 2 phương pháp permutation test và permutation anova để xét xem có sự khác nhau giữa tỉ lệ người bệnh ở các biến đó theo nhóm, trung bỉnh hay không, đồng thời quan sát xem giữa các biến độc lập có mối quan hệ tương quan nào không.

Cuối cùng ta sẽ xét thử xem liệu một vài biến định lượng(BMI,...) có sự khác biệt rõ rệt, ảnh hưởng mạnh đến biến mục tiêu hay không.

2.1 Cơ sở lý thuyết

2.1.1 Odds Ratio (OR)

Odds Ratio (OR) là một thước đo thống kê được sử dụng để xác định mức độ liên quan giữa hai biến trong các nghiên cứu quan sát, đặc biệt là trong nghiên cứu trường hợp - đối chứng (case-control studies).

Odds Ratio được tính dựa trên bảng hai chiều (2x2) như sau:

	Biến mục tiêu (+)	**Biến mục tiêu (-)**
Biến trạng thái có (+)	a	b
Biến trạng thái không (-)	c	d

$$OR = \frac{a \times d}{b \times c}$$

Kiểm định Odds Ratio:

Để kiểm định sự khác biệt của Odds Ratio so với giá trị giả thuyết, sử dụng kiểm định log-rank :

+ Khoảng Tin Cậy cho OR:

$$\ln(OR) \pm Z_{\alpha/2} \times \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}$$

+ Sau đó, chuyển đổi ngược lại bằng hàm mũ để có khoảng tin cậy cho OR:

$$\text{CI}_{95\%} = \left(e^{\ln(OR) - Z_{\alpha/2} \times SE},\ e^{\ln(OR) + Z_{\alpha/2} \times SE}\right)$$

+ Trong đó:

$Z_{\alpha/2}$: Giá trị phân vị chuẩn tương ứng với mức ý nghĩa α (thường là 1.96 cho $\alpha = 0.05$).

SE: Sai số chuẩn của $\ln(OR)$.

Giả thuyết liên quan đến Odds Ratio

Trong kiểm định giả thuyết về Odds Ratio, các giả thuyết như sau:

$$\begin{cases} H_0 : OR = 1 & \text{(Không có sự liên quan giữa hai trạng thái)} \\ H_1 : OR \neq 1 & \text{(Có sự liên quan giữa hai trạng thái)} \end{cases}$$

Nếu khoảng tin cậy 95% không chứa giá trị 1, ta bác bỏ giả thuyết không H_0 và kết luận rằng có sự liên quan đáng kể giữa hai biến.

Ý nghĩa của Odds Ratio (OR) trong kiểm định

Khi đã tính toán Odds Ratio (OR) và kiểm tra khoảng tin cậy, ý nghĩa của giá trị OR được giải thích như sau:

- **Khi** $OR = 1$**:** Điều này có nghĩa là không có mối liên quan giữa hai biến trạng thái. Yếu tố nghiên cứu không làm tăng hoặc giảm xác suất xảy ra kết quả.
- **Khi** $OR > 1$**:** Điều này cho thấy rằng biến trạng thái (+) có liên quan đến xác suất tăng của kết quả (+).
  - Ví dụ: Nếu $OR = 2$, nghĩa là khả năng xảy ra kết quả (+) khi có trạng thái (+) gấp 2 lần so với khi không có trạng thái (+).
- **Khi** $OR < 1$**:** Điều này cho thấy rằng biến trạng thái (+) có liên quan đến xác suất giảm của kết quả (+).
  - Ví dụ: Nếu $OR = 0.5$, nghĩa là khả năng xảy ra kết quả (+) khi có trạng thái (+) chỉ bằng một nửa so với khi không có trạng thái (+).

2.1.2 Permutation test và permutation anova

Cả 2 phương pháp này sẽ có những giả thuyết như sau :

- **Permutation test**

  Với μ_1, μ_2 là trung bình giá trị của biến cần xét theo một biến hoặc nhóm biến nào đó mang hai thuộc tính.

  Ví dụ: Trung bình BMI giữa nhóm nam và nữ với μ_1 là trung bình BMI của nhóm nam và μ_2 là trung bình BMI của nhóm nữ.

– **TH1: Kiểm định hai phía**

$$\begin{cases} H_0 : \mu_1 = \mu_2 & \text{(Không có khác nhau giữa hai nhóm)} \\ H_1 : \mu_1 \neq \mu_2 & \text{(Có khác nhau giữa hai nhóm)} \end{cases}$$

Nếu $p_{value} < 0.05$, ta sẽ bác bỏ giả thuyết H_0, chấp nhận đối thuyết H_1.

– **TH2: Kiểm định phía bên trái**

$$\begin{cases} H_0 : \mu_1 \geq \mu_2 & \text{(Nhóm 1 có trung bình lớn hơn hoặc bằng nhóm 2)} \\ H_1 : \mu_1 < \mu_2 & \text{(Nhóm 1 có trung bình bé hơn nhóm 2)} \end{cases}$$

Nếu $p_{value} < 0.05$, ta sẽ bác bỏ giả thuyết H_0, chấp nhận đối thuyết H_1.

- **Permutation ANOVA**

Với $\mu_1, \mu_2, \ldots, \mu_j$ là trung bình giá trị của biến cần xét theo một biến hoặc nhóm biến nào đó mang hơn hai thuộc tính.

Ví dụ: Trung bình BMI giữa các nhóm tuổi được xếp thành 13 nhóm.

Giả thuyết kiểm định

$$\begin{cases} H_0 : \mu_1 = \mu_2 = \mu_3 = \cdots = \mu_j & \text{(Giữa các nhóm không có sự khác biệt)} \\ H_1 : \mu_i \neq \mu_j \ \forall i \neq j & \text{(Giữa các nhóm có sự khác biệt)} \end{cases}$$

Nếu $p_{value} < 0.05$, ta sẽ bác bỏ giả thuyết H_0, chấp nhận đối thuyết H_1.

2.1.3 Chi bình phương

Công thức Chi bình phương

Giá trị thống kê χ^2 được tính theo công thức:

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

Trong đó:

- O_i: Tần suất quan sát được trong ô thứ i.
- E_i: Tần suất kỳ vọng trong ô thứ i, được tính bằng:

$$E_i = \frac{\text{Tổng hàng} \times \text{Tổng cột}}{\text{Tổng tất cả các giá trị}}$$

2.2 Tiến hành áp dụng lên dữ liệu

2.2.1 Tiến hành dùng phương pháp kiểm định Chi bình phương để đánh giá độ ảnh hưởng của tát cả các biến độc lập lên biến mục tiêu :

Ta thu được kết quả được xếp theo thứ tự từ giá trị "Score" lớn nhất đến bé nhất như sau :

Feature	Score
gen_hlth	19340.2343
high_bp	15825.1795
bmi	13103.7019
diff_walk	10203.0112
high_chol	9640.9845
age	8975.6578
heart_diseaseor_attack	6708.0364
phys_hlth	6436.6757
income	5154.3553
education	2853.7023
phys_activity	2460.9016
stroke	2315.7016
chol_check	1338.3465
ment_hlth	1315.1697
hvy_alcohol_consump	1036.6095
smoker	503.2836
veggies	436.4558
sex	244.3055
no_docbc_cost	179.8283
fruits	148.9966
any_healthcare	146.1926

Nhận xét : Biến có "Score" càng lớn thì khả năng ảnh hưởng của biến đến biến mục tiêu càng mạnh. Song kiểm định này chỉ tương đối chính xác đối với các biến định tính nên ta sẽ dùng nhiều kiểm định khác đối với từng nhóm biến khác nhau.

Biến định lượng BMI có giá trị "Score" khá lớn nằm trong top 3, có thể ảnh hưởng rõ rệt đến việc phân loại biến mục tiêu, lý do bởi vì đây là biến định lượng duy có điểm cao, trong khi các biến cao điểm đều là biến nhị phân.

2.2.2 Phân loại biến

Nhằm xác định những nhóm biến dựa trên kiểu dữ liệu mà biến đó thể hiện như nhóm biến nhị phân xác định trạng thái (0,1),nhóm biến phân bậc và nhóm biến định lượng

2.2.3 Tiến hành kiểm định

a) Sử dụng kiểm định OR đối với nhóm biến trạng thái và kết quả chuẩn đoán bệnh tiểu đường

Loại biến	Danh sách biến
Biến nhị phân	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
Biến phân bậc	gen_hlth, age, education, income
Biến định lượng	bmi, ment_hlth, phys_hlth

Bảng 2: Danh sách các biến theo nhóm

Với 0 là biến chỉ trạng thái đó không xuất hiện trên người khảo sát và 1 biểu thị cho trạng thái đó được ghi nhận là có trên người khảo sát

Nhìn vào bảng phân bố phía dưới ta thấy:

+Nhóm không có tỉ lệ khác biệt: smoker ,sex

+Nhóm chiếm tỉ lệ cao ở mục tiểu đường và tiền tiểu đường:high_bp,high_chol, diff_walkheart_diseaseor_attackstroke

+Nhóm chiếm tỉ lệ cao ở mục không tiểu đường : phys_activity,fruits,veggies, hvy_alcohol_consump,any_healthcare

Biến nhị phân	Diabetes = 0	Diabetes = 1	Diabetes = 2
high_bp	0.6048 / 0.3952	0.3709 / 0.6291	0.2477 / 0.7523
high_chol	0.6047 / 0.3953	0.3791 / 0.6209	0.3305 / 0.6695
chol_check	0.0473 / 0.9527	0.0134 / 0.9866	0.0069 / 0.9931
smoker	0.5449 / 0.4551	0.5072 / 0.4928	0.4808 / 0.5192
stroke	0.9645 / 0.0355	0.9428 / 0.0572	0.9069 / 0.0931
heart_diseaseor_attack	0.9200 / 0.0800	0.8566 / 0.1434	0.7762 / 0.2238
phys_activity	0.2459 / 0.7541	0.3217 / 0.6783	0.3715 / 0.6285
fruits	0.3815 / 0.6185	0.3975 / 0.6025	0.4158 / 0.5842
veggies	0.1972 / 0.8028	0.2312 / 0.7688	0.2451 / 0.7549
hvy_alcohol_consump	0.9321 / 0.0679	0.9551 / 0.0449	0.9763 / 0.0237
any_healthcare	0.0564 / 0.9436	0.0549 / 0.9451	0.0405 / 0.9595
no_docbc_cost	0.9106 / 0.0894	0.8706 / 0.1294	0.8934 / 0.1066
diff_walk	0.8515 / 0.1485	0.7224 / 0.2776	0.6263 / 0.3737
sex	0.5678 / 0.4322	0.5625 / 0.4375	0.5227 / 0.4773

Bảng 3: Tỉ lệ so sánh giữa các biến trạng thái với biến mục tiêu diabetes_012

Ta tiến hành dùng phương pháp OR để khảo sát và thu được kết quả :

Biến nhị phân	So sánh	Giá trị OR	CI thấp (95%)	CI cao (95%)
high_bp	0 vs 1	2.5956	2.4440	2.7574

high_bp	0 vs 2	4.6494	4.5307	4.7720
high_bp	1 vs 2	1.7912	1.6793	1.9101
high_chol	0 vs 1	2.5045	2.3587	2.6600
high_chol	0 vs 2	3.0976	3.0240	3.1729
high_chol	1 vs 2	1.2368	1.1607	1.3176
chol_check	0 vs 1	3.6499	2.8658	4.7432
chol_check	0 vs 2	7.1804	6.3307	8.1849
chol_check	1 vs 2	1.9673	1.4732	2.5873
smoker	0 vs 1	1.1631	1.0972	1.2330
smoker	0 vs 2	1.2930	1.2638	1.3227
smoker	1 vs 2	1.1117	1.0456	1.1820
stroke	0 vs 1	1.6497	1.4506	1.8680
stroke	0 vs 2	2.7879	2.6691	2.9113
stroke	1 vs 2	1.6899	1.4882	1.9270
heart_diseaseor_attack	0 vs 1	1.9272	1.7709	2.0942
heart_diseaseor_attack	0 vs 2	3.3182	3.2198	3.4196
heart_diseaseor_attack	1 vs 2	1.7217	1.5811	1.8776
phys_activity	0 vs 1	0.6878	0.6462	0.7323
phys_activity	0 vs 2	0.5518	0.5387	0.5653
phys_activity	1 vs 2	0.8023	0.7514	0.8564
fruits	0 vs 1	0.9348	0.8808	0.9923
fruits	0 vs 2	0.8667	0.8468	0.8870
fruits	1 vs 2	0.9271	0.8708	0.9869
veggies	0 vs 1	0.8169	0.7625	0.8758
veggies	0 vs 2	0.7565	0.7365	0.7771
veggies	1 vs 2	0.9261	0.8610	0.9954
hvy_alcohol_consump	0 vs 1	0.6461	0.5599	0.7412
hvy_alcohol_consump	0 vs 2	0.3332	0.3102	0.3575
hvy_alcohol_consump	1 vs 2	0.5158	0.4426	0.6038
any_healthcare	0 vs 1	1.0285	0.9070	1.1717
any_healthcare	0 vs 2	1.4148	1.3375	1.4976
any_healthcare	1 vs 2	1.3756	1.1968	1.5748
no_docbc_cost	0 vs 1	1.5149	1.3870	1.6516
no_docbc_cost	0 vs 2	1.2161	1.1713	1.2623
no_docbc_cost	1 vs 2	0.8028	0.7326	0.8811
diff_walk	0 vs 1	2.2032	2.0628	2.3519
diff_walk	0 vs 2	3.4201	3.3354	3.5065
diff_walk	1 vs 2	1.5523	1.4508	1.6618
sex	0 vs 1	1.0215	0.9631	1.0833
sex	0 vs 2	1.1994	1.1724	1.2272
sex	1 vs 2	1.1742	1.1040	1.2491

Bảng 4: Odds Ratio và khoảng tin cậy các biến trạng thái

Từ đó ta thấy:

+Biến có ảnh hưởng tiêu cực (Odds Ratio > 1, CI > 1): high_bp,high_chol, chol_check,smoker,stroke,heart_diseaseor_attack,diff_walk,any_healthcare, no_docbc_cost.

+Biến có ảnh hưởng tích cực (Odds Ratio < 1, CI < 1): veggies, hvy_alcohol_consump, phys_activity,fruits

+Biến không có ý nghĩa (CI chứa 1): sex

b) Áp dụng phương pháp permutation anova với nhóm biến phân bậc

Lập bảng thống kê các biến thuộc nhóm biến phân bậc với biến mục tiêu ta thu được :

gen_hlth	count	mean
1	34907	0.0740
2	77536	0.178
3	73714	0.385
4	31546	0.653
5	12078	0.787

age	count	mean
1	5512	0.0321
2	7068	0.0473
3	10025	0.0698
4	12234	0.114
5	14050	0.161
6	17299	0.219
7	23140	0.284
8	27301	0.331
9	29736	0.406
10	29168	0.468
11	22041	0.489
12	15394	0.468
13	16813	0.408

education	count	mean
1	174	0.552
2	4040	0.625
3	9467	0.518
4	61158	0.383
5	66499	0.330
6	88443	0.248

income	**count**	**mean**
1	9792	0.519
2	11757	0.555
3	15922	0.474
4	19957	0.429
5	25345	0.377
6	35001	0.322
7	40189	0.278
8	71818	0.210

Nhận xét:

+ gen_hlth (tình trạng sức khỏe) có mối liên hệ mạnh mẽ với tỉ lệ mắc bệnh, với tỉ lệ mắc bệnh tăng dần từ người có sức khỏe rất tốt đến rất xấu.

+ age (tuổi tác) cũng là yếu tố quan trọng, với tỉ lệ mắc bệnh tăng theo độ tuổi.

+ education (trình độ học vấn) và income (thu nhập) có mối liên hệ nghịch với tỉ lệ mắc bệnh, với nhóm thu nhập và trình độ học vấn thấp có tỉ lệ mắc bệnh cao hơn.

Ta tiến hành kiểm định bằng anova permutation với giả thuyết đề cập ở phần 2.1.2, và thu được kết quả :

Nhóm biến phân loại	**Giá trị p-value**
`gen_hlth`	0.000
`age`	0.000
`education`	0.000
`income`	0.000

Với kết quả trên ta bác bỏ giả thuyết là trung bình tỉ lệ mắc bệnh ở các nhóm được xét là giống nhau nhau, tức là có sự khác nhau về khả năng mắc bệnh khi xét theo từng nhóm.

c) Dùng phương pháp permutation test để kiểm định liệu có sự khác nhau giữa giới tính và việc tập luyện thể thao ảnh hưởng thế nào đến bmi,sức khỏe thể chất,...

Ta có bảng thống kê như sau :

Bảng 5: Bảng tổng hợp theo `phys_activity` và `bmi`

`phys_activity`	**count**	tb	dlc
0	61270	30.1	7.57
1	168511	28.2	6.40

Bảng 6: Bảng tổng hợp theo `phys_activity` và `ment_hlth`

phys_activity	count	tb	dlc
0	61270	4.86	9.30
1	168511	3.01	6.98

Bảng 7: Bảng tổng hợp theo `phys_activity` và `phys_hlth`

phys_activity	count	tb	dlc
0	61270	7.67	11.3
1	168511	3.59	7.78

Bảng 8: Bảng tổng hợp theo `sex` và `bmi`

sex	count	tb	dlc
0	128854	28.5	7.26
1	100927	28.9	6.12

Bảng 9: Bảng tổng hợp theo `sex` và `ment_hlth`

sex	count	tb	dlc
0	128854	4.08	8.17
1	100927	2.77	7.02

Bảng 10: Bảng tổng hợp theo `sex` và `phys_hlth`

sex	count	tb	dlc
0	128854	5.03	9.21
1	100927	4.22	8.81

Ta thấy rằng :

Hoạt động thể chất:

+ Người vận động có BMI trung bình thấp hơn

Giới tính:

+ Nam giới có BMI cao hơn nữ giới.

+ Nữ giới có sức khỏe tinh thần và thể chất dễ bị kém hơn nam.

Tiến hành kiểm định bằng phương pháp permutation test (alter = "two side"), tức kiểm định 2 phía:

Dựa vào bảng kết quả ta thấy :

Bảng 11: Tóm tắt kiểm định giữa các cặp biến và giá trị p-value

Cặp biến	**p-value**
`high_bp` vs `bmi`	0.383
`high_chol` vs `bmi`	0.350
`sex` vs `ment_hlth`	0.000
`sex` vs `phys_activity`	0.000
`phys_hlth` vs `bmi`	0.423
`phys_hlth` vs `ment_hlth`	0.000
`phys_hlth` vs `phys_activity`	0.000

+ `high_bp` vs `bmi` (p-value = 0.383): Không có sự khác biệt ý nghĩa, BMI không liên quan đáng kể đến cao huyết áp.

+ `high_chol` vs `bmi` (p-value = 0.35): Không có sự khác biệt ý nghĩa, BMI không liên quan đáng kể đến cholesterol cao.

+ `sex` vs `ment_hlth` (p-value = 0): Có sự khác biệt đáng kể, nữ giới thường có sức khỏe tinh thần kém hơn nam giới.

+ `sex` vs `phys_activity` (p-value = 0): Có sự khác biệt đáng kể, nam giới vận động nhiều hơn nữ giới.

+ `phys_hlth` vs `bmi` (p-value = 0.423): Không có sự khác biệt ý nghĩa, BMI không liên quan đáng kể đến sức khỏe thể chất.

+ `phys_hlth` vs `ment_hlth` (p-value = 0): Có sự khác biệt đáng kể, sức khỏe thể chất và tinh thần liên quan chặt chẽ.

+ `phys_hlth` vs `phys_activity` (p-value = 0): Có sự khác biệt đáng kể, hoạt động thể chất giúp cải thiện sức khỏe thể chất.

Tiến hành kiểm định bằng phương pháp permutation test (alter = "left"), tức kiểm định phía trái ta thu được:

Bảng 12: So sánh kết quả kiểm định giữa các cặp biến

Cặp biến	**p-value**
`sex vs ment_hlth`	1
`sex vs phys_activity`	0
`phys_hlth vs ment_hlth`	1
`phys_hlth vs phys_activity`	1

Kết luận:

+ `sex vs ment_hlth`(p-value = 1): Không có bằng chứng nữ giới có sức khỏe tinh thần kém hơn nam giới, dù phân tích trước cho thấy số ngày không tốt cao hơn ở nữ.

+ `sex vs phys_activity`(p-value = 0): Nam giới vận động nhiều hơn nữ giới, phù hợp với các phân tích trước đó.

+ `phys_hlth vs ment_hlth`(p-value = 1): Không có bằng chứng sức khỏe thể chất tốt hơn tinh thần, dù phân tích trước cho thấy có mối liên hệ.

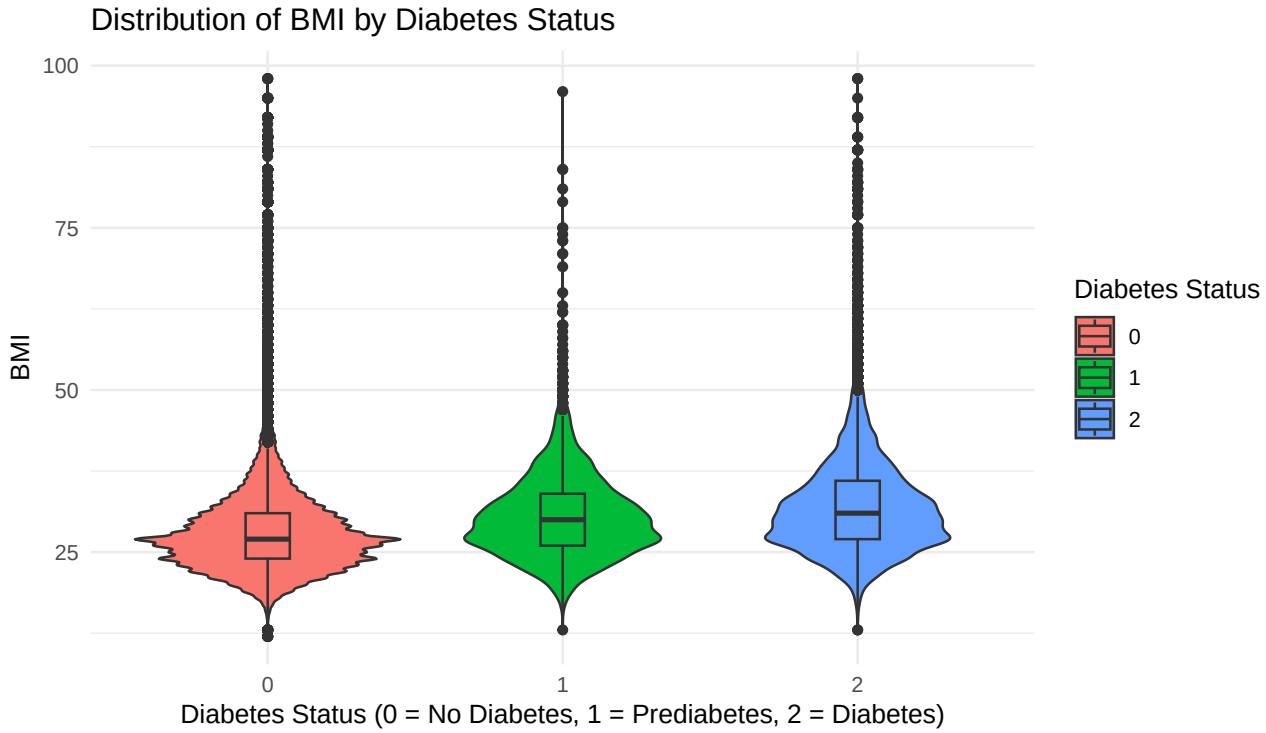
+ `phys_hlth vs phys_activity` (p-value = 1): Không thấy mối liên hệ mạnh giữa vận động và sức khỏe thể chất, mâu thuẫn với các quan sát trước

d) Xét sự ảnh hưởng của `bmi` lên biến mục tiêu

Ta xét bảng thống kê và biểu đồ violin plot giữa biến mục tiêu và BMI:

Bảng 13: Bảng thống kê giữa biến mục tiêu và BMI

Nhóm	**Tiểu đường**	**Không Tiểu đường**	**Tiền tiểu đường**
Trung bình `bmi`	30.5	25.8	28.4
Độ lệch chuẩn `bmi`	6.2	4.7	5.3



Nhận xét :

+ BMI tăng rõ rệt từ không tiểu đường → tiền tiểu đường → tiểu đường, thể hiện BMI là yếu tố nguy cơ mạnh đối với bệnh tiểu đường.

+ Mức độ chênh lệch BMI: Chênh lệch lớn hơn giữa không tiểu đường và tiền tiểu đường, cho thấy giai đoạn đầu của rối loạn chuyển hóa ảnh hưởng mạnh đến BMI.

+ BMI là có thể là yếu tố nguy cơ liên quan chặt chẽ đến nguy cơ từ không tiểu đường đến tiền tiểu đường và tiểu đường.

Tiến hành dùng ANOVA permutation với `bmi` và biến mục tiêu, ta thu được:

Kết quả kiểm định: $p\_value = 0$

Ta bác bỏ giả thuyết "Không có sự khác biệt ý nghĩa thống kê về giá trị BMI trung bình giữa các nhóm (Không tiểu đường, Tiền tiểu đường, và Tiểu đường)".

Kết quả này phù hợp với nhận xét từ bảng thống kê trước đó, khi BMI trung bình tăng rõ rệt từ nhóm không tiểu đường → tiền tiểu đường → tiểu đường.

2.3 Kết luận cuối cùng:

Tất cả các biến đều có sự khác nhau và có độ ảnh hưởng lớn đến biến mục tiêu, trừ giới tính.

BMI khác biệt rõ rệt giữa 3 nhóm không bệnh, tiền tiểu đường và tiểu đường.

Có thể có sự liên hệ, tương quan nhẹ giữa các biến độc lập được xét.

Vậy ta có thể rút ra : BMI cao, thiếu vận động thể chất, bệnh nền và lối sống (sinh hoạt, ăn uống) là những yếu tố liên quan đến nguy cơ mắc bệnh tiểu đường.

3 A/B Testing (Resampling Method)

3.1 Ý tưởng

Phân chia dữ liệu xử lý làm 3 phần:

1. **Các yếu tố liên quan đến sức khỏe:** Phân tích các yếu tố ảnh hưởng trực tiếp đến sức khỏe có thể góp phần làm tăng nguy cơ mắc bệnh tiểu đường (`high_bp`,`high_chol`...).
2. **Thói quen và bệnh dẫn đến nguy cơ mắc bệnh tiểu đường:** Đánh giá một số thói quen (như hút thuốc, đã từng bị đột quỵ) và bệnh lý (như bệnh tim mạch vành (CHD) hoặc nhồi máu cơ tim (MI)) liên quan đến khả năng mắc bệnh.
3. **Ảnh hưởng của các yếu tố kinh tế - xã hội:** Xem xét tác động của các yếu tố kinh tế - xã hội (như thu nhập, giáo dục, tuổi) đến tỷ lệ mắc bệnh tiểu đường.

Ta tính tỷ lệ giữa 1 biến ta xét với `diabetes_012` trước rồi mới thực hiện trực quan. Trực quan hóa dữ liệu để đánh giá mức độ liên quan giữa biến `diabetes_012` với các biến khác trong tập dữ liệu. Ví dụ:

- `diabetes_012` với `high_bp` (huyết áp cao),
- `diabetes_012` với `high_chol` (cholesterol cao),
- `diabetes_012` với `income` (thu nhập)...

Cuối cùng, thực hiện phương pháp **Resampling Method** để xác định xem giữa hai biến có mối liên quan với nhau hay không.

3.2 Loading and Processing Data

Đọc dữ liệu

Để bắt đầu phân tích, chúng ta đã tải dữ liệu từ tệp
diabetes_012_health_indicators_BRFSS2015.csv.

Dữ liệu được xử lý bằng cách sử dụng các phương pháp làm sạch và chọn lọc để giữ lại những biến quan trọng, bao gồm các yếu tố liên quan đến bệnh tiểu đường như huyết áp cao, cholesterol cao, chỉ số BMI, thói quen hút thuốc, tiền sử đột quỵ, bệnh tim mạch, và các yếu tố kinh tế - xã hội như thu nhập, tuổi tác và trình độ học vấn.

Sau khi đọc dữ liệu, chúng ta đã làm sạch dữ liệu bằng cách loại bỏ các giá trị trùng lặp và kiểm tra cấu trúc của dữ liệu để đảm bảo không có vấn đề gì với các bản ghi. Cụ thể, các giá trị trùng lặp đã được loại bỏ để đảm bảo tính chính xác của dữ liệu đầu vào.

Chuyển đổi dữ liệu

Các bước tiếp theo là chuyển đổi các biến phân loại thành kiểu dữ liệu `factor` trong R để dễ dàng phân tích và trực quan hóa. Các biến được chuyển đổi bao gồm:

- `high_bp`: Chuyển đổi giá trị 0 và 1 thành các nhãn "No High BP" và "High BP" tương ứng.
- `high_chol`: Tương tự, chuyển đổi giá trị 0 và 1 thành "No High Chol" và "High Chol".
- `chol_check`: Chuyển thành "No Check" và "Checked".
- `phys_activity`: Chuyển thành "No" và "Yes" dựa trên thói quen hoạt động thể chất.
- `diabetes_012`: Phân loại thành ba mức độ "No Diabetes", "Pre-Diabetes" và "Diabetes".
- `smoker`: Chuyển thành "No Smoke" và "Smoked".
- `stroke`: Chuyển thành "No Stroke" và "Stroked".
- `heart_diseaseor_attack`: Được phân loại thành "No CHD or No MI" và "CHD or MI".
- `age`: Chuyển đổi thành các nhóm tuổi từ 18 - 24 đến 80 hoặc hơn.

- `education`: Chuyển đổi thành các mức độ giáo dục từ "Never attended school" đến "College graduate".
- `income`: Phân loại thu nhập từ "Less than $10,000" đến "$75,000 or more".
- `sex`: Chuyển thành hai giá trị "Female" và "Male".
- `gen_hlth`: Phân loại sức khỏe chung thành các mức "Excellent", "Very Good", "Good", "Fair", và "Poor".

Rows	229,781
Columns	14
`diabetes_012`	`<fct>` No Diabetes, No Diabetes, No Diabetes, No Diabetes, ...
`high_bp`	`<fct>` High BP, No High BP, High BP, High BP, High BP, ...
`high_chol`	`<fct>` High Chol, No High Chol, High Chol, No High Chol, ...
`chol_check`	`<fct>` Checked, No Check, Checked, Checked, Checked, ...
`bmi`	`<dbl>` 40, 25, 28, 27, 24, 25, 30, 25, 30, 24, ...
`smoker`	`<fct>` Smoked, Smoked, No Smoke, No Smoke, No Smoke, ...
`stroke`	`<fct>` No Stroke, No Stroke, No Stroke, No Stroke, ...
`heart_diseaseor_attack`	`<fct>` No CHD or No MI, No CHD or No MI, No CHD or No MI, ...
`phys_activity`	`<fct>` No, Yes, No, Yes, Yes, Yes, No, Yes, No, ...
`sex`	`<fct>` Female, Female, Female, Female, Female, Male, ...
`age`	`<ord>` Age 60–64, Age 50–54, Age 60–64, Age 70–74, ...
`education`	`<ord>` Grade 12 or GED (High school graduate), College 4 years, ...
`income`	`<ord>` "$15,000 to $20,000", "Less than $10,000", "$75,000 and more", ...
`gen_hlth`	`<fct>` Poor, Good, Poor, Very Good, Very Good, ...

Nhận xét

Dữ liệu `diabetes_012_health_indicators_BRFSS2015.csv` bao gồm 22 biến, trong đó chúng ta đã loại bỏ 8 biến không cần thiết cho phân tích. Sau khi xử lý, dữ liệu đã được chuyển đổi thành các kiểu dữ liệu thích hợp và các biến phân loại đã được mã hóa thành các giá trị có thể hiểu được.

Dữ liệu hiện tại có đầy đủ các thông tin quan trọng cần thiết để phân tích các yếu tố ảnh hưởng đến nguy cơ mắc bệnh tiểu đường, từ các yếu tố sức khỏe đến các yếu tố kinh tế - xã hội. Việc tiền xử lý dữ liệu đã giúp giảm thiểu các vấn đề liên quan đến dữ liệu thiếu hoặc không hợp lệ, giúp chúng ta tiếp tục với các phân tích sâu hơn.

3.3 Thực hiện A/B Testing theo Resampling Method

Phương pháp Resampling sẽ được áp dụng để kiểm tra mối quan hệ giữa các yếu tố sức khỏe và nguy cơ mắc bệnh tiểu đường. Trong quá trình này, chúng ta sẽ sử dụng các kỹ thuật như kiểm định giả thuyết và phân tích thống kê mô tả để đánh giá sự liên kết giữa các biến trong bộ dữ liệu. Sau đây là các bước thực hiện A/B testing.

3.3.1 Tiến hành thống kê suy luận các yếu tố liên quan đến sức khỏe

Thống kê mô tả được thực hiện cho các yếu tố `bmi`, `high_bp`, `high_chol`, `phys_activity` và `diabetes_012`. Dưới đây là một số nhận xét từ kết quả thống kê mô tả:

bmi
Min.: 12.00
1st Qu.: 24.00
Median: 27.00
Mean: 28.69
3rd Qu.: 32.00
Max.: 98.00

high_bp	**high_chol**	**phys_activity**	**diabetes_012**
No High BP: 125,359	No High Chol: 128,273	No: 61,270	No Diabetes: 190,055
High BP: 104,422	High Chol: 101,508	Yes: 168,511	Pre-Diabetes: 4,629
			Diabetes: 35,097

Nhận xét

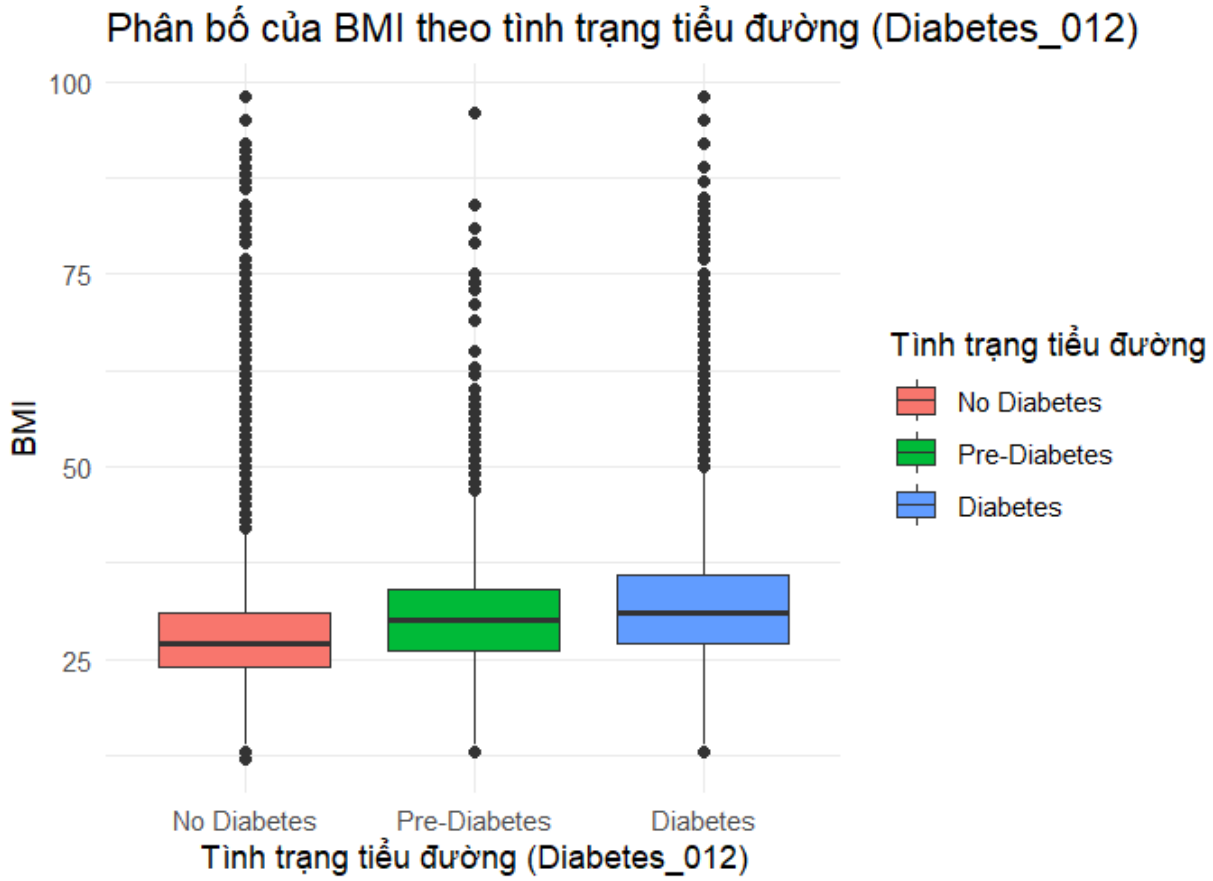
- **Biến `bmi`**: Giá trị `bmi` dao động từ 12.00 đến 98.00, với sự chênh lệch lớn. Trung bình và độ lệch chuẩn tương đối gần nhau.
- **Biến `high_bp`**: Phân chia rõ rệt giữa nhóm có và không có huyết áp cao, cho thấy sự phổ biến của huyết áp cao trong dân cư.
- **Biến `high_chol`**: Tương tự `high_bp`, với tỷ lệ lớn người mắc cholesterol cao, phản ánh nguy cơ tiểu đường.
- **Biến `phys_activity`**: Phân chia rõ rệt giữa nhóm có và không có hoạt động thể chất, ảnh hưởng đến nguy cơ mắc bệnh.
- **Biến `diabetes_012`**: Phân loại ba nhóm với tỷ lệ phân hóa rõ rệt về nguy cơ mắc bệnh tiểu đường.

→ Những nhận xét trên cho thấy sự phân hóa rõ rệt giữa các yếu tố sức khỏe và nguy cơ mắc bệnh tiểu đường trong bộ dữ liệu. Tiếp theo, chúng ta sẽ tiếp tục với các phân tích A/B testing để xác định mối quan hệ giữa các yếu tố này và nguy cơ mắc bệnh tiểu đường.

A/B testing (Resampling Method)

Vì phương pháp Resampling method được học chỉ áp dụng cho biến phản đối là định tính và biến còn lại là định lượng cho nên ở đây, ta chỉ áp dụng với `bmi`, còn các biến khác, ta chỉ vẽ hình tỷ lệ các nhóm và nhận xét sự ảnh hưởng của chúng lên `diabetes_012` .

Kiểm tra mối liên hệ giữa hai biến `diabetes_012` và `bmi`



Hình 1: Phân bố của BMI theo tình trạng tiểu đường (`Diabetes_012`).

Nhận xét: Nhìn vào biểu đồ, ta thấy chỉ số `BMI` ở nhóm mắc bệnh tiểu đường cao hơn hai nhóm còn lại. Điều này cho thấy người có `BMI` cao càng dễ mắc bệnh tiểu đường.

Giả thuyết kiểm định:

- H_0: Không có mối quan hệ giữa `bmi` và tình trạng tiểu đường (`diabetes_012`).
- H_1: Có mối quan hệ giữa `bmi` và tình trạng tiểu đường (`diabetes_012`).

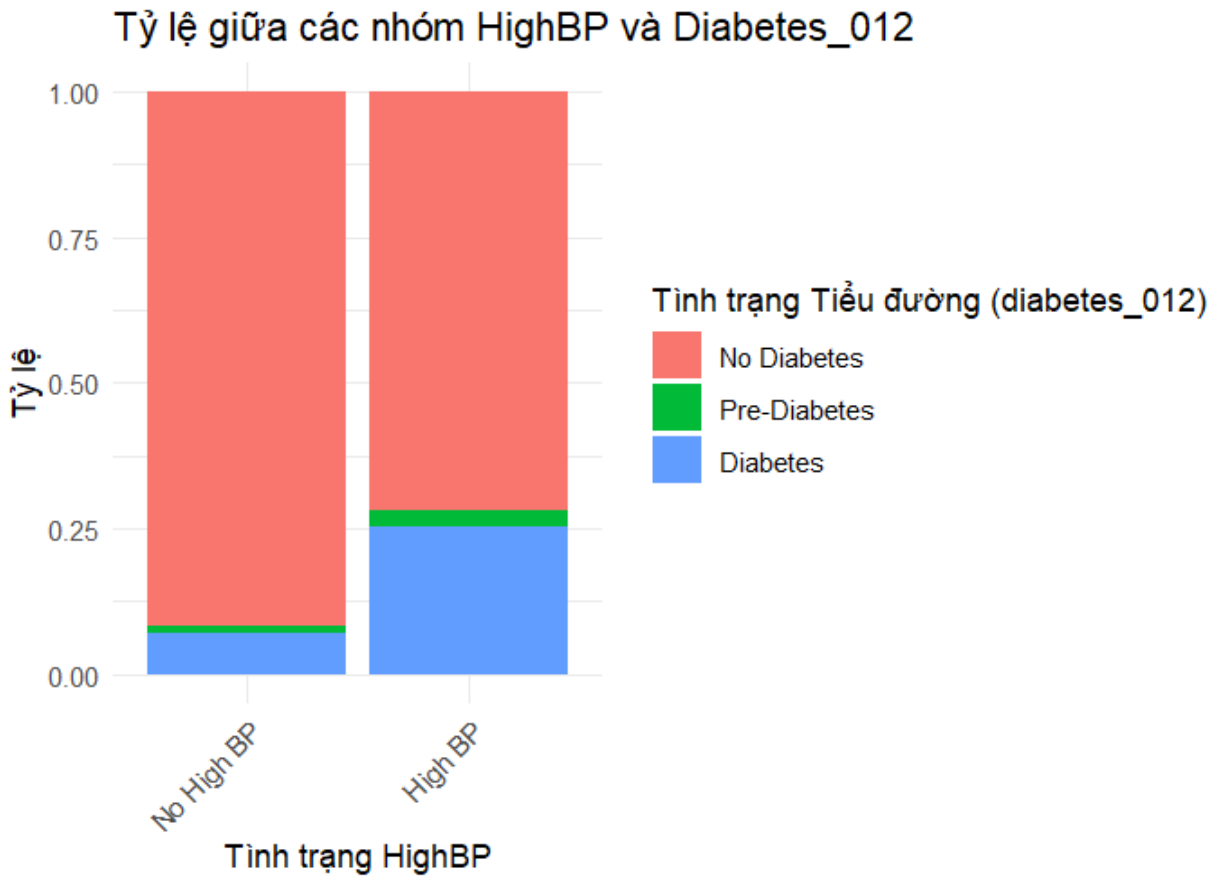
Phương pháp kiểm định: Sử dụng phương pháp kiểm định Resampling để xác định sự khác biệt giữa chỉ số trung bình `bmi` của các nhóm trong `diabetes_012`. Kết quả kiểm định cho ra giá trị *P-value*.

Kết quả:

- Giá trị $P\text{-}value = 0$.
- Vì *P-value* nhỏ hơn mức ý nghĩa thông thường ($\alpha = 0.05$), ta bác bỏ giả thuyết H_0.

Kết luận: Có mối quan hệ giữa chỉ số `bmi` và tình trạng tiểu đường (`diabetes_012`). Chỉ số `BMI` cao là yếu tố nguy cơ quan trọng dẫn đến tình trạng mắc bệnh tiểu đường.

Kiểm tra mối liên hệ giữa hai biến `diabetes_012` và `high_bp`



Hình 2: Biểu đồ tỷ lệ giữa hai nhóm HighBP và Diabetes_012.

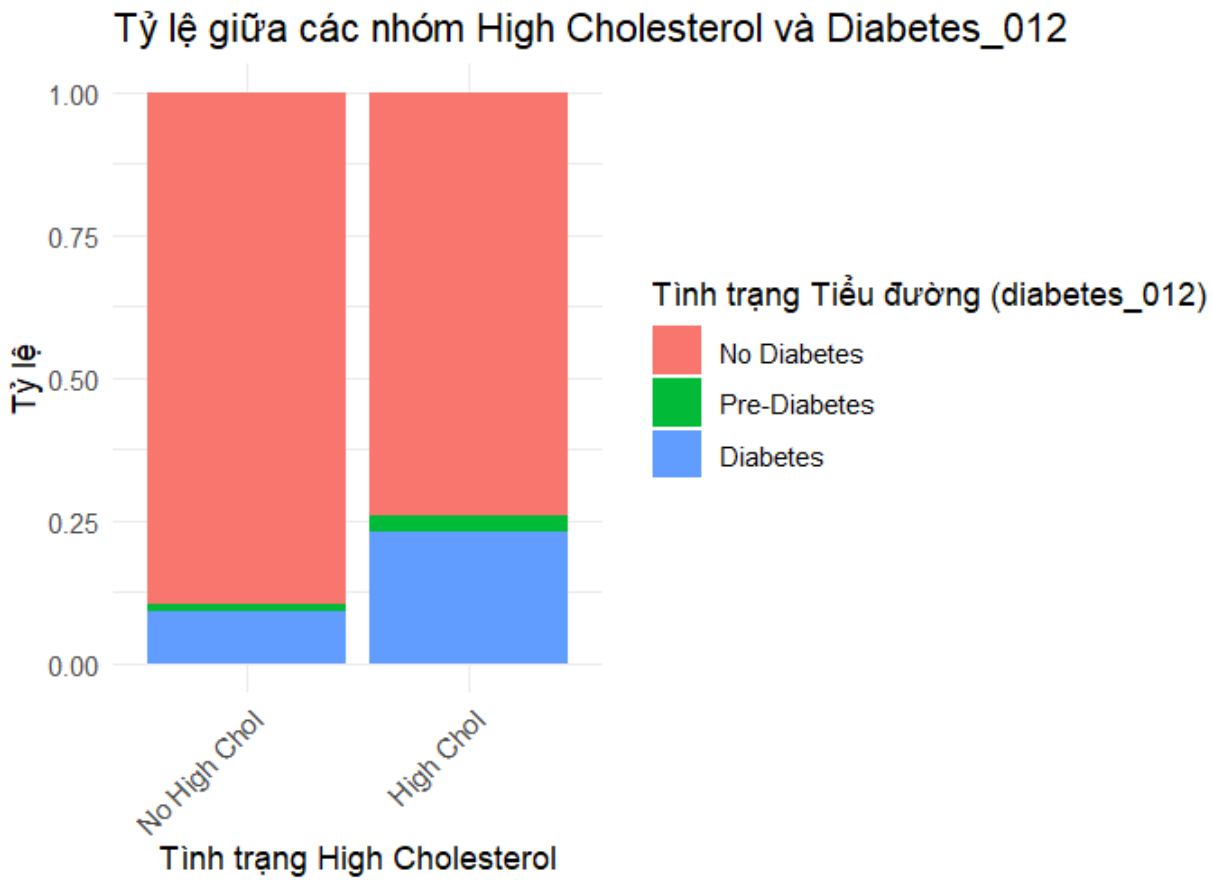
Nhận xét:

- Nhóm `No Diabetes`: tỷ lệ những người không bị bệnh tiểu đường ở nhóm `No High BP` cao hơn so với nhóm `High BP`.
- Nhóm `Pre-Diabetes`: tỷ lệ những người tiền tiểu đường ở nhóm `No High BP` thấp hơn so với nhóm `High BP`.
- Nhóm `Diabetes`: tỷ lệ những người mắc bệnh tiểu đường ở nhóm `High BP` cao hơn so với nhóm `No High BP`.

→ Từ những điều trên ta thấy có mối liên quan giữa tình trạng huyết áp cao và nguy cơ mắc tiểu đường (huyết áp càng cao thì nguy cơ mắc bệnh tiểu đường càng cao).

→ Không dùng được resampling method do phương pháp không áp dụng được cho hai biến định tính.

Kiểm tra mối quan hệ giữa hai biến `diabetes_012` và `high_chol`



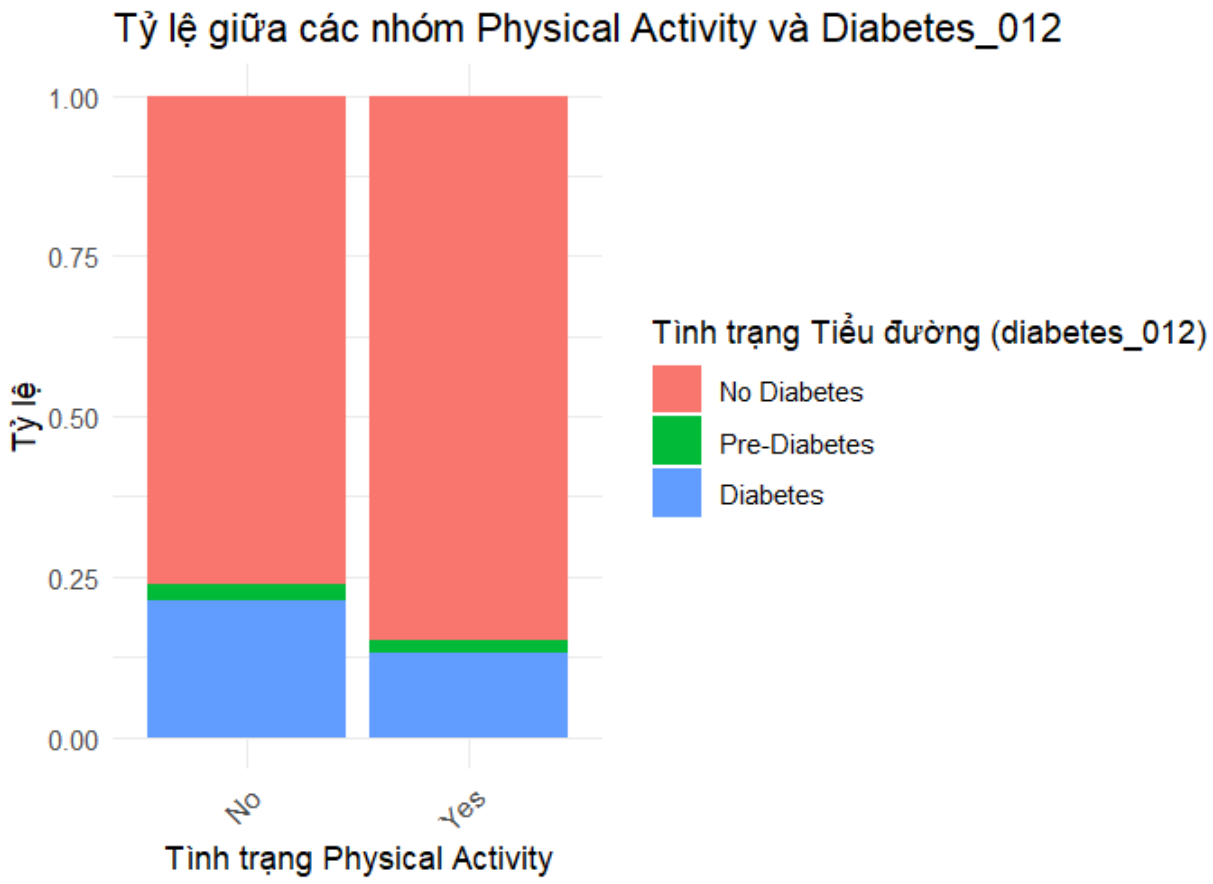
Hình 3: Tỷ lệ giữa các nhóm High Cholesterol và `Diabetes_012`.

Nhận xét:

- **Nhóm No Diabetes**: Tỷ lệ những người không bị bệnh tiểu đường ở nhóm không có cholesterol cao (`No High Chol`) cao hơn so với nhóm có cholesterol cao (`High Chol`).
- **Nhóm Pre-Diabetes**: Tỷ lệ những người tiền tiểu đường ở nhóm `No High Chol` thấp hơn so với nhóm `High Chol`.
- **Nhóm Diabetes**: Tỷ lệ những người mắc bệnh tiểu đường ở nhóm `High Chol` cao hơn so với nhóm `No High Chol`.

Kết luận: Từ những điều trên, ta thấy có mối liên quan giữa tình trạng cholesterol cao và nguy cơ mắc tiểu đường (cholesterol càng cao thì nguy cơ mắc bệnh tiểu đường càng cao). Phương pháp *resampling* không áp dụng được do cả hai biến đều mang tính chất định tính.

Kiểm tra mối quan hệ giữa hai biến `diabetes_012` và `phys_activity`



Hình 4: Tỷ lệ giữa các nhóm Physical Activity và `Diabetes_012`.

Nhận xét:

- **Nhóm No Diabetes**: Tỷ lệ những người không bị bệnh tiểu đường ở nhóm không hoạt động thể chất thấp hơn so với nhóm có hoạt động thể chất.
- **Nhóm Pre-Diabetes**: Tỷ lệ những người tiền tiểu đường ở nhóm không hoạt động thể chất gần như tương đương với nhóm có hoạt động thể chất.
- **Nhóm Diabetes**: Tỷ lệ những người mắc bệnh tiểu đường ở nhóm có hoạt động thể chất thấp hơn so với nhóm không hoạt động thể chất.

Kết luận: Từ những điều trên, ta thấy có mối liên quan giữa nhóm hoạt động thể chất trong vòng 30 ngày và nguy cơ mắc tiểu đường (có hoạt động thể chất thì tỷ lệ mắc bệnh tiểu đường thấp hơn). Phương pháp *resampling* không áp dụng được do cả hai biến đều mang tính chất định tính.

3.3.2 Phân tích một số thói quen và bệnh còn lại dẫn tới nguy cơ bệnh tiểu đường (`bmi`, `smoker`, `heart_diseaseor_attack`, `diabetes_012`)

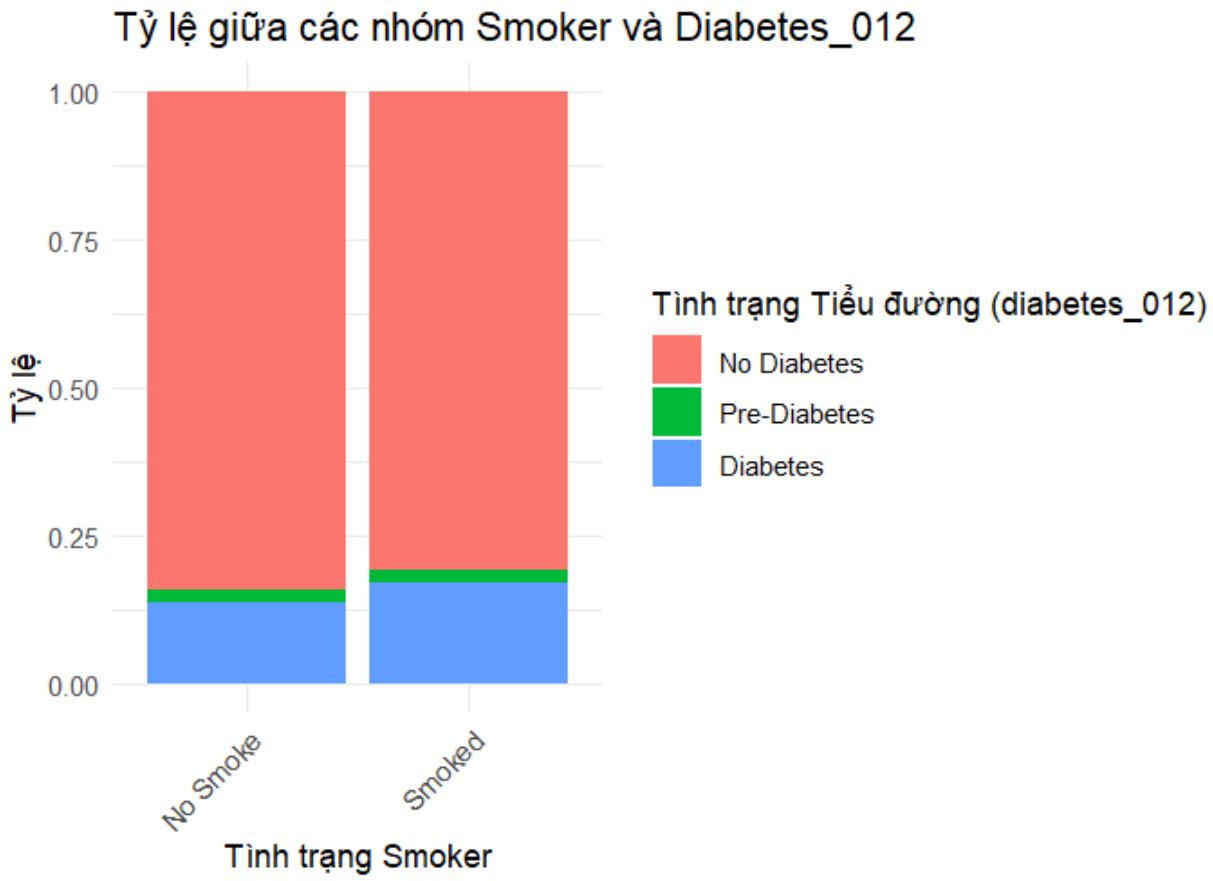
Thống kê mô tả:

smoker	stroke	heart_diseaseor_attack	diabetes_012
No Smoke: 122,781 Smoked: 107,000	No Stroke: 219,497 Stroked: 10,284	No CHD or No MI: 206,064 CHD or MI: 23,717	No Diabetes: 190,055 Pre-Diabetes: 4,629 Diabetes: 35,097

Nhận xét:

- Các biến như `smoker`, `stroke`, `heart_diseaseor_attack`, `diabetes_012` có sự chênh lệch giữa các nhóm trong cùng một thuộc tính.

Kiểm tra mối quan hệ giữa hai biến `diabetes_012` và `smoker`



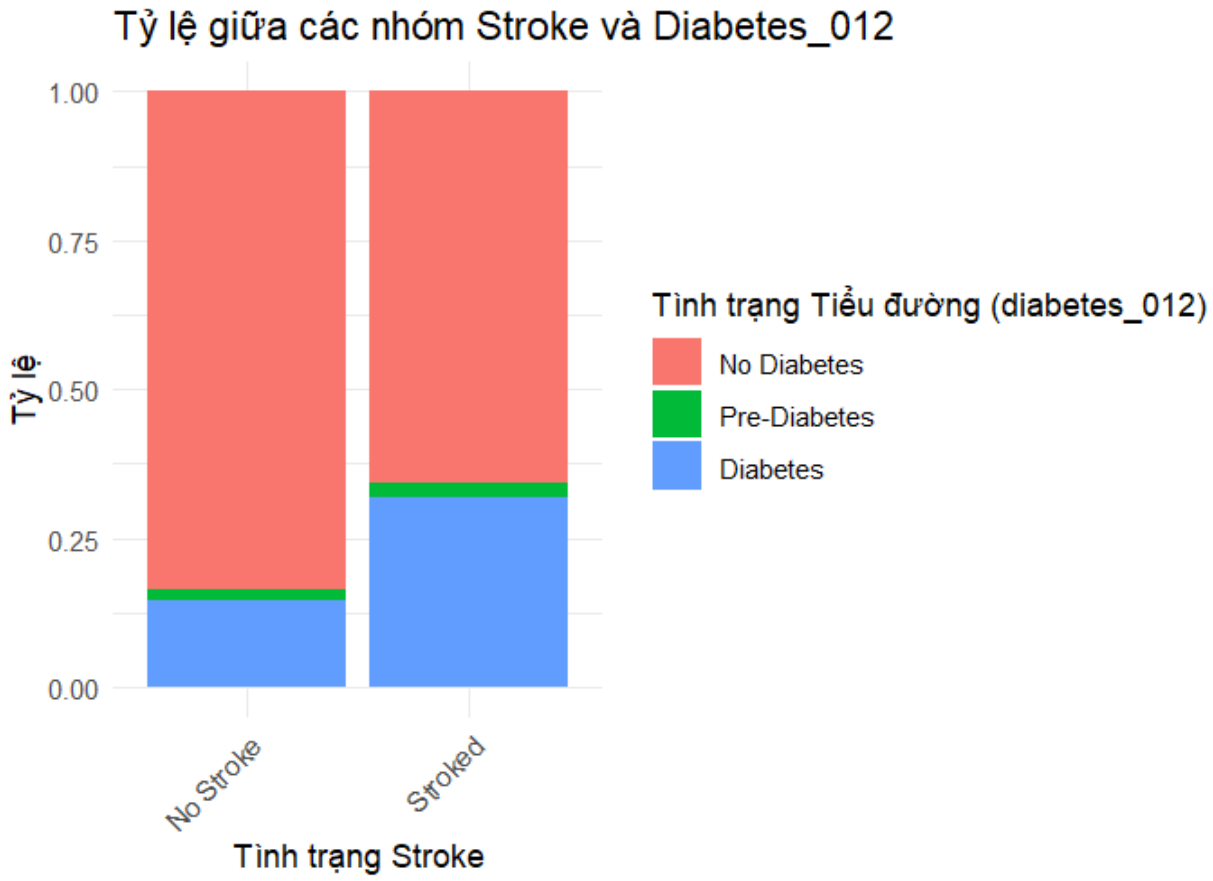
Hình 5: Tỷ lệ giữa các nhóm Smoker và `Diabetes_012`.

Nhận xét:

- **Nhóm No Diabetes**: Tỷ lệ những người không bị bệnh tiểu đường ở nhóm No Smoke cao hơn so với nhóm Smoke.
- **Nhóm Pre-Diabetes**: Tỷ lệ những người tiền tiểu đường ở nhóm No Smoke thấp hơn một chút so với nhóm Smoke nhưng không đáng kể.
- **Nhóm Diabetes**: Tỷ lệ những người mắc bệnh tiểu đường ở nhóm Smoke cao hơn so với nhóm No Smoke.

Kết luận: Có mối liên quan giữa những người đã hút 100 điếu thuốc trong cuộc đời mình và nguy cơ mắc tiểu đường. Hút thuốc càng nhiều thì nguy cơ mắc bệnh tiểu đường càng cao. Không dùng được resampling method do phương pháp không áp dụng được cho hai biến định tính.

Kiểm tra mối quan hệ giữa hai biến `stroke` và `diabetes_012`



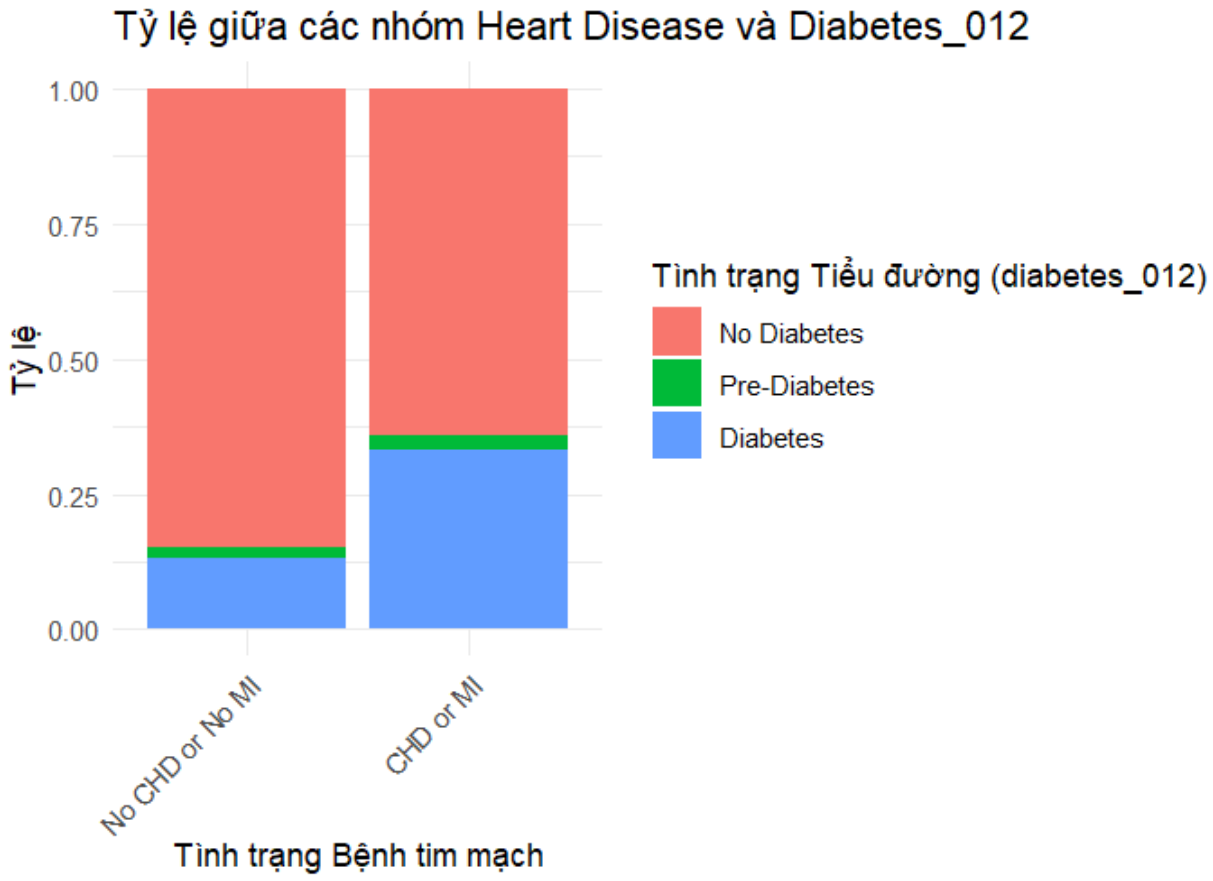
Hình 6: Tỷ lệ giữa các nhóm Stroke và `Diabetes_012`.

Nhận xét:

- **Nhóm No Diabetes**: Tỷ lệ những người không bị bệnh tiểu đường ở nhóm No Stroke cao hơn so với nhóm Stroke.
- **Nhóm Pre-Diabetes**: Tỷ lệ những người tiền tiểu đường ở nhóm No Stroke thấp hơn một chút so với nhóm Stroke nhưng không đáng kể.
- **Nhóm Diabetes**: Tỷ lệ những người mắc bệnh tiểu đường ở nhóm Stroke cao hơn so với nhóm No Stroke.

Kết luận: Có mối liên quan giữa những người từng đột quỵ và nguy cơ mắc tiểu đường. Những người từng đột quỵ có nguy cơ mắc bệnh tiểu đường cao hơn. Không dùng được resampling method do phương pháp không áp dụng được cho hai biến định tính.

Kiểm tra mối quan hệ giữa hai biến `diabetes_012` và `heart_diseaseor_attack`



Hình 7: Tỷ lệ giữa các nhóm Heart Disease và `Diabetes_012`.

Nhận xét:

- **Nhóm No Diabetes**: Tỷ lệ những người không bị bệnh tiểu đường ở nhóm No CHD or No MI cao hơn so với nhóm CHD or MI.

- **Nhóm Pre-Diabetes**: Tỷ lệ những người tiền tiểu đường ở nhóm No CHD or No MI thấp hơn một chút so với nhóm CHD or MI nhưng không đáng kể.
- **Nhóm Diabetes**: Tỷ lệ những người mắc bệnh tiểu đường ở nhóm CHD or MI cao hơn so với nhóm No CHD or No MI.

Kết luận: Có mối liên quan giữa những người mắc bệnh tim mạch vành (CHD) hoặc nhồi máu cơ tim (MI) và nguy cơ mắc tiểu đường. Những người từng mắc một trong hai bệnh trên có nguy cơ mắc bệnh tiểu đường cao hơn. Không dùng được resampling method do phương pháp không áp dụng được cho hai biến định tính.

3.3.3 Phân tích ảnh hưởng của các yếu tố kinh tế - xã hội đến tỷ lệ mắc bệnh tiểu đường (`age`, `education`, `income`, `diabetes_012`)

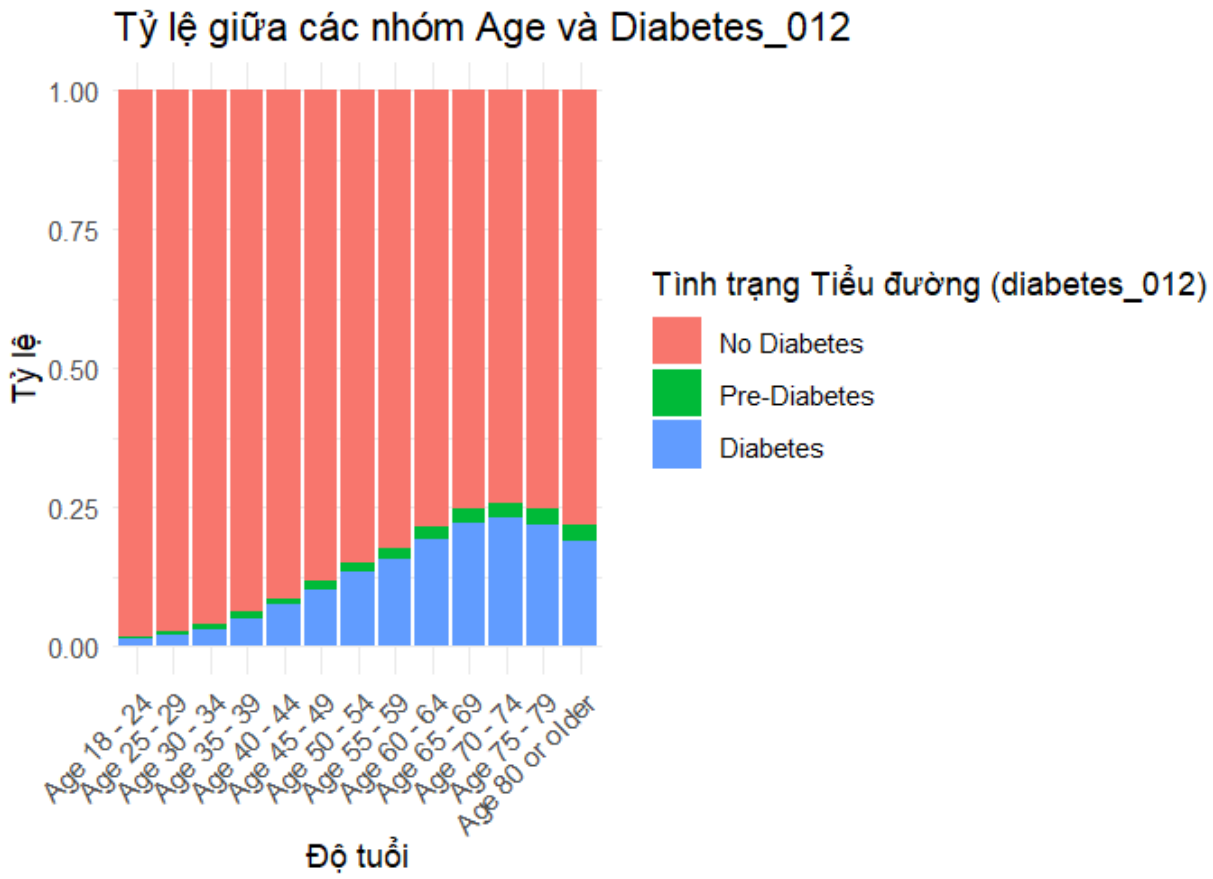
Thống kê mô tả:

age	**education**
Age 60 - 64: 29,736	Never attended school or only kindergarten: 174
Age 65 - 69: 29,168	Grades 1 - 8 (Elementary): 4,040
Age 55 - 59: 27,301	Grades 9 - 11 (Some high school): 9,467
Age 50 - 54: 23,140	Grade 12 or GED (High school graduate): 61,158
Age 70 - 74: 22,041	College 1 year to 3 years (Some college or technical): 66,499
Age 45 - 49: 17,299	College 4 years or more (College graduate): 88,443
(Other): 81,096	
income	**diabetes_012**
$75,000 or more: 71,818	No Diabetes: 190,055
$50,000 to $75,000: 40,189	Pre-Diabetes: 4,629
$35,000 to $50,000: 35,001	Diabetes: 35,097
$25,000 to $35,000: 25,345	
$20,000 to $25,000: 19,957	
$15,000 to $20,000: 15,922	
(Other): 21,549	

Nhận xét:

- **age**: Dữ liệu chủ yếu tập trung vào người trung niên và cao tuổi (45 tuổi trở lên), trong đó nhóm "Other" chiếm tỷ lệ lớn nhất.
- **education**: Phần lớn có trình độ cao (tốt nghiệp đại học hoặc hơn), trong khi các nhóm có học vấn thấp hơn chiếm tỷ lệ nhỏ.
- **income**: Đa số thuộc nhóm thu nhập cao (trên 75.000 USD/năm), nhưng vẫn có đủ sự phân bố ở các mức thu nhập thấp hơn.
- **diabetes_012**: Phần lớn không mắc bệnh, nhưng vẫn có một số lượng người mắc tiểu đường và tiền tiểu đường.

Kiểm tra mối quan hệ giữa hai biến `diabetes_012` và `age`:



Hình 8: Tỷ lệ giữa các nhóm Age và Diabetes_012.

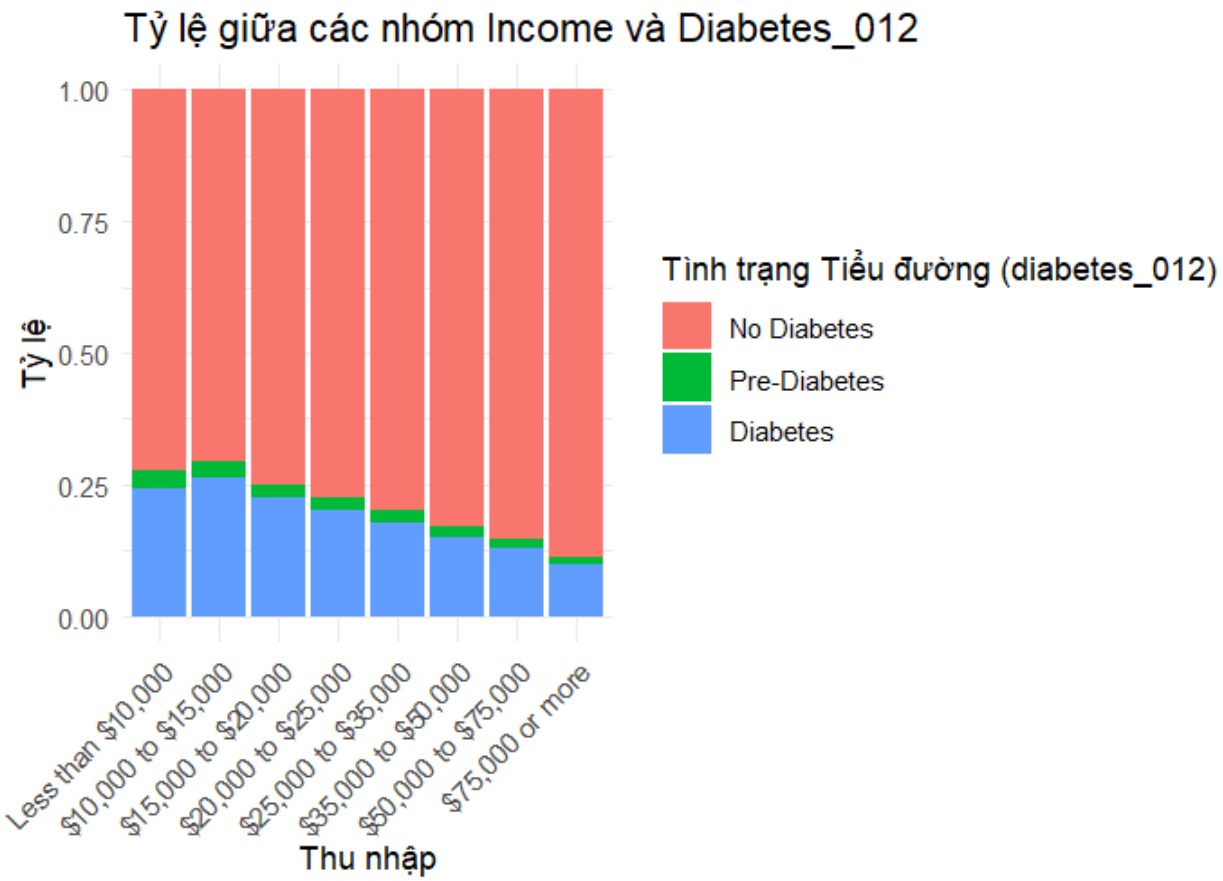
Nhận xét:

- Nhóm No Diabetes: Tỷ lệ những người không bị bệnh tiểu đường có xu hướng giảm dần từ 18 đến 74 tuổi.
- Nhóm Pre-Diabetes: Tỷ lệ những người tiền tiểu đường có xu hướng tăng dần theo độ tuổi nhưng không rõ ràng.
- Nhóm Diabetes: Tỷ lệ những người mắc bệnh tiểu đường có xu hướng tăng dần từ 18 đến 74 tuổi.

→ Từ những điều trên ta thấy có mối liên quan giữa nhóm tuổi và nguy cơ mắc tiểu đường (người càng lớn tuổi thì càng dễ mắc bệnh tiểu đường).

→ Không dùng được resampling method do phương pháp không áp dụng được cho hai biến định tính.

Kiểm tra mối quan hệ giữa hai biến `diabetes_012` và `income`:



Hình 9: Tỷ lệ giữa các nhóm Income và Diabetes_012.

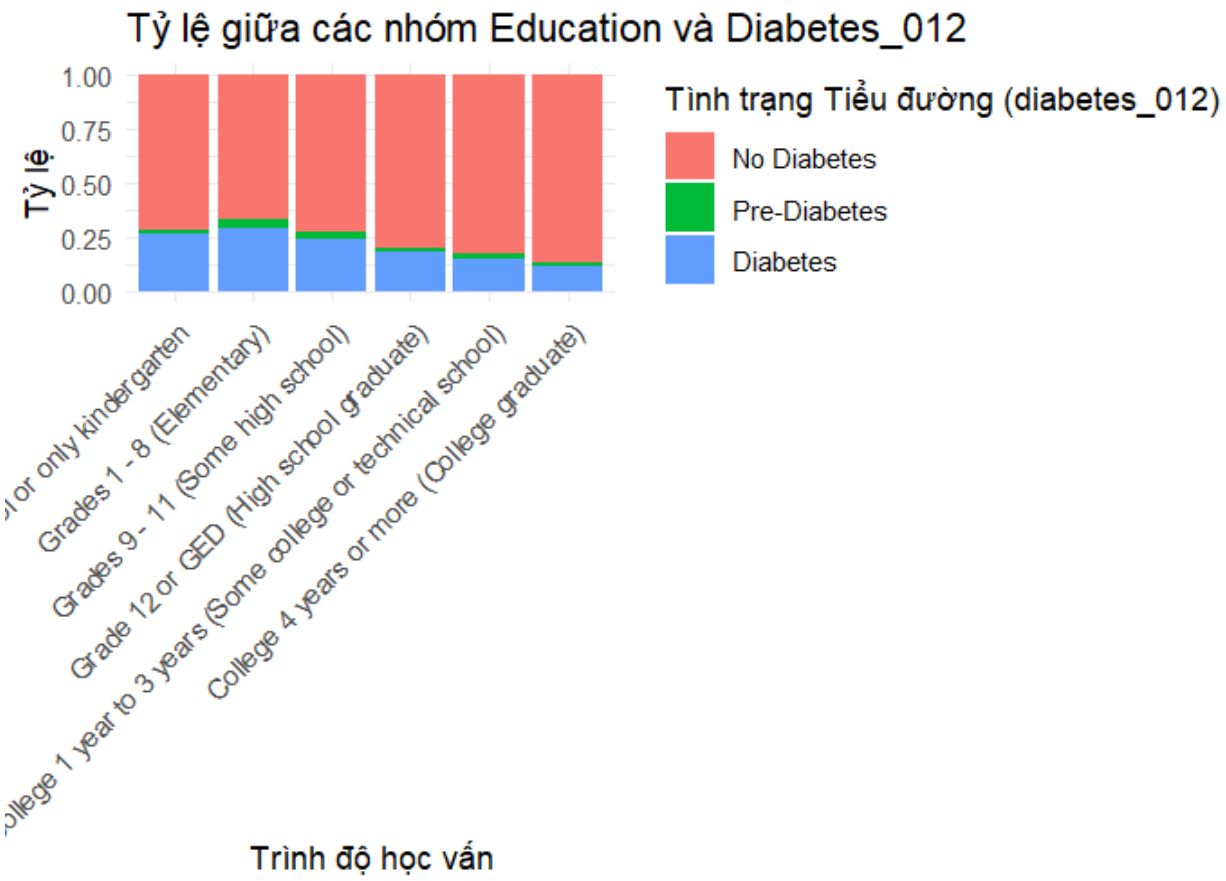
Nhận xét:

- Nhóm No Diabetes: Tỷ lệ những người không bị bệnh tiểu đường có xu hướng tăng bắt đầu từ nhóm thu nhập 10,000$ - 15,000$ trở đi.
- Nhóm Pre-Diabetes: Tỷ lệ những người tiền tiểu đường có xu hướng giảm dần nhưng không rõ ràng.
- Nhóm Diabetes: Tỷ lệ những người mắc bệnh tiểu đường có xu hướng giảm bắt đầu từ nhóm thu nhập 10,000$ - 15,000$ trở đi.

→ Từ những điều trên ta thấy có mối liên quan giữa nhóm thu nhập và nguy cơ mắc tiểu đường (người có thu nhập càng thấp thì có nguy cơ mắc bệnh tiểu đường càng cao).

→ Không dùng được resampling method do phương pháp không áp dụng được cho hai biến định tính.

Kiểm tra mối quan hệ giữa hai biến `diabetes_012` và `education`:



Hình 10: Tỷ lệ giữa các nhóm Education và Diabetes_012.

Nhận xét:

- Nhóm No Diabetes: Tỷ lệ những người không bị bệnh tiểu đường có xu hướng tăng theo trình độ học vấn bắt đầu từ nhóm lớp 1-8 trở đi.
- Nhóm Pre-Diabetes: Tỷ lệ những người tiền tiểu đường có xu hướng giảm dần nhưng không rõ ràng.
- Nhóm Diabetes: Tỷ lệ những người mắc bệnh tiểu đường có xu hướng giảm theo trình độ học vấn bắt đầu từ nhóm lớp 1-8 trở đi.

→ Từ những điều trên ta thấy có mối liên quan giữa nhóm thu nhập và nguy cơ mắc tiểu đường (người có trình độ học vấn càng cao thì có nguy cơ mắc bệnh tiểu đường càng thấp).

→ Không dùng được resampling method do phương pháp không áp dụng được cho hai biến định tính.

3.4 Đánh giá

Dựa vào các biểu đồ và kết quả phân tích, chúng ta nhận thấy rõ mối liên quan giữa các yếu tố như BMI, huyết áp, cholesterol, độ tuổi... với nguy cơ mắc bệnh tiểu đường. Những người có BMI cao, huyết áp cao, và mức cholesterol cao thường có tỷ lệ mắc bệnh tiểu đường cao hơn. Kết quả thống kê cũng chỉ ra rằng các yếu tố này có ảnh hưởng đáng kể đến khả năng mắc bệnh, từ đó giúp xây dựng các biện pháp giúp giảm tỷ lệ hoặc phát hiện bệnh tiểu đường sớm hơn.

4 Mục tiêu ứng dụng các phương pháp phân loại

Xây dựng mô hình phân loại nhằm mục đích sử dụng các thông số sức khỏe cần thiết và tối ưu hóa nhất số lượng biến cần dùng để chẩn đoán được liệu một bệnh nhân đang trong tình trạng nào của bệnh tiểu đường:

0 : Không bị tiểu đường.

1 : Giai đoạn tiền tiểu đường.

2 : Giai đoạn bị tiểu đường.

Như trong phần A/B testing có nhắc đến số lượng các quan sát trong mỗi biến sau khi xử lí dữ liệu lặp, số lượng quán sát trong mỗi nhóm bệnh là như sau:

Bảng 15: Số lượng quan sát trong mỗi nhóm bệnh

Nhóm	**0**	**1**	**2**
Số lượng	190055	4629	35097
tỷ lệ	0.827	0.02	0.152

Ta thấy rằng số lượng giữa các nhóm chênh lệch rất lớn, nhất là nhóm giai đoạn tiền tiểu đường, do đó, nếu xây dựng mô hình hồi quy để chẩn đoán cho cả ba nhóm sẽ là vấn đề khá khó khăn vì số lượng không đủ tốt để mô hình có độ chính xác cao.

Tuy nhiên, mục tiêu là có thể xây dựng mô hình chẩn đoán bệnh, do đó, nhóm sẽ thực hiện mục tiêu này trên hai trường hợp:

I . Trường hợp 1: Xây dựng mô hình chẩn đoán nhóm bệnh trên hai nhóm "Không bị tiểu đường" và "Mắc bệnh tiểu đường" trên cơ sở gộp nhóm bệnh 1 và 2 thành một nhóm,

II . Trường hợp 2: Xây dựng mô hình chẩn đoán nhóm bệnh trên cả ba nhóm,

Ngoài việc xây dựng mô hình phân loại bằng phương pháp *Mô hình hồi quy logistic* và *Mô hình multinominal logistic* ứng với hai trường hợp trên thì nhóm cũng sử dụng các phương pháp phân loại khác như *Phân loại Naive Bayes*, *Phân loại LDA và QDA*.

Hơn nữa, để ứng dụng hiệu quả phương pháp đưa ra, nhóm đã áp dụng phương pháp *Cân bằng dữ liệu* để đưa số lượng các nhóm trong biến phản hồi về trạng thái cân bằng và áp dụng các mô hình phân loại nêu trên để xây dựng mô hình chẩn đoán sau khi cân bằng dữ liệu. Và cuối cùng, đó là đưa ra sự so sánh và đánh giá mức độ hiệu quả.

5 Chọn biến cho các phương pháp phân loại

Để chọn ra được các dữ liệu biến thích hợp dùng cho mô hình phân loại, ta sẽ kết hợp kết quả của ba trường hợp sau:

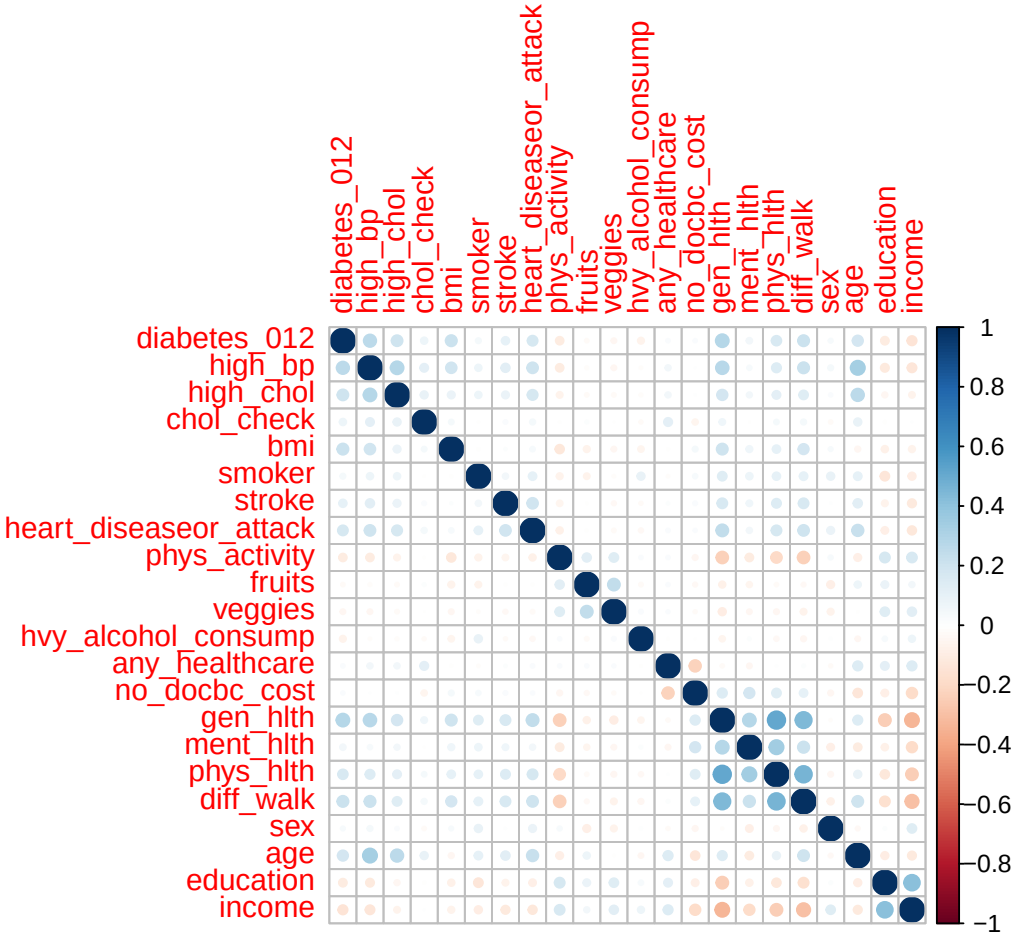
Bảng 16: Các biến được sử dụng từ các trường hợp

Biến	**A/B testing**	**AIC 2 nhóm**	**AIC 3 nhóm**	**Tìm hiểu**
high_bp	x	x	x	x
high_chol	x	x	x	x
chol_check	x	x	x	
bmi	x	x	x	x
smoker	x	x	x	x
stroke	x	x	x	x
heart_diseaseor_attack	x	x	x	x
phys_activity	x	x	x	
fruits				
veggies		x		x
hvy_alcohol_consump	x	x	x	x
any_healthcare		*	x	
no_docbc_cost		*	x	
gen_hlth	x	x	x	x
ment_hlth	x	*	x	x
phys_hlth	x	x	x	x
diff_walk	x	x	x	x
sex		x	x	
age	x	x	x	x
education	x		x	
income	x	x	x	x
Tổng	16	19	19	15

Dấu "x" biểu thị cho các biến là phù hợp, dấu "*" thể hiện rằng các biến có ý nghĩa đóng góp ít hơn hẳn so với các biến được chọn khác.

Và quan sát bảng hệ số tương quan và biểu đồ tương quan sau:

Hình 11: Biểu đồ tương quan



Bảng 17: Hệ số tương quan giữa các biến với diabetes_012

	diabetes_012
high_bp	0.262
high_chol	0.203
chol_check	0.076
bmi	0.212
smoker	0.047
stroke	0.100
heart_diseaseor_attack	0.171
phys_activity	-0.103
fruits	-0.025
veggies	-0.043
hvy_alcohol_consump	-0.067

	diabetes_012
any_healthcare	0.025
no_docbc_cost	0.024
gen_hlth	0.285
ment_hlth	0.058
phys_hlth	0.160
diff_walk	0.211
sex	0.032
age	0.185
education	-0.108
income	-0.147

Qua đó, ta xét thấy những biến phù hợp nhất bao gồm 15 biến sau:

Bảng 18: Các biến được sử dụng từ các trường hợp

high_bp	high_chol	chol_check
bmi	smoker	stroke
heart_diseaseor_attack	phys_activity	hvy_alcohol_consump
gen_hlth	ment_hlth	phys_hlth
diff_walk	age	income

Bảng trên là những biến được chọn qua các A/B testing và góp phần xây dựng mô hình tốt, phù hợp thực tế và có độ tương quan trên ngưỡng ± 0.05.

6 Mô hình logistic

Mô hình hồi quy logistic có dạng như sau:

$$\log\left(\frac{p(\boldsymbol{x})}{1-p(\boldsymbol{x})}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$

Vì có sự xuất hiện của rất nhiều chỉ số `bmi` quá lớn trong dữ liệu nên ta sẽ xây dựng mô hình hồi quy logistic với hai dữ liệu: một dữ liệu gốc và dữ liệu của những bệnh nhân có `bmi` dưới 70 và so sánh chúng.

Vì để đảm bảo kiểm tra lại kết quả thu được từ mô hình xây dựng nên ta sẽ chia data gốc thành hai tập train và tập test với tỷ lệ 9:1. Để đảm bảo có sự phân chia ngẫu nhiên nhất thì với mỗi lần xây dựng mô hình, ta sẽ chia ngẫu nhiên lại tập train và tập test một lần nữa.

Đa phần các biến cần trong xây dựng mô hình đều là biến định tính, do đó, việc factor các biến ấy là cần thiết, tuy nhiên với những biến chỉ có hai nhóm, việc factor chúng mang lại kết quả ước lượng của β tương tự với việc chưa factor vì chúng mang giá trị 0 và 1. Vì vậy, trong phần xây dựng mô hình logistic lẫn mô hình multinominal logistic, ta chỉ factor những biến có ba nhóm trở lên như: diabetes_012, `gen_hlth`, `age` và `income`.

6.1 Dữ liệu gốc

6.1.1 Ước lượng mô hình

Với việc sử dụng 15 biến đã chọn trong phần chọn biến, mô hình xây dựng được các ước lượng, trong đó có biến `smoker` có p_value$= 0.08 > \alpha = 0.05$. Hơn nữa, dựa vào tìm hiểu thực tế thì hút thuốc nó có thể có ảnh hưởng một cách nào đó đến tim mạch, nhưng ta không thể khẳng định bệnh nhân bị các bệnh về tim do hút thuốc và cũng có một biến nói lên trạng thái về bệnh tim của bệnh nhân nên ta sẽ bỏ qua biến `smoker`. Thực hiện hồi quy lại lần nữa, ta có kết quả ước lượng trong bảng (19).

Bảng 19: Hệ số hồi quy của mô hình hồi quy logistic 1

Variable	**Estimate**	**Std. Error**	**z value**	**Pr(>|z|)**
Intercept	-7.501	0.134	-56.177	< 2e-16
high_bp	0.666	0.015	45.524	< 2e-16
high_chol	0.544	0.014	39.774	< 2e-16
chol_check	1.204	0.065	18.549	< 2e-16
bmi	0.056	0.001	61.101	< 2e-16
stroke	0.125	0.026	4.897	9.73e-07
heart_diseaseor_attack	0.258	0.018	14.335	< 2e-16
phys_activity	-0.030	0.014	-2.079	0.0376
hvy_alcohol_consump	-0.716	0.037	-19.559	< 2e-16
gen_hlth2	0.598	0.032	18.657	< 2e-16
gen_hlth3	1.207	0.031	38.673	< 2e-16
gen_hlth4	1.656	0.034	48.680	< 2e-16
gen_hlth5	1.812	0.041	43.726	< 2e-16
ment_hlth	-0.003	0.001	-3.487	0.00049
phys_hlth	-0.004	0.001	-4.954	7.28e-07
diff_walk	0.116	0.017	6.809	9.81e-12
age2	0.186	0.134	1.390	0.1644
age3	0.402	0.121	3.318	0.00091
age4	0.819	0.115	7.124	1.05e-12
age5	1.032	0.112	9.177	< 2e-16
age6	1.282	0.111	11.595	< 2e-16
age7	1.464	0.109	13.390	< 2e-16
age8	1.550	0.109	14.227	< 2e-16
age9	1.768	0.109	16.269	< 2e-16
age10	1.915	0.109	17.626	< 2e-16
age11	1.944	0.109	17.825	< 2e-16
age12	1.874	0.110	17.071	< 2e-16
age13	1.710	0.110	15.559	< 2e-16
income2	-0.036	0.036	-1.012	0.3115
income3	-0.094	0.035	-2.722	0.0065
income4	-0.115	0.034	-3.408	0.00065
income5	-0.194	0.033	-5.868	4.41e-09
income6	-0.246	0.032	-7.627	2.41e-14
income7	-0.274	0.032	-8.480	< 2e-16
income8	-0.360	0.031	-11.487	< 2e-16

Nhận xét: Ta thấy rằng ngoại trừ ước lượng cho nhóm tuổi age2 và lương income2 thì các ước lượng khác đều có p-value rất bé, do đó, các biến dùng để hồi quy cho diabetes_012 đều có ý nghĩa thống kê với mức ý nghĩa $\alpha = 5\%$ đối với mô hình logistic, và nó hữu ích trong việc phân loại các đối tượng mắc bệnh tiểu đường. Tuy nhiên với nhóm bệnh nhân có

nhóm `age2` hoặc `income2` hoặc cả hơn thì cần cân nhắc hơn khi sử dụng mô hình này.

Về ý nghĩa của mỗi ước lượng, ta sẽ tính tỷ lệ cược - odds cho mỗi biến trong bảng (23).

Bảng 20: Tỷ lệ cược - Odds của từng biến mô hình hồi quy logistic 1

Biến	Tỷ lệ cược
(Intercept)	0.0006
high_bp	1.9466
high_chol	1.7221
chol_check	3.3333
bmi	1.0581
stroke	1.1336
heart_diseaseor_attack	1.2949
phys_activity	0.9707
hvy_alcohol_consump	0.4886
gen_hlth2	1.8184
gen_hlth3	3.3443
gen_hlth4	5.2372
gen_hlth5	6.1220
ment_hlth	0.9970
phys_hlth	0.9960
diff_walk	1.1231

Biến	Tỷ lệ cược
age2	1.2040
age3	1.4950
age4	2.2673
age5	2.8058
age6	3.6036
age7	4.3247
age8	4.7117
age9	5.8616
age10	6.7899
age11	6.9866
age12	6.5129
age13	5.5280
income2	0.9642
income3	0.9101
income4	0.8914
income5	0.8235
income6	0.7817
income7	0.7603
income8	0.6976

Nhận xét:

Đối với các biến có tỷ lệ cược lớn hơn 1:

- `high_bp` (tỷ lệ cược = 1.9466): Khi mức huyết áp cao tăng lên một đơn vị, nguy cơ mắc bệnh tiểu đường sẽ tăng lên khoảng 94.66%. Điều này cho thấy huyết áp cao là yếu tố nguy cơ quan trọng đối với bệnh tiểu đường.
- `high_chol` (tỷ lệ cược = 1.7221): Khi mức cholesterol cao tăng lên một đơn vị, nguy cơ mắc bệnh tiểu đường sẽ tăng lên khoảng 72.21%. Điều này cho thấy cholesterol là yếu tố nguy cơ quan trọng đối với bệnh tiểu đường.
- `chol_check` (tỷ lệ cược = 3.3333): Những người kiểm tra mức cholesterol thường xuyên có nguy cơ mắc bệnh tiểu đường cao gấp khoảng 3.33 lần so với những người không kiểm tra. Điều này có thể ám chỉ rằng việc kiểm tra cholesterol thường xuyên giúp phát hiện bệnh hoặc có liên hệ với việc các cá nhân có nguy cơ cao mắc bệnh tiểu đường được giám sát sức khỏe kỹ hơn.
- `bmi` (tỷ lệ cược = 1.0581): Tăng 1 đơn vị `bmi` sẽ làm nguy cơ mắc bệnh tiểu đường tăng lên 5.81%. Điều này chứng tỏ việc duy trì một chỉ số `bmi` khỏe mạnh đóng vai trò quan trọng trong việc phòng ngừa bệnh tiểu đường.

- `stroke` (tỷ lệ cược = 1.1336): Những người đã từng bị đột quỵ có nguy cơ mắc bệnh tiểu đường cao hơn 13.36% so với những người chưa bị đột quỵ.
- `heart_diseaseor_attack` (tỷ lệ cược = 1.2949): Những người có tiền sử bệnh tim có khả năng mắc bệnh tiểu đường cao hơn 29.49%.
- Nhóm `gen_hlth`: Các biến sức khỏe tổng quát (`gen_hlth2` đến `gen_hlth5`) có tỷ lệ cược lớn hơn 1, phản ánh rằng các mức đánh giá sức khỏe từ 1 đến 5 làm tăng nguy cơ mắc bệnh tiểu đường rõ rệt. Cụ thể, nguy cơ này gia tăng từ 81.84% (`gen_hlth2`) đến 512.20% (`gen_hlth5`), cho thấy sức khỏe tổng quát kém là một yếu tố nguy cơ quan trọng.
- `diff_walk` (tỷ lệ cược = 1.1231): Những người có khó khăn trong việc di chuyển có nguy cơ mắc bệnh tiểu đường cao hơn 12.31% so với những người không gặp vấn đề về di chuyển.
- Nhóm `age`: Các biến tuổi (`age2` đến `age13`) thể hiện sự gia tăng nguy cơ mắc bệnh tiểu đường theo tuổi tác. Cụ thể, từ độ tuổi `age2` (nguy cơ tăng 20.40%) đến độ tuổi `age10` (nguy cơ tăng 578.99%), điều này cho thấy rằng người lớn tuổi có nguy cơ cao mắc bệnh tiểu đường hơn so với những người trẻ tuổi. Tuy nhiên, nguy cơ này giảm nhẹ ở độ tuổi `age13` (nguy cơ tăng 452.80%), mặc dù vẫn rất cao đối với người cao tuổi.

Đối với các biến có tỷ lệ cược bé hơn 1:

Các biến `ment_hlth`, và `phys_hlth` có tỷ lệ cược khá gần 1, điều này có thể cho thấy các yếu tố này không phải là yếu tố quyết định mạnh mẽ trong việc làm tăng hoặc giảm nguy cơ mắc bệnh tiểu đường. Đồng thời, các yếu tố liên quan đến thu nhập thấp (`income2` đến `income8`) cho thấy khả năng mắc bệnh giảm dần khi thu nhập giảm, điều này có thể là một yếu tố phụ, nhưng không có ảnh hưởng mạnh mẽ.

Nhìn chung, các kết quả này cho thấy rằng các yếu tố sức khỏe như huyết áp, cholesterol, bệnh về tim mạch, và tuổi tác có ảnh hưởng lớn hơn nhiều đến nguy cơ mắc bệnh tiểu đường so với các yếu tố lối sống và kinh tế.

6.1.2 Tiên đoán

Dự đoán mô hình với ngưỡng 0.5 **trên tập test**: kết quả thu được cho dự báo mô hình của dữ liệu gốc trên tập test là 83.50% tỉ lệ phân loại đúng hai nhóm của dữ liệu. Đây là một tỉ lệ dự đoán khá tốt tuy nhiên với tỉ lệ nhóm 0 xuất hiện hơn 80% trong dữ liệu thì đây vẫn là một con số mang nhiều nghi vấn, ta lấy ngẫu nhiên 10 giá trị đầu trong tập test và so sánh với dự đoán của mô hình:

Bảng 21: Bảng nhóm tiên đoán và dữ liệu thực của mô hình hồi quy logistic

Nhóm tiên đoán	Dữ liệu thực
0	0
0	0
0	0
0	0
0	0
0	0
1	1
0	1
0	1
1	0

Nhận xét: kết quả dự đoán có xu hướng dự đoán nghiêng về nhóm 0 nhiều hơn. Vì thế, để cải thiện phần nào vấn đề này, nhất là vấn đề về cân bằng tỷ lệ giữa các nhóm trong dữ liệu, ta dùng phương pháp cân bằng mẫu để xem sự cải thiện về tiên đoán của mô hình ở phần sau.

6.2 Dữ liệu với `bmi` < 70

6.2.1 Ước lượng mô hình

Ta bỏ đi hai biến `smoker` và `phys_activity` do p-value trong trường hợp này vượt trên mức ý nghĩa 5%, ta có kết quả ước lượng trong bảng (22).

Bảng 22: Hệ số hồi quy của mô hình hồi quy logistic 2

Variable	Estimate	Std. Error	z value	Pr(>|z|)
Intercept	-7.822	0.135	-58.129	<2e-16
`high_bp`	0.633	0.015	43.070	<2e-16
`high_chol`	0.545	0.014	39.714	<2e-16
`chol_check`	1.180	0.065	18.292	<2e-16
`bmi`	0.069	0.001	66.772	<2e-16
`stroke`	0.133	0.026	5.142	2.71e-7
`heart_diseaseor_attack`	0.260	0.018	14.386	<2e-16
`hvy_alcohol_consump`	-0.688	0.037	-18.847	<2e-16
`gen_hlth2`	0.586	0.032	18.275	<2e-16
`gen_hlth3`	1.168	0.031	37.347	<2e-16
`gen_hlth4`	1.620	0.034	47.613	<2e-16
`gen_hlth5`	1.796	0.041	43.291	<2e-16
`ment_hlth`	-0.003	0.001	-3.114	0.0018
`phys_hlth`	-0.004	0.001	-4.783	1.72e-6
`diff_walk`	0.085	0.017	4.946	7.56e-7
`age2`	0.195	0.135	1.438	0.150

Variable	Estimate	Std. Error	z value	Pr(>|z|)
age3	0.407	0.123	3.301	0.001
age4	0.804	0.117	6.871	6.39e-12
age5	1.024	0.115	8.934	<2e-16
age6	1.288	0.113	11.430	<2e-16
age7	1.476	0.112	13.233	<2e-16
age8	1.576	0.111	14.190	<2e-16
age9	1.792	0.111	16.170	<2e-16
age10	1.934	0.111	17.449	<2e-16
age11	2.002	0.111	18.007	<2e-16
age12	1.934	0.112	17.294	<2e-16
age13	1.773	0.112	15.834	<2e-16
income2	-0.078	0.036	-2.170	0.030
income3	-0.112	0.035	-3.228	0.001
income4	-0.158	0.034	-4.647	3.36e-6
income5	-0.214	0.033	-6.440	1.19e-10
income6	-0.287	0.032	-8.867	<2e-16
income7	-0.314	0.032	-9.701	<2e-16
income8	-0.391	0.031	-12.444	<2e-16

Nhận xét: Ta có nhận xét giống với mô hình với dữ liệu gốc, tuy nhiên, ở mô hình này biến `phys_activity` lại không có ý nghĩa thống kê.

Về ý nghĩa của mỗi ước lượng, ta sẽ tính tỷ lệ cược - odds cho mỗi biến trong bảng (23).

Bảng 23: Tỷ lệ cược - Odds của từng biến mô hình hồi quy logistic 2

Variable	Odds Ratio
(Intercept)	0.0004
high_bp	1.8826
high_chol	1.7245
chol_check	3.2553
bmi	1.0718
stroke	1.1418
heart_diseaseor_attack	1.2976
hvy_alcohol_consump	0.5025
gen_hlth2	1.7968
gen_hlth3	3.2140
gen_hlth4	5.0550
gen_hlth5	6.0278
ment_hlth	0.9973
phys_hlth	0.9961
diff_walk	1.0882
age2	1.2151
age3	1.5025

Variable	Odds Ratio
age4	2.2353
age5	2.7833
age6	3.6261
age7	4.3745
age8	4.8378
age9	6.0036
age10	6.9148
age11	7.4037
age12	6.9205
age13	5.8901
income2	0.9245
income3	0.8940
income4	0.8542
income5	0.8075
income6	0.7506
income7	0.7303
income8	0.6767

Nhận xét: Các tỷ số cược trong mô hình này không sai khác quá nhiều so với mô hình trước nên ta sẽ không nhận xét lặp lại.

6.2.2 Tiên đoán

Dự đoán mô hình trên tập test: Kết quả thu được cho dự báo mô hình với ngưỡng 0.5 trên tập test là 83.60% tỉ lệ phân loại đúng hai nhóm của dữ liệu. Đây là một tỉ lệ dự đoán khá tốt so với mô hình trên dữ liệu gốc, do đó, việc các chỉ số `bmi` xuất hiện quá lớn và quá bất thường không quá ảnh hưởng đến kết quả tiên đoán. Tuy nhiên, giống với mô hình ban đầu, đây là một con số mang nhiều nghi vấn với tỉ lệ nhóm 0 xuất hiện hơn 80% trong dữ liệu.

6.3 Kết luận 1

Vậy ta chọn mô hình với dữ liệu gốc, nhưng xét thấy biến `phys_activity` có mức p-value quá cao nên ta sẽ dùng 13 biến (bỏ `smoker`, `phys_activity`) để hồi quy cho mô hình dữ liệu gốc.

Xét mô hình 13 biến với dữ liệu ban đầu, mô hình hồi quy cho các hệ số ước lượng đều có ý nghĩa thống kê ở mức 5% (trừ `age2` và `income2`), xác suất dự đoán trên tập test là 83.54% và bảng tỷ lệ cược - odds của từng biến trong mô hình như sau:

Theo như bảng tỷ lệ cược, các yếu tố như huyết áp, kiểm tra cholesterol, tuổi tác (đặc biệt là các nhóm tuổi lớn), và sức khỏe tổng quát có sự thay đổi đáng kể trong tỷ lệ cược, cụ thể là tăng, so với mô hình `bmi` < 70 thì khả năng xét vào nhóm nguy cơ bị tiểu đường tăng

Bảng 24: Tỷ lệ cược - Odds của từng biến mô hình hồi quy logistic 3

Variable	Odds Ratio
(Intercept)	0.0005
high_bp	1.9471
high_chol	1.7221
chol_check	3.3286
bmi	1.0583
stroke	1.1335
heart_diseaseor_attack	1.2943
hvy_alcohol_consump	0.4888
gen_hlth2	1.8202
gen_hlth3	3.3521
gen_hlth4	5.2580
gen_hlth5	6.1548
ment_hlth	0.9971
phys_hlth	0.9961
diff_walk	1.1274
age2	1.2051
age3	1.4972

Variable	Odds Ratio
age4	2.2729
age5	2.8135
age6	3.6151
age7	4.3391
age8	4.7273
age9	5.8804
age10	6.8107
age11	7.0118
age12	6.5400
age13	5.5537
income2	0.9644
income3	0.9103
income4	0.8917
income5	0.8234
income6	0.7812
income7	0.7594
income8	0.6961

từ khoảng 10% - 20% trên mỗi biến này, điều này càng khẳng định hơn kết luận ban đầu: các yếu tố sức khỏe như huyết áp, cholesterol, bệnh về tim mạch, và tuổi tác, sức khỏe tổng quát có ảnh hưởng lớn hơn nhiều đến nguy cơ mắc bệnh tiểu đường so với các yếu tố lối sống và kinh tế.

6.4 Đánh giá mô hình

Như kết luận ở phần trước, ta đã chọn mô hình với dữ liệu gốc, bây giờ, ta sẽ đánh giá mô hình này.

Để đánh giá độ chính xác của mô hình logistic, chúng ta sẽ sử dụng phân tích đường cong ROC và diện tích dưới đường cong (AUC).

Chỉ số AUC thu được là:

$$\text{AUC} = 0.8032.$$

Khoảng tin cậy 95% cho AUC là: (0.0.8008; 0.0.8055).

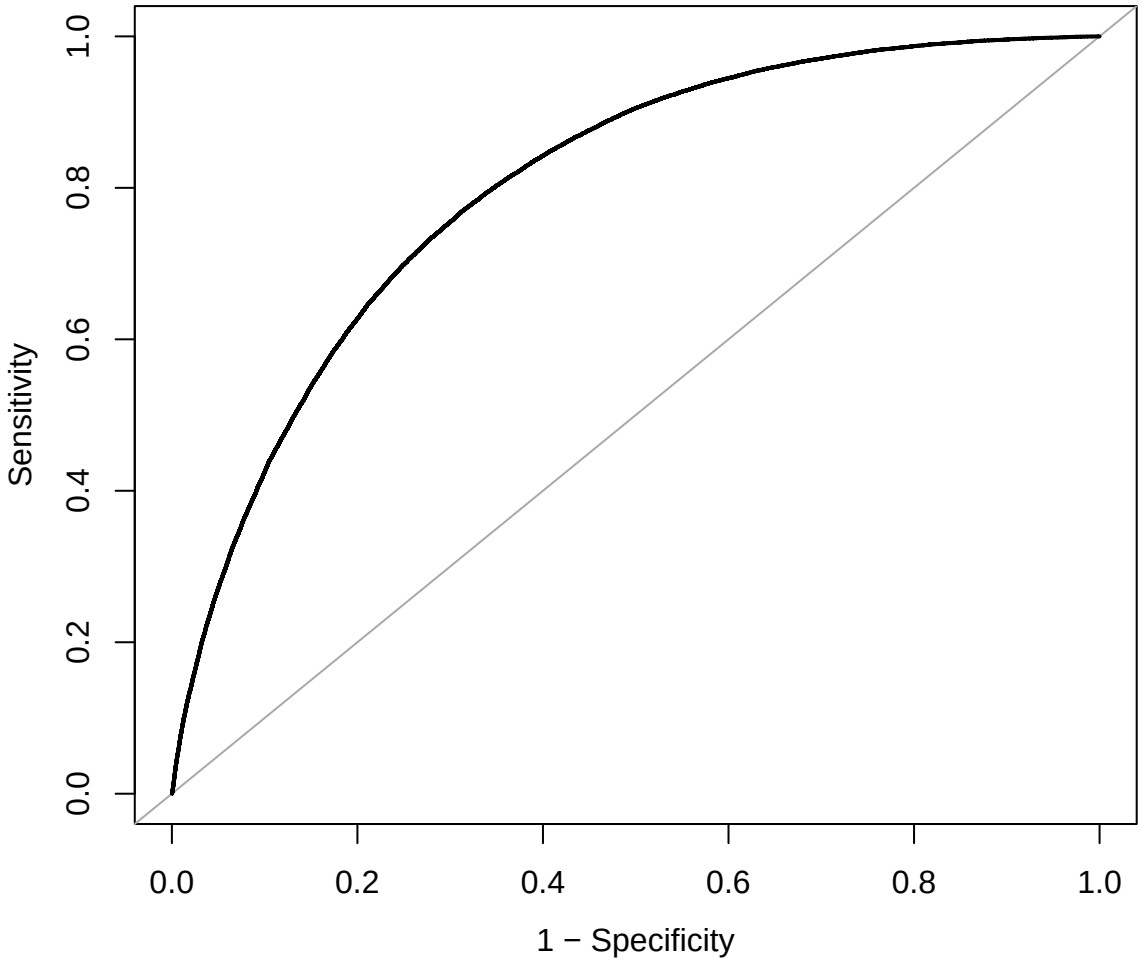
Nhận xét:

- Giá trị AUC ở mức tốt cho thấy mô hình có khả năng phân loại dự đoán đúng khoảng 80.32% giữa các nhóm mắc bệnh và không mắc bệnh.
- Với AUC = 0.8032, và khoảng tin cậy khá hẹp (0.0.8008; 0.0.8055), ta có thể tự tin rằng mô hình ổn định và cho kết quả tốt trong việc phân loại.

- Đây cũng là một minh chứng rằng mô hình không bị quá nhạy cảm với dữ liệu mẫu, và hiệu quả của nó sẽ giữ vững trong các lần thử nghiệm khác.

Đường cong ROC thu được là:

Hình 12: Đường cong ROC



Ngưỡng tối ưu dựa trên phương pháp *Youden index* hoặc phương pháp *closest top left* thể hiện ở bảng (25).

Bảng 25: Ngưỡng phân loại tối ưu trên hai phương pháp Youden index và closest top left

	Closest top left	Youden index
threshold	0.1818	0.1667
specificity	0.7162	0.6885
sensitivity	0.7400	0.7700
pred_prob	0.7026	0.7203

Nhận xét:

- **Youden index**:
  - Specificity = 0.6885, thấp hơn so với Closest top left, nhưng
  - Sensitivity = 0.7700, cao hơn, giúp phát hiện mẫu dương tính tốt hơn.
- **Closest top left**:
  - Specificity = 0.7162, cao hơn so với Youden index, làm giảm tỷ lệ âm tính giả.
  - Sensitivity = 0.7400, thấp hơn, cho thấy khả năng phát hiện mẫu dương tính bị giảm.
- Qua đó, ta thấy ngưỡng tối ưu từ phương pháp Closest top left có vẻ cân bằng hơn cho việc dự đoán phân loại.
- AUC giúp đánh giá mô hình tổng quát, trong khi các ngưỡng phân loại tối ưu cho phép lựa chọn những ngưỡng phù hợp để tối ưu hóa độ nhạy và độ đặc hiệu theo mục tiêu bài toán nên sẽ có sự sai giữa các cách đánh giá mô hình này.

6.5 Dữ liệu cân bằng

Số lượng quán sát trong mỗi nhóm bệnh sau khi cân bằng dữ liệu như sau:

Bảng 26: Số lượng quan sát trong mỗi nhóm bệnh sau khi cân bằng dữ liệu

Nhóm	**0**	**1**	**2**
Số lượng	190055	95027	95027

Việc bố trí dữ liệu để nhóm 0 $\approx$ nhóm 1 + nhóm 2, vì ta đang sử dụng hồi quy cho hai nhóm.

6.5.1 Ước lượng mô hình

Từ việc cân bằng dữ liệu, các biến phân loại không còn chia nhóm theo quy tắc hữu hạn theo số đếm nên ta sẽ để dạng biến liên tục để thực hiện hồi quy, hơn nữa việc phân loại giúp nhìn rõ hơn các trị nhưng ta đã biết được xu hướng thuận nghịch của từng biến tác động với biến phản hồi theo hình thức tỉ lệ thuận hay nghịch ở phần hồi quy logistic, ngoài trừ biến `age` có biến động tăng ở đoạn từ 10 lên 11 và giảm nhẹ từ 11 xuống 12, 13, nếu không factor thì ta không rõ được điểm này tuy nhiên điều này không quá ảnh hưởng vì mặc dù giảm nhưng tỉ lệ cược có được vẫn rất cao và ảnh hưởng mạnh như ở điểm cao nhất là `age` = 11.

Ta có kết quả ước lượng trong bảng (27).

Bảng 27: Hệ số hồi quy của mô hình hồi quy logistic 3

Variable	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-6.4153	0.0478	-134.341	< 2e-16
high_bp	0.4875	0.0085	57.366	< 2e-16
high_chol	0.5812	0.0081	71.753	< 2e-16
chol_check	1.4523	0.0357	40.703	< 2e-16
bmi	0.0704	0.0007	99.527	< 2e-16
stroke	-0.1153	0.0186	-6.194	5.86e-10
heart_diseaseor_attack	0.0561	0.0126	4.443	8.86e-06
hvy_alcohol_consump	-0.7347	0.0201	-36.483	< 2e-16
gen_hlth	0.4835	0.0051	94.856	< 2e-16
ment_hlth	-0.0041	0.0006	-7.179	7.04e-13
phys_hlth	-0.0110	0.0005	-20.540	< 2e-16
diff_walk	0.0084	0.0114	0.738	0.46
age	0.1559	0.0016	94.575	< 2e-16
income	-0.0468	0.0021	-22.819	< 2e-16

Nhận xét: Ta thấy rằng ngoại trừ ước lượng cho `diff_walk` thì các ước lượng khác đều có p-value rất bé, do đó, các biến dùng để hồi quy cho diabetes_012 đều có ý nghĩa thống kê với mức ý nghĩa $\alpha = 5\%$ đối với mô hình logistic, và nó hữu ích trong việc phân loại các đối tượng mắc bệnh tiểu đường. Tuy nhiên với `diff_walk` thì p-value thu được khá cao, điều này ngược lại so với lúc hồi quy logistic với dữ liệu gốc, có thể giải thích do việc cân bằng mẫu, và bản thân `diff_walk` cũng có tương quan khá cao với hai biến là `gen_hlth` và `phys_hlth` nên ta cân nhắc loại bỏ biến này và ước lượng lại lần nữa. Kết quả thu được trong bảng (28).

Bảng 28: Hệ số hồi quy của mô hình hồi quy logistic 4

Variable	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-6.4185	0.0476	-134.972	< 2e-16
high_bp	0.4876	0.0085	57.383	< 2e-16
high_chol	0.5812	0.0081	71.752	< 2e-16
chol_check	1.4524	0.0357	40.707	< 2e-16
bmi	0.0704	0.0007	100.492	< 2e-16
stroke	-0.1146	0.0186	-6.164	7.08e-10
heart_diseaseor_attack	0.0565	0.0126	4.483	7.36e-06
hvy_alcohol_consump	-0.7347	0.0201	-36.485	< 2e-16
gen_hlth	0.4842	0.0050	96.464	< 2e-16
ment_hlth	-0.0040	0.0006	-7.148	8.82e-13
phys_hlth	-0.0109	0.0005	-21.164	< 2e-16
age	0.1561	0.0016	95.756	< 2e-16
income	-0.0471	0.0020	-23.176	< 2e-16

Về ý nghĩa của mỗi ước lượng, ta sẽ tính tỷ lệ cược - odds cho mỗi biến trong bảng (29).

Bảng 29: Tỷ lệ cược - Odds của từng biến mô hình hồi quy logistic 4

Biến	Tỷ lệ cược
(Intercept)	0.0016
high_bp	1.6284
high_chol	1.7882
chol_check	4.2733
bmi	1.0730
stroke	0.8917
heart_diseaseor_attack	1.0581
hvy_alcohol_consump	0.4796
gen_hlth	1.6228
ment_hlth	0.9960
phys_hlth	0.9892
age	1.1689
income	0.9540

Nhận xét:

- `stroke` tỉ lệ cược giảm từ 1.1335 ở mô hình dữ liệu gốc xuống 0.8917, tức là từ những người đã bị đột quỵ (có tỉ lệ cao) có nguy cơ mắc bệnh tiểu đường cao hơn 13.35% so với những người chưa bị đột quỵ thì lúc này lại thấp hơn 10.83%, việc này cho thấy đây giống như một biến nhiễu, không đánh giá tốt được mô hình.
- `chol_check` tăng từ 3.3286 lên 4.2733, chứng tỏ rằng đây là biến vô cùng quan trọng trong việc phân loại.
- Các biến `high_bp`, `heart_diseaseor_attack` có sự giảm nhẹ, tuy nhiên, tỷ số cược của chúng vẫn trên 1, do đó, việc này cho thấy chúng là những yếu tố rất quan trọng trong việc dự đoán phân loại.
- Các biến `hvy_alcohol_consump`, `ment_hlth`, `phys_hlth` thì giảm nhẹ, việc này khẳng định vai trò làm giảm nguy cơ mắc bệnh tiểu đường của chúng hơn so mới mô hình đầu.
- Các biến `gen_hlth`, `age`, `income` tuy không được factor nhưng với tỷ số cược có được cho thấy chúng vẫn giữa tốt vai trò lăm tăng hoặc nguy cơ mắc bệnh tiểu đường giống như mô hình dữ liệu gốc.

Và vẫn như vậy, các yếu tố sức khỏe như huyết áp, cholesterol, bệnh về tim mạch, và tuổi tác có ảnh hưởng lớn hơn nhiều đến nguy cơ mắc bệnh tiểu đường so với các yếu tố lối sống và kinh tế, yếu tố đột quỵ có lẽ là không phù hợp để chẩn đoán phân loại.

6.5.2 Tiên đoán

Dự đoán mô hình với ngưỡng 0.5 **trên tập test**: kết quả thu được cho dự báo mô hình của dữ liệu gốc trên tập test là 70.84% tỉ lệ phân loại đúng hai nhóm của dữ liệu. Đây là một tỉ lệ dự đoán khá ổn, ta lấy ngẫu nhiên 10 giá trị trong tập test và so sánh với dự đoán của mô hình:

Bảng 30: Bảng nhóm tiên đoán và dữ liệu thực của mô hình hồi quy logistic 4

Nhóm tiên đoán	**Dữ liệu thực**
0	0
1	1
0	0
0	1
0	0
1	1
0	0
1	1
0	1
0	1

Nhận xét: kết quả dự đoán có xu hướng dự đoán vẫn nghiêng về nhóm 0 nhiều hơn, tuy nhiên, ở mô hình này, trạng thái tiên đoán vẫn ổn định hơn.

6.5.3 Đánh giá mô hình

Để đánh giá độ chính xác của mô hình logistic, chúng ta sẽ sử dụng phân tích đường cong ROC và diện tích dưới đường cong (AUC).

Chỉ số AUC thu được là:

$$\text{AUC} = 0.7755.$$

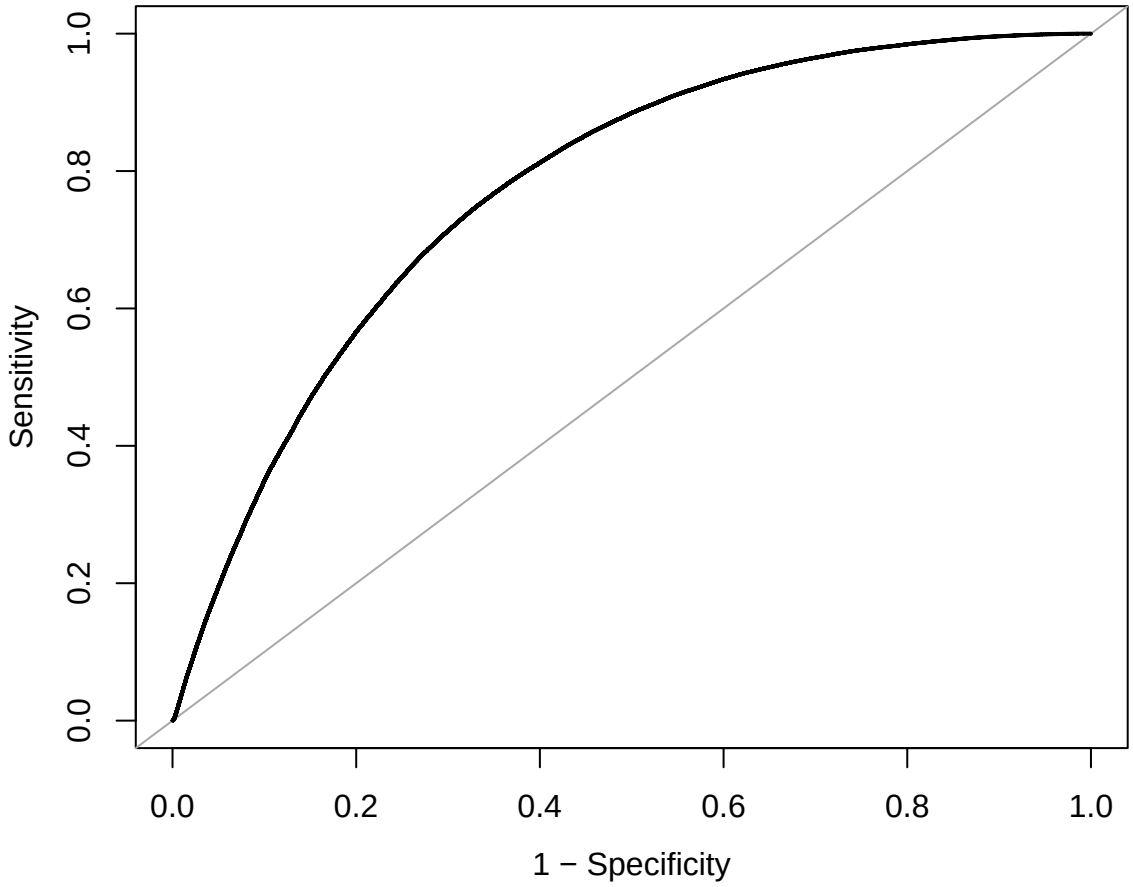
Khoảng tin cậy 95% cho AUC là: (0.7739; 0.7770).

Nhận xét:

- Giá trị AUC ở mức ổn cho thấy mô hình có khả năng phân loại dự đoán đúng khoảng 77.55% giữa các nhóm mắc bệnh và không mắc bệnh.
- Với khoảng tin cậy khá hẹp (0.7739; 0.7770), ta có thể tự tin rằng mô hình ổn định và cho kết quả tốt trong việc phân loại.

Đường cong ROC thu được là:

Hình 13: Đường cong ROC



Nhận xét: So với ROC thu được ở phần dữ liệu gốc, ROC trên có vẻ gần hơn với đường trung bình.

Ngưỡng tối ưu dựa trên phương pháp *Youden index* hoặc phương pháp *closest top left* thể hiện ở bảng (31).

Bảng 31: Ngưỡng phân loại tối ưu trên hai phương pháp Youden index và closest top left

	Youden index	Closest top left
threshold	0.4789	0.5079
specificity	0.6548	0.6889
sensitivity	0.7646	0.7274
pred_prob	0.7098	0.7081

Nhận xét: Hầu như các chỉ số đều giảm so với trước, tuy nhiên xác suất dự đoán thì lại gần nhau hơn cho các phương pháp, chứng tỏ kết quả trả về thực sự có sự ổn định hơn.

6.6 Kết luận 2

Các biến như smoke, `phys_activity`, `diff_walk` có sự ảnh hưởng quá ít đến việc phân loại hai nhóm, dù trong một số trường hợp chúng vẫn có ý nghĩa thống kê trong mô hình, tuy nhiên, nếu loại bỏ, ta vẫn thu được một mô hình khác với xác suất dự đoán đúng vẫn khá cao dù trong trường hợp mẫu như thế nào.

Biến `stroke` đang chưa thể hiện được vai trò của mình, điển hình là nó thay đổi vai trò thuận nghịch với hai trường hợp dữ liệu gốc và dữ liệu cân bằng.

Ngoài ra, các biến hồi quy khác, dù đóng góp mạnh hay yếu nhưng chúng đều giữ được vai trò làm tăng hoặc làm giảm nguy cơ mắc bệnh tiểu đường, hơn nữa, nếu nghiên cứu kĩ hơn, ta sẽ thấy các yếu tố này so với những biểu hiện mắc bệnh, nguyên nhân được công bố về bệnh tiểu đường thì đều hợp lý và là những yếu tố hợp lí nhất.

Qua đó, ta sẽ nhận xét hiệu suất của hai mô hình dữ liệu gốc và dữ liệu cân bằng sau khi đã loại bỏ 4 biến smoke, `phys_activity`, `diff_walk` và `stroke` nêu trên qua bảng (32) và (33).

Bảng 32: Các chỉ số đánh giá của mô hình dữ liệu gốc

		AUC = 0.8029
	Youden index	Closest top left
threshold	0.1698	0.1803
specificity	0.6935	0.7120
sensitivity	0.7639	0.7438
pred_prob	0.7057	0.7175

Bảng 33: Các chỉ số đánh giá của mô hình dữ liệu cân bằng

		AUC = 0.7754
	Youden index	Closest top left
threshold	0.4832	0.5069
specificity	0.6602	0.6879
sensitivity	0.7592	0.7289
pred_prob	0.7098	0.7084

Như vậy, sau khi loại bỏ 4 biến so với 15 biến đã chọn ban đầu thì việc phân loại, dù cho kết quả có bé hơn ban đầu, nhưng số liệu vẫn thể hiện ở mức tốt, hơn nữa còn đảm bảo được sự ổn định.

7 Mô hình multinominal logistic

7.1 Dữ liệu gốc

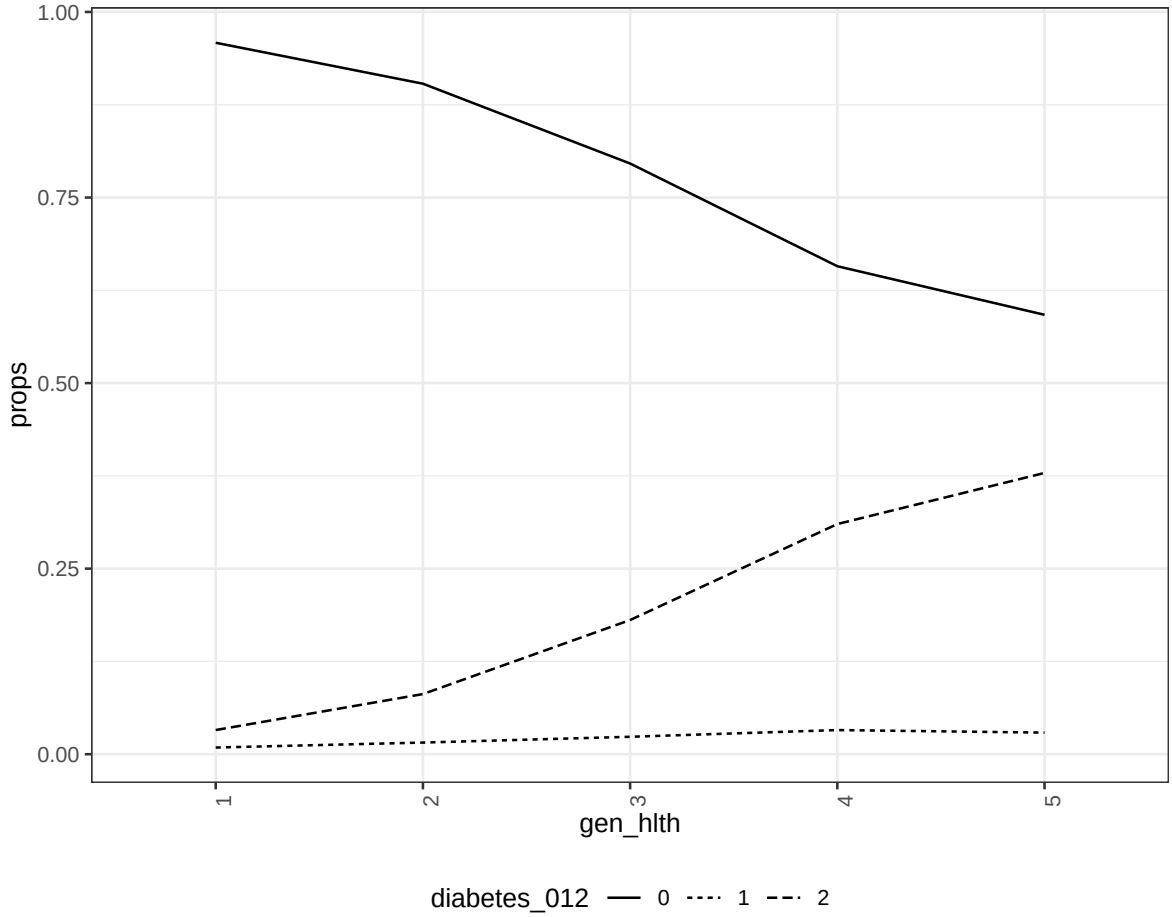
Mô hình hồi quy multinomial logistic với k nhóm:

$$\log\left(\frac{\pi(\boldsymbol{x}_j)}{\pi(\boldsymbol{x}_1)}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p, \quad j = 2, \cdots, k.$$

Nhóm $j = 1$ được chọn làm lớp tham chiếu, tức là mô hình sẽ ước tính các tham số cho các lớp $j = 2, 3, \ldots, k$ so với lớp tham chiếu $j = 1$. Trong phần này ta chọn nhóm không bị tiểu đường (0) để làm nhóm tham chiếu.

Sử dụng các biến `gen_hlth`, `age` và `income`, ta tính tỷ lệ các nhóm bệnh trong các khoảng tương ứng để xem sự ảnh hưởng của chúng.

Hình 14: Biểu đồ tỷ lệ các nhóm bệnh theo `gen_hlth`

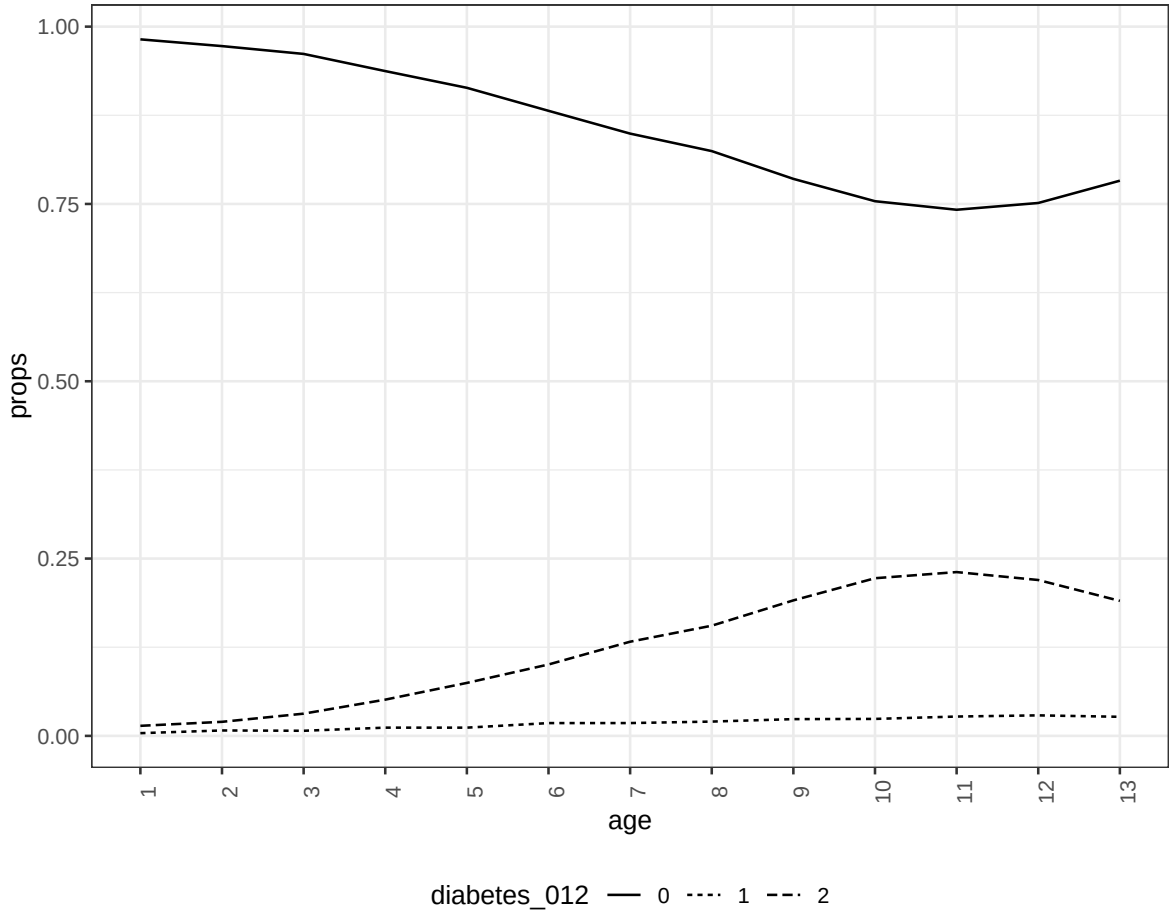


Nhận xét

- Tỷ lệ mắc bệnh tiểu đường tăng dần khi tình trạng sức khỏe giảm. Nhóm `gen_hlth` = 1 có tỷ lệ mắc tiểu đường thấp, trong khi nhóm `gen_hlth` = 5 có tỷ lệ mắc bệnh cao nhất.
- Tỷ lệ tiền tiểu đường không được phản ánh quá rõ ràng bởi vì số lượng quá ít, tuy nhiên nếu nhìn kĩ ta sẽ thấy sự tăng lên nhẹ từ mức sức khỏe 1 tới mức sức khỏe 4 và giảm nhẹ ở mức 5.

- Dù số lượng nhóm bệnh 1 và 2 rất thấp, tỉ lệ tính ra được cũng rất nhỏ nên chưa nhìn quá rõ ràng, nhưng với từng ấy quan trắc vẫn thấy được xu hướng tăng mức độ bệnh theo sức khỏe tổng quát.

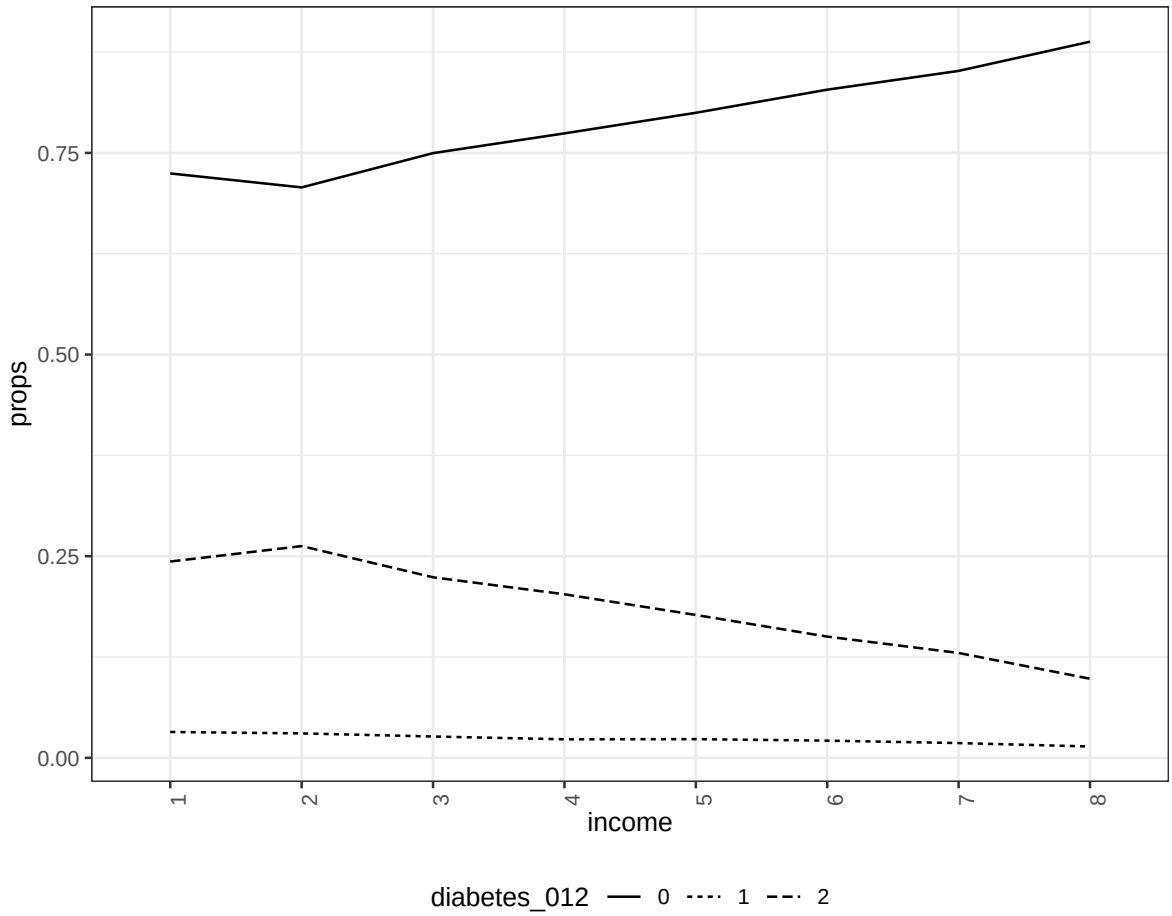
Hình 15: Biểu đồ tỷ lệ các nhóm bệnh theo age



Nhận xét

- Tỷ lệ mắc bệnh tiểu đường tăng dần khi tuổi tác tăng. Nhóm age = 1 có tỷ lệ mắc tiểu đường thấp, trong khi nhóm age = 11 có tỷ lệ mắc bệnh cao nhất và giảm nhẹ về phía nhóm age = 13, điều này cũng được nhắc đến ở các phần trước.
- Tỷ lệ tiền tiểu đường không được phản ánh quá rõ ràng bởi vì số lượng quá ít, tuy nhiên nếu nhìn kĩ ta sẽ thấy sự tăng lên nhẹ từ nhóm 1 tới nhóm tuổi 13.
- Dù số lượng nhóm bệnh 1 và 2 rất thấp, tỉ lệ tính ra được cũng rất nhỏ nên chưa nhìn quá rõ ràng, nhưng với từng ấy quan trắc vẫn thấy được xu hướng tăng mức độ bệnh theo sức khỏe tổng quát.

Hình 16: Biểu đồ tỷ lệ các nhóm bệnh theo `income`



Nhận xét

- Tỷ lệ mắc bệnh tiểu đường giảm dần khi mức lương tăng lên, có đoạn gấp khúc ở `income` $= 2$, điều này cũng đã được thấy ở phần mô hình hồi quy logistic, ở điểm `income` $= 2$, mô hình đều không có ý nghĩa thống kê có lẽ là vì vậy.
- Tỷ lệ tiền tiểu đường vẫn không được rõ ràng nhưng vẫn có xu hướng giảm dần khi `income` tăng.
- Dù số lượng nhóm bệnh 1 và 2 rất thấp, tỉ lệ tính ra được cũng rất nhỏ nên chưa nhìn quá rõ ràng, nhưng với từng ấy quan trắc vẫn thấy được xu hướng tăng mức độ bệnh theo sức khỏe tổng quát.

7.1.1 Ước lượng mô hình

Ta có kết quả ước lượng trong bảng sau:

Bảng 34: Hệ số hồi quy của mô hình multinominal logistic

Nhóm 1		**Nhóm 2**	
Variable	Coefficient (β)	Variable	Coefficient (β)
(Intercept)	-8.0890	(Intercept)	-8.0324
high_bp	0.3415	high_bp	0.7280
high_chol	0.5358	high_chol	0.5422
chol_check	0.7898	chol_check	1.2812
bmi	0.0470	bmi	0.0579
smoker	-0.0476	smoker	-0.0236
stroke	-0.1191	stroke	0.1425
heart_diseaseor_attack	0.0280	heart_diseaseor_attack	0.2860
phys_activity	0.0086	phys_activity	-0.0268
hvy_alcohol_consump	-0.1819	hvy_alcohol_consump	-0.8013
gen_hlth2	0.4042	gen_hlth2	0.6694
gen_hlth3	0.7302	gen_hlth3	1.3196
gen_hlth4	1.0317	gen_hlth4	1.7848
gen_hlth5	0.9386	gen_hlth5	1.9953
ment_hlth	0.0079	ment_hlth	-0.0045
phys_hlth	-0.0032	phys_hlth	-0.0046
diff_walk	0.0092	diff_walk	0.1417
age2	0.8217	age2	0.1014
age3	0.5691	age3	0.4914
age4	1.1070	age4	0.8890
age5	1.0352	age5	1.1343
age6	1.4742	age6	1.3477
age7	1.4211	age7	1.5677
age8	1.5391	age8	1.6588
age9	1.7063	age9	1.8819
age10	1.7486	age10	2.0389
age11	1.9104	age11	2.0923
age12	1.9298	age12	1.9798
age13	1.8692	age13	1.7903
income2	-0.1234	income2	-0.0458
income3	-0.2020	income3	-0.0619
income4	-0.3392	income4	-0.0843
income5	-0.2871	income5	-0.1729
income6	-0.3727	income6	-0.2351
income7	-0.4155	income7	-0.2422
income8	-0.5765	income8	-0.3201

Để hiểu rõ hơn về ý nghĩa của các hệ số, ta tính exp() của các ước lượng hệ số, cũng chính là tỷ số rủi ro tương đối - relative risk ratio (RRR) của từng biến:

Bảng 36: Tỷ số rủi ro tương đối - relative risk ratio (RRR) của 15 biến

Nhóm 1		**Nhóm 2**	
Variable	**RRR**	**Variable**	**RRR**
Intercept	0.0003	Intercept	0.0003
high_bp	1.4070	high_bp	2.0708
high_chol	1.7089	high_chol	1.7198
chol_check	2.2030	chol_check	3.6008
bmi	1.0481	bmi	1.0596
smoker	0.9535	smoker	0.9767
stroke	0.8877	stroke	1.1532
heart_diseaseor_attack	1.0284	heart_diseaseor_attack	1.3311
phys_activity	1.0086	phys_activity	0.9736
hvy_alcohol_consump	0.8337	hvy_alcohol_consump	0.4487
gen_hlth2	1.4981	gen_hlth2	1.9531
gen_hlth3	2.0755	gen_hlth3	3.7419
gen_hlth4	2.8057	gen_hlth4	5.9581
gen_hlth5	2.5563	gen_hlth5	7.3548
ment_hlth	1.0079	ment_hlth	0.9955
phys_hlth	0.9968	phys_hlth	0.9954
diff_walk	1.0093	diff_walk	1.1523
age2	2.2743	age2	1.1067
age3	1.7666	age3	1.6346
age4	3.0253	age4	2.4328
age5	2.8157	age5	3.1089
age6	4.3673	age6	3.8484
age7	4.1417	age7	4.7955
age8	4.6604	age8	5.2529
age9	5.5088	age9	6.5662
age10	5.7464	age10	7.6819
age11	6.7557	age11	8.1037
age12	6.8880	age12	7.2412
age13	6.4832	age13	5.9911
income2	0.8839	income2	0.9552
income3	0.8171	income3	0.9399
income4	0.7123	income4	0.9191
income5	0.7505	income5	0.8412
income6	0.6888	income6	0.7905
income7	0.6600	income7	0.7849
income8	0.5619	income8	0.7261

Nhận xét:

- **high_bp:**
  - Nhóm tiền tiểu đường có nguy cơ cao hơn 1.407 lần (40.7%) so với nhóm không bị tiểu đường.
  - Nhóm tiểu đường có nguy cơ cao hơn 2.070 lần (107.0%) so với nhóm không bị tiểu đường.
- **high_chol:** Cả hai nhóm đều có nguy cơ tương đối cao (khoảng 1.7 lần (70%)), cho thấy cholesterol cao là một yếu tố nguy cơ đáng kể đối với cả tiền tiểu đường và tiểu đường.
- **chol_check:**
  - Nhóm tiền tiểu đường có nguy cơ cao hơn 2.203 lần (120.3%) khi kiểm tra cholesterol, cho thấy việc kiểm tra có thể liên quan đến tăng nguy cơ tiền tiểu đường.
  - Nhóm tiểu đường có nguy cơ cao hơn 3.601 lần (260.08%) khi kiểm tra cholesterol, cho thấy việc kiểm tra có mối quan hệ mạnh với tăng nguy cơ tiểu đường.
- **bmi:** Mỗi đơn vị tăng của bmi làm tăng nguy cơ vào nhóm tiền tiểu đường lên 1.048 lần (4.8%) và vào nhóm tiểu đường lên 1.060 lần (6%), cho thấy ảnh hưởng tăng dần từ thừa cân/béo phì.
- **smoker:** Tác động của việc hút thuốc có xu hướng giảm nhẹ nguy cơ vào nhóm 1 (0.953) và nhóm 2 (0.976). Điều này khá vô lý.
- **stroke:** Tác động của việc từng bị đột quỵ có xu hướng giảm nhẹ nguy cơ vào nhóm 1 (0.8877) nhưng tăng nguy cơ vào nhóm 2 (1.153). Mặc dù tác động tăng giảm không cao nhưng nó cho thấy xu hướng bị đột quỵ sẽ tăng nguy cơ tiểu đường, bỏ qua tiền tiểu đường.
- **phys_activity:** Ngược lại với stroke, hoạt động thể chất lại tăng nguy cơ tiền tiểu đường và giảm nguy cơ tiểu đường, con số ảnh hưởng cũng rất rất nhỏ, tăng giảm nguy cơ đều dưới 2.5%.
- **hvy_alcohol_consump:** Là yếu tố giảm nguy cơ rõ rệt ở cả hai nhóm, với 0.8337 ở nhóm 1 và 0.4487 ở nhóm 2, đặc biệt liên quan mạnh hơn đến bệnh tiểu đường. Mặc dù các nghiên cứu cho thấy rằng uống một lượng rượu vừa phải có thể làm giảm nguy cơ mắc bệnh tiểu đường, trong trường hợp này cũng chỉ đánh giá theo thang từ 0 đến 1, nên chưa thể hiện được mức độ sử dụng rượu cao, ở đây ta chỉ ghi nhận ở mức 0 đến mức 1, không sử dụng rượu đến sử dụng rượu nhưng trong mức độ cho phép.
- **gen_hlth:**
  - Tình trạng sức khỏe càng kém (gen_hlth3 đến gen_hlth5), nguy cơ thuộc nhóm 1 và nhóm 2 càng cao.

  - `gen_hlth5`: Tăng nguy cơ thuộc nhóm 1 lên 2.556 lần (155.6%) và nhóm 2 lên 7.354 lần (635.4%).
  - Mức độ ảnh hưởng mạnh hơn ở nhóm 2, cho thấy sức khỏe tổng quát kém có liên hệ mật thiết với nguy cơ bệnh tiểu đường.
- **`ment_hlth`:** Giống với hoạt động thể chất, `ment_hlth` làm tăng nguy cơ tiền tiểu đường và giảm nguy cơ tiểu đường, con số ảnh hưởng cũng rất rất nhỏ, tăng giảm nguy cơ đều dưới 1%.
- **`age`:**
  - Nguy cơ tăng dần theo tuổi ở cả hai nhóm.
  - Nhóm tuổi `age2` đến `age4` có nguy cơ mắc bệnh tiền tiểu đường cao hơn so với các nhóm tuổi cao hơn.
  - Nhóm tuổi 60+ (`age8` đến `age13`) có nguy cơ cao vượt trội (gấp $5-8$ lần so với nhóm trẻ hơn), đặc biệt ở nhóm 2.
- **`income`:** Thu nhập cao hơn liên quan đến nguy cơ mắc bệnh thấp hơn ở cả hai nhóm.

Kết luận:

- Các yếu tố như huyết áp cao, cholesterol cao, kiểm tra cholesterol, `bmi`, sức khỏe tổng quát kém, và tuổi cao đóng vai trò quan trọng trong nguy cơ mắc tiền tiểu đường và tiểu đường.
- Các yếu tố bảo vệ bao gồm thu nhập cao và tiêu thụ rượu có hạn mức vừa phải cũng được thể hiện rõ trong mối quan hệ với RRR thấp hơn.
- Tác động của từng yếu tố mạnh hơn khi chuyển từ nhóm 1 (tiền tiểu đường) sang nhóm 2 (tiểu đường), cho thấy nguy cơ tích lũy khi bệnh tiến triển.
- Sự ảnh hưởng của `smoker` khá bất thường, hơn nữa, `stroke`, `phys_activity` và `ment_hlth` cũng thế, hơn nữa, sự ảnh hưởng của 3 biến này đều không cao.
- Với các biến không thực sự gây ảnh hưởng hoặc ảnh hưởng trái ngược với thực tế, ta sẽ bỏ các biến này ra khỏi mô hình và ước lượng lại hệ số hồi quy, ta sẽ kiểm tra lại bảng tỷ số rủi ro tương đối - relative risk ratio của từng biến xem có xuất hiện bất thường khác hay không, xem bảng (38).

Bảng 38: Tỷ số rủi ro tương đối - relative risk ratio (RRR) của 11 biến

Nhóm 1		Nhóm 2	
Variable	**RRR**	**Variable**	**RRR**
Intercept	0.0003	Intercept	0.0003
high_bp	1.404	high_bp	2.078
high_chol	1.714	high_chol	1.716
chol_check	2.203	chol_check	3.612
bmi	1.048	bmi	1.060
heart_diseaseor_attack	1.012	heart_diseaseor_attack	1.346
hvy_alcohol_consump	0.833	hvy_alcohol_consump	0.444
gen_hlth2	1.499	gen_hlth2	1.951
gen_hlth3	2.079	gen_hlth3	3.739
gen_hlth4	2.831	gen_hlth4	5.942
gen_hlth5	2.619	gen_hlth5	7.289
phys_hlth	0.998	phys_hlth	0.995
diff_walk	1.011	diff_walk	1.156
age2	2.261	age2	1.106
age3	1.753	age3	1.630
age4	2.992	age4	2.430
age5	2.787	age5	3.114
age6	4.315	age6	3.860
age7	4.065	age7	4.821
age8	4.548	age8	5.293
age9	5.341	age9	6.648
age10	5.518	age10	7.809
age11	6.439	age11	8.278
age12	6.542	age12	7.433
age13	6.146	age13	6.192
income2	0.879	income2	0.957
income3	0.809	income3	0.944
income4	0.705	income4	0.922
income5	0.740	income5	0.846
income6	0.678	income6	0.795
income7	0.649	income7	0.790
income8	0.552	income8	0.732

Nhận xét: Bảng (38) cho thấy tất cả các biến đều thể hiện đúng vai trò như trong nhận xét của mô hình ban đầu. Ta sẽ tiếp tục với mô hình 11 biến này.

7.1.2 Tiên đoán

Dự đoán mô hình trên tập test: Kết quả thu được cho dự báo mô hình trên tập test là 82.92% tỉ lệ phân loại đúng ba nhóm của dữ liệu. Đây là một tỉ lệ dự đoán khá tốt tuy nhiên vẫn còn nhiều bất cập do số lượng nhóm 0 quá lớn, mô hình có xu hướng dự đoán sang nhóm 0 nhiều hơn, hoặc giả như toàn bằng 0 thì tỉ số nhóm 0 trong dữ liệu cũng đã trên 80%, hơn nữa, nhóm 1 và nhóm 2 quá ít, do đó, phân loại gần như chỉ đúng cho nhóm 0, vì vậy, độ tin cậy lúc này vẫn chưa cao, ta có thể xem 10 giá trị dự đoán ngẫu so với dữ liệu trên tập test ở bảng (40).

Bảng 40: Bảng tiên đoán xác suất cho các nhóm bệnh tiểu đường

Quan sát	Nhóm 0	Nhóm 1	Nhóm 2	Nhóm tiên đoán	Dữ liệu thực
1	0.790	0.025	0.185	0	0
2	0.883	0.014	0.104	0	0
3	0.815	0.025	0.161	0	0
4	0.818	0.023	0.159	0	1
5	0.962	0.009	0.029	0	0
6	0.895	0.019	0.086	0	0
7	0.886	0.016	0.098	0	0
8	0.722	0.024	0.254	0	2
9	0.747	0.028	0.225	0	0
10	0.818	0.018	0.164	0	0

Nhận xét: Như vậy, với 10 giá trị ngẫu nhiên có 80% nhóm 0 và 20% cho nhóm 1 và nhóm 2 thì mô hình dự đoán 100% vào nhóm 0, vẫn suy ra được đúng 80% kết quả. Điều này là một vấn đề khá lớn, do đó, ta sẽ áp dụng phương pháp *Cân bằng dữ liệu* để giải quyết vấn đề chênh lệch số lượng các nhóm và ước lượng lại mô hình.

7.2 Dữ liệu cân bằng

Số lượng quán sát trong mỗi nhóm bệnh sau khi cân bằng dữ liệu như sau:

Bảng 41: Số lượng quan sát trong mỗi nhóm bệnh sau khi cân bằng dữ liệu

Nhóm	**0**	**1**	**2**
Số lượng	190055	190055	190055

7.2.1 Ước lượng mô hình

Từ việc cân bằng dữ liệu, các biến phân loại không còn chia nhóm theo quy tắc hữu hạn theo số đếm nên ta sẽ để dạng biến liên tục để thực hiện hồi quy, ta sẽ xem bảng tỷ số rủi ro tương đối - relative risk ratio của trường hợp 15 biến, xem với dữ liệu cân bằng thì các tác động 'lạ' của các biến smoke, `stroke` và `ment_hlth` có hiệu quả khác hay không. Ta có kết quả tỷ số rủi ro tương đối - relative risk ratio của trường hợp 15 biến trong bảng (42).

Bảng 42: Tỷ số rủi ro tương đối - relative risk ratio (RRR) của 15 biến

Nhóm 1		**Nhóm 2**	
Variable	**RRR**	**Variable**	**RRR**
Intercept	0.003	Intercept	0.000
high_bp	1.366	high_bp	1.918
high_chol	1.780	high_chol	1.798
chol_check	3.867	chol_check	6.214
bmi	1.070	bmi	1.095
smoker	0.994	smoker	1.008
stroke	0.684	stroke	1.066
heart_diseaseor_attack	0.873	heart_diseaseor_attack	1.213
phys_activity	1.016	phys_activity	0.988
hvy_alcohol_consump	0.571	hvy_alcohol_consump	0.346
gen_hlth	1.470	gen_hlth	1.997
ment_hlth	1.000	ment_hlth	0.989
phys_hlth	0.988	phys_hlth	0.986
diff_walk	0.947	diff_walk	1.007
age	1.175	age	1.207
income	0.943	income	0.966

Nhận xét: Ta thấy rằng kết quả thu được khá giống với nhận xét mô hình 15 ở dữ liệu gốc, các tỷ số có sự chênh lệch nhẹ nhưng không thay đổi vai trò của nó, ngoài ra biến heart_diseaseor_attack cũng đang có tỷ số rủi ro tương đối - relative risk ratio giảm cho cả hai nhóm, và sự giảm này làm đảo lộn vai trò của nó trong cả hai nhóm, nếu việc giảm nguy cơ tiền tiểu đường đi kèm với khả năng tăng về bệnh tiểu đường cao hơn, thì ta có thể cân nhắc giữ lại nhưng trong trường hợp này, cả hai chỉ số ở cả hai nhóm đều có xu hướng giảm so với mô hình 15 biến. Do đó, ta sẽ chọn mô hình 10 biến như ở phần dữ liệu gốc. Kết quả thu được như sau:

Bảng 44: Tỷ số rủi ro tương đối - relative risk ratio (RRR) của 10 biến

Nhóm 1		**Nhóm 2**	
Variable	**RRR**	**Variable**	**RRR**
Intercept	0.003	Intercept	0.000
high_bp	1.348	high_bp	1.944
high_chol	1.760	high_chol	1.809
chol_check	3.897	chol_check	6.217
bmi	1.071	bmi	1.095
hvy_alcohol_consump	0.573	hvy_alcohol_consump	0.341
gen_hlth	1.449	gen_hlth	2.013
phys_hlth	0.987	phys_hlth	0.984
diff_walk	0.928	diff_walk	1.004
age	1.168	age	1.223
income	0.945	income	0.970

Nhận xét: Như các nhận xét trước, các yếu tố sức khỏe như huyết áp, cholesterol, bệnh về tim mạch, và tuổi tác và sức khỏe tổng quát có ảnh hưởng lớn hơn nhiều đến nguy cơ mắc bệnh tiểu đường so với các yếu tố lối sống và kinh tế.

7.2.2 Tiên đoán

Dự đoán mô hình trên tập test: kết quả thu được cho dự báo mô hình của dữ liệu gốc trên tập test là 51.92% tỉ lệ phân loại đúng ba nhóm của dữ liệu. Đây là một tỉ lệ dự đoán khá tệ, ta lấy ngẫu nhiên 10 giá trị trong tập test và so sánh với dự đoán của mô hình:

Bảng 46: Kết quả Dự đoán và Dữ liệu thực

STT	Nhóm 0	Nhóm 1	Nhóm 2	Dự đoán	Thực tế
1	0.341	0.325	0.333	0	0
2	0.192	0.393	0.415	2	0
3	0.075	0.363	0.561	2	2
4	0.371	0.427	0.202	1	0
5	0.460	0.328	0.213	0	0
6	0.533	0.325	0.142	0	0
7	0.624	0.267	0.109	0	0
8	0.264	0.364	0.372	2	0
9	0.346	0.386	0.268	1	0
10	0.236	0.359	0.405	2	0

Bảng 47: Ma trận nhầm lẫn - Dự đoán và Dữ liệu thực trên tập kiểm tra

Dự đoán	Nhóm 0	Nhóm 1	Nhóm 2
0	12314	5309	3186
1	3247	5622	4168
2	3487	8014	11670

Dự đoán	Nhóm 0	Nhóm 1	Nhóm 2
0	59.18%	25.51%	15.31%
1	24.91%	43.12%	31.97%
2	15.05%	34.59%	50.36%

Dự đoán mô hình trên chính dữ liệu:

Bảng 48: Ma trận nhầm lẫn - Dự đoán và dữ liệu thực

Dự đoán	0	1	2
0	109577	48017	28088
1	29707	49761	38503
2	31723	73332	104440

Dự đoán	Nhóm 0	Nhóm 1	Nhóm 2
0	59.07%	25.82%	15.11%
1	25.24%	42.13%	32.64%
2	14.99%	34.68%	49.42%

Nhận xét: Mô hình dự đoán cho nhóm 0 hoặc 2 tốt hơn hẳn so với nhóm 1, kết quả dự đoán có xu hướng dự đoán có sự chênh lệch giữa hai nhóm kề nhau như 0-1, 1-2, nếu dự đoán rơi vào 0 hoặc 2, ta có thể phân vân, hoặc xây dựng một mô hình hồi quy khác tiên đoán cho hai nhóm, nếu dự đoán rơi vào 1, điều này khó để chắc chắn nó đang nằm ở trường hợp nào.

7.2.3 Kết luận

Đối với mô hình multinominal logistic, mô hình trên dữ liệu gốc với khả năng tiên đoán gần như là phân loại vào nhóm 0, bởi vì sự chênh lệch dữ liệu, do đó mà độ tin cậy gần như rất thấp, xác suất phân loại vào 0 như bảng (40) cho ta thấy, nhóm 0 luôn là hơn 70%. Do đó mà khi ta xây dựng mô hình hồi quy trên dữ liệu cân bằng, mặc dù xác suất dự đoán đúng rất thấp chỉ khoảng 50%, tuy nhiên thì dữ liệu vẫn tiên đoán khá ổn cho nhóm 0 và nhóm 2, hoặc nếu dữ liệu được cải thiện hơn về sự đồng đều cho các nhóm, có lẽ độ chính xác của mô hình sẽ tăng hơn nữa, vì ta cũng thấy rằng trong nhóm tiên đoán là 0, xác suất tiên đoán sai vào nhóm 2 rất thấp và ngược lại, hầu hết các khả năng dự đoán sai do dự đoán vào nhóm 1.

Về mô hình, sau khi loại bỏ các biến cho kết quả khác với thực tế và ít có phần ảnh hưởng hoặc ảnh hưởng "nhiễu", thì ta thu được mô hình multinominal logistic với 10 biến. Về khả năng dự đoán của mô hình.

- **Nhóm 0:** Khả năng dự đoán Nhóm 0 là khá tốt. Hầu hết các trường hợp thực tế thuộc Nhóm 0 đều được mô hình dự đoán đúng, cho thấy mô hình khá mạnh trong việc phân biệt nhóm này so với các nhóm khác.
- **Nhóm 1 và Nhóm 2:** Khả năng dự đoán Nhóm 1 và Nhóm 2 có vẻ kém hơn. Mặc dù mô hình đã đạt được một mức độ chính xác hợp lý đối với Nhóm 1 (42.13%) và Nhóm 2 (49.42%), các tỷ lệ này vẫn thấp hơn Nhóm 0. Điều này cho thấy mô hình có sự nhầm lẫn khi phân loại các nhóm 1 và 2 với nhau.
- **Tổng thể:** Mô hình có thể hoạt động tốt đối với Nhóm 0 nhưng cần cải thiện để tăng độ chính xác đối với Nhóm 1 và Nhóm 2. Điều này có thể bao gồm việc cải thiện chất lượng dữ liệu hoặc thử nghiệm các thuật toán phân loại khác để nâng cao hiệu quả phân loại đối với tất cả các nhóm.
- **Khả năng áp dụng trong thực tế:** Nếu mục đích của mô hình là phân loại tốt Nhóm 0, thì mô hình hiện tại có thể đáp ứng được yêu cầu. Tuy nhiên, nếu cần độ chính xác cao hơn cho Nhóm 1 và Nhóm 2, có thể cần cải tiến thêm hoặc thử nghiệm với các kỹ thuật khác.

8 Phân loại Naive Bayes

Mô hình Naive Bayes dựa trên ước lượng xác suất hậu nghiệm $\hat{p}_j$ có dạng như sau:

$$\hat{p}_{j,\text{NB}}(x) = \frac{\hat{f}_{j1}(x_1) \times \hat{f}_{j2}(x_2) \times \cdots \times \hat{f}_{jp}(x_p) \times \hat{\pi}_j}{\sum_{i=1}^{k} \hat{f}_{i1}(x_1) \times \hat{f}_{i2}(x_2) \times \cdots \times \hat{f}_{ip}(x_p) \times \hat{\pi}_i}.$$

Khi đó, dựa vào ước lượng xác suất hậu nghiệm, ta có ước lượng nhóm như sau:

- **K = 2**
$$\hat{Y}(x) = \begin{cases} 1 & \text{nếu } \hat{p}_{1,\text{NB}}(x) > p_0, \\ 0 & \text{nếu } \hat{p}_{1,\text{NB}}(x) \leq p_0, \end{cases}$$
- **K ≥ 3**
$$\hat{Y}(x) = \arg \max_{j=1,\ldots,K} \hat{p}_{j,\text{NB}}(x),$$
tức là nhóm j có xác suất hậu nghiệm lớn nhất.

Ý tưởng đối với dữ liệu này là ta sẽ chia thành 2 trường hợp để xử lý dữ liệu cũng như là quan sát phân loại Naive Bayes:

- **TH1:** Gộp 2 nhóm Pre-Diabetes và Diabetes thành 1 nhóm rồi xây dựng mô hình cho 2 nhóm với dữ liệu gốc.
- **TH2:** Xây dựng mô hình cho 3 nhóm với dữ liệu đã được cân bằng.

Trường hợp 1

Với trường hợp này, ta sẽ thay thế nhóm Pre-Diabetes thành nhóm Diabetes. Khi đó dữ liệu sẽ được số lượng quan sát mới như sau:

Bảng 49: Số lượng quan sát đối với 2 nhóm No Diabetes và Diabetes

Category	Count
No Diabetes	190,055
Diabetes	39,726

Tiếp đến, ta sẽ xây dựng mô hình dựa trên số lượng quan sát mới này. Vì dữ liệu quá lớn cũng như để đảm bảo kết quả nên ta sẽ chia tập train và tập test với tỉ lệ 9:1.

Ta sử dụng các biến đã được chọn làm biến giải thích để xây dựng mô hình phân loại Naive Bayes.

Vì lượng dữ liệu lớn nên ta sẽ quan sát 20 giá trị đầu của kết quả xác suất hậu nghiệm.

Bảng 50: Xác suất hậu nghiệm của 20 quan sát đầu tiên

Quan sát	No Diabetes	Diabetes	Quan sát	No Diabetes	Diabetes
1	0.4255	0.5745	11	0.9999	0.0001
2	0.9692	0.0308	12	0.9811	0.0189
3	0.9944	0.0056	13	0.1673	0.8327
4	0.7880	0.2120	14	0.9661	0.0339
5	0.9808	0.0192	15	0.9989	0.0011
6	0.8776	0.1224	16	0.8628	0.1372
7	0.3973	0.6027	17	0.9996	0.0004
8	0.9930	0.0070	18	0.9734	0.0266
9	0.9797	0.0203	19	0.9230	0.0770
10	0.2874	0.7126	20	0.9992	0.0008

Nhận xét: Dựa vào bảng xác suất từ phân loại Naive Bayes, ta có thể dự đoán được với 20 quan sát đầu tiên ai là người mắc bệnh tiểu đường. Ta sẽ kiểm tra độ chính xác của mô hình bằng ma trận phân loại (confusion matrix) cũng như các chỉ số là precision, recall, kappa, marco-F1.

	No Diabetes	Diabetes
Precision	0.8545	0.5024
Recall	0.9557	0.2156

	Value
Accuracy	0.8286
Kappa	0.2214
Macro-F1	0.2498

Nhận xét: Với độ chính xác là 0.8286, ta thấy rằng mô hình với bộ dữ liệu này là dự đoán tốt. Precision của nhóm No Diabetes là 0.8545, cho thấy mô hình phân loại nhóm này là tốt. Còn đối với nhóm Diabetes thì mô hình phân loại chưa tốt.

Chỉ số Recall cho ta biết độ bỏ sót dữ liệu của mô hình. Với nhóm No Diabetes thì mô hình gần như không bỏ sót dữ liệu nào, còn đối với nhóm Diabetes thì mô hình có khả năng bỏ sót nhiều dữ liệu. Điều này 1 phần là do nhóm No Diabetes chiếm phần lớn dữ liệu.

Chỉ số Kappa của mô hình thấp cho thấy dự đoán của mô hình hoàn toàn độc lập với phân loại thực tế.

Và cuối cùng là chỉ số Macro-F1, cho thấy các lớp dự đoán kém, ở đây là lớp Diabetes.

Trường hợp 2

Nhận xét: Ta thấy rằng số lượng của Pre-Diabetes chiếm phần nhỏ nhất trong tập dữ liệu, do đó ta sẽ xử lý phần mất cân bằng trong dữ liệu này bằng cách dùng phương pháp

Bảng 51: Số lượng quan sát cho từng nhóm

	No Diabetes	**Pre-Diabetes**	**Diabetes**
Số lượng	190055	4629	35097

data generation với thuật toán SMOTE cho 3 nhóm No Diabetes, Pre-Diabetes và nhóm Diabetes.

Kết quả sau khi áp dụng thuật toán SMOTE

Bảng 52: Số lượng quan sát cho từng nhóm

	No Diabetes	**Pre-Diabetes**	**Diabetes**
Số lượng	190055	190055	190055

Và ta cũng có dữ liệu của 20 quan sát đầu tiên.

Bảng 53: Dữ liệu xác suất của 10 quan sát đầu

Quan sát	**No Diabetes**	**Pre-Diabetes**	**Diabetes**
1	0.6172	0.2585	0.1243
2	0.2480	0.3896	0.3624
3	0.9248	0.0667	0.0085
4	0.0026	0.1037	0.8937
5	0.0446	0.1731	0.7823
6	0.9739	0.0231	0.0030
7	0.4828	0.2694	0.2478
8	0.5608	0.2809	0.1583
9	0.9998	0.0001	0.0001
10	0.2304	0.4306	0.3390

Với trường hợp này ta cũng sẽ kiểm tra độ chính xác của mô hình bằng ma trận phân loại (confusion matrix) cũng như các chỉ số là precision, recall, kappa, marco-F1.

	No Diabetes	Pre-Diabetes	Diabetes
Precision	0.5361	0.5637	0.5250
Recall	0.8613	0.2896	0.4577

	Value
Accuracy	0.5376
Kappa	0.3055
Macro F1	0.1565

Nhận xét: Với trường hợp 3 nhóm đã được cân bằng, ta thấy độ chính xác của mô hình đã giảm xuống đáng kể, chỉ còn 0.5376. Chỉ số Precision cho thấy rằng khả năng phân loại của mô hình này đối với 3 nhóm là không tốt.

Bảng 54: Dữ liệu xác suất của 10 quan sát kế tiếp

Quan sát	No Diabetes	Pre-Diabetes	Diabetes
11	0.6636	0.2073	0.1291
12	0.8459	0.0702	0.0839
13	0.3050	0.2096	0.4854
14	0.5127	0.2047	0.2826
15	0.0805	0.2516	0.6679
16	0.2289	0.2536	0.5175
17	0.3429	0.4207	0.2364
18	0.9761	0.0210	0.0029
19	0.5966	0.2695	0.1339
20	0.9561	0.0416	0.0023

Chỉ số Recall ở trong trường hợp này là không tốt với 2 nhóm Pre-Diabetes và nhóm Diabetes, nghĩa là khả năng bỏ sót dữ liệu của mô hình với 2 nhóm này là vô cùng lớn

Với chỉ số Kappa và Macro-F1, cũng cho ta nhận thấy mô hình này không tối ưu dự đoán đối với các lớp, mô hình cũng độc lập với phân loại thực tế.

9 Phân loại LDA và QDA

9.1 Lời dẫn

Naive Bayes là một mô hình phân loại dựa trên giả định về tính độc lập giữa các biến giải thích trong một nhóm danh mục bất kỳ. Tuy nhiên, giả định này thường bị vi phạm bởi dữ liệu thực tế. Các nhà nghiên cứu đã đề xuất hướng tiếp cận mới bằng cách ước lượng hàm mật độ xác suất đồng thời dựa trên phân phối chuẩn nhiều chiều. Dựa trên cơ sở đó, các nhà nghiên cứu đã phát triển một kỹ thuật phân loại khác. Đó là kỹ thuật phân tích phân biệt - Discriminant Analysis. Trong bài báo cáo, ta sẽ tập trung vào hai kỹ thuật phân tích phân biệt, đó là:

- Phân tích phân biệt tuyến tính - Linear Discriminant Analysis (LDA);
- Phân tích phân biệt bậc hai - Quadratic Discriminant Analysis (QDA).

Ta sẽ ứng dụng hai kỹ thuật phân tích phân biệt LDA và QDA vào bộ dữ liệu thực tế `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 liên quan đến việc đánh giá sức khỏe, cụ thể hơn là đánh giá khả năng mắc bệnh tiểu đường. Sau quá trình xử lý giá trị khuyết và trùng lặp dữ liệu, thực tế bộ dữ liệu mới chứa thông tin khảo sát của $229,781$ người dân.

Biến mục tiêu là `Diabetes_012` mô tả tình trạng bệnh tiểu đường của một người với các mức độ:

- nhóm người không bị bệnh tiểu đường - gán nhãn 0;

- nhóm người tiền tiểu đường - gán nhãn 1;
- nhóm người bị bệnh tiểu đường - gán nhãn 2.

Mục tiêu của ta là xây dựng các mô hình dựa trên những hướng tiếp cận vấn đề khác nhau, sau đó đánh giá kết quả của mỗi mô hình nói riêng và so sánh giữa các mô hình nói chung.

Ta sử dụng phương pháp tập xác thực (Cross Validation Approach), với ý tưởng chia bộ dữ liệu gốc thành hai tập dữ liệu: một tập dùng để huấn luyện mô hình (training data) với 70% dữ liệu được lấy ngẫu nhiên không lặp từ bộ dữ liệu gốc và 30% dữ liệu còn lại là tập kiểm tra (testing data) dùng để so sánh giá trị thực tế với giá trị ước lượng từ mô hình trên tập huấn luyện.

Các hướng xây dựng mô hình cho cả hai phương pháp LDA và QDA như sau:

- Với bộ dữ liệu gốc (Không cân bằng giữa các nhóm)
  - Xây dựng mô hình phân loại 3 nhóm mức độ của biến `Diabetes_012`, cụ thể ta sẽ:
    - Xây dựng mô hình với tất cả 22 biến;
    - Xây dựng mô hình với 15 biến đã được chọn trong phần (4);
  - Xây dựng mô hình phân loại 2 nhóm mức độ của biến `Diabetes_012` sau khi biến đổi với 15 biến đã chọn.
- Với bộ dữ liệu đã cân bằng nhóm
  - Xây dựng mô hình phân loại 3 nhóm mức độ của biến `Diabetes_012` với 15 biến đã chọn;
  - Xây dựng mô hình phân loại 2 nhóm mức độ của biến `Diabetes_012` sau khi biến đổi với 15 biến đã chọn.

Ta xây dựng mô hình LDA và QDA được hỗ trợ trong thư viện `MASS` trên phần mềm `R`.

9.2 LDA

Mô hình 1: Mô hình LDA được xây dựng dựa trên tất cả 22 biến từ bộ dữ liệu gốc (không cân bằng) để phân loại 3 nhóm của biến mục tiêu `Diabetes_012`.

Từ kết quả ước lượng 55, ta thấy trong tập training:

- Có khoảng 82.7% quan sát được phân loại vào nhóm 0, tức là không có mắc bệnh tiểu đường;
- Có khoảng 2.05% quan sát được phân loại vào nhóm 1, tức tiền tiểu đường;
- Có khoảng 15.26% quan sát được phân loại vào nhóm 2, tức là có mắc bệnh tiểu đường;

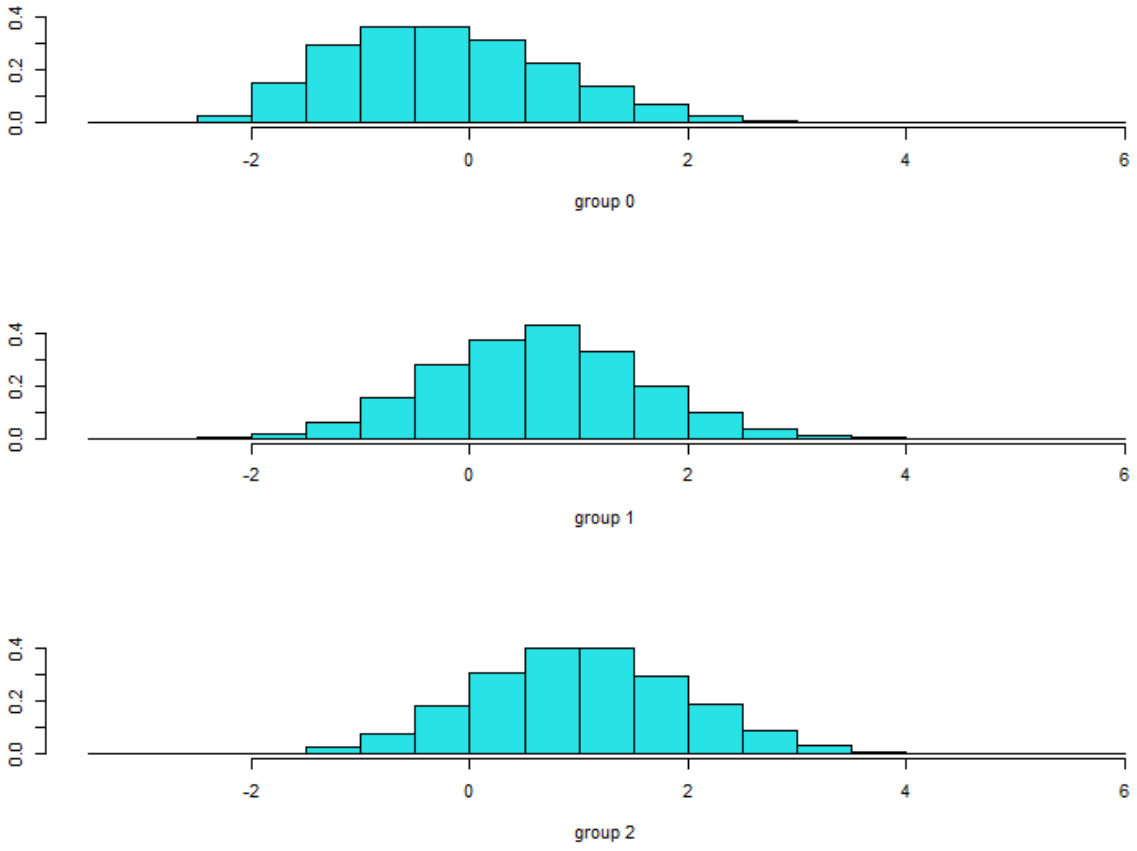
Mô hình LDA phân loại dựa trên các trục phân biệt tuyến tính, nếu có k nhóm thì số trục phân biệt sẽ là $k-1$ trục. Chúng đóng vai trò quan trọng trong việc giảm số chiều dữ liệu và phân loại.

Bảng 55: Xác suất tiên nghiệm của mô hình 1.

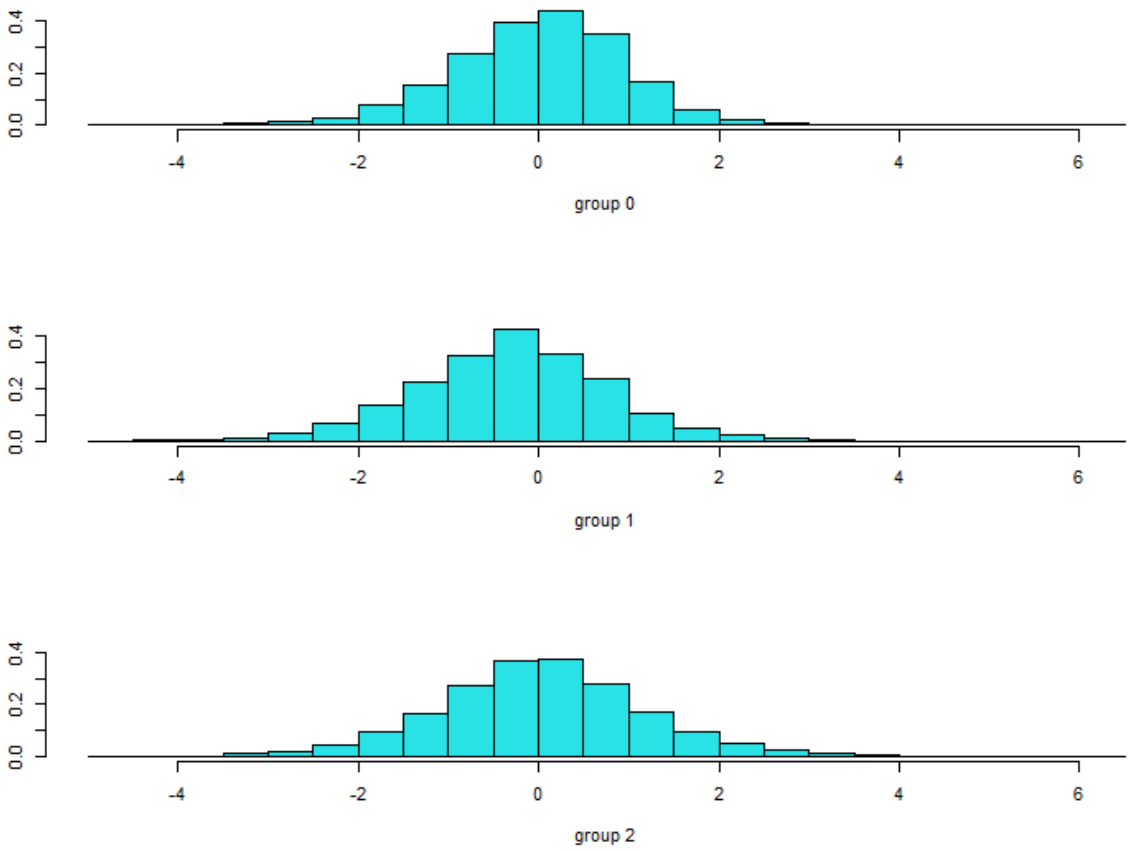
Nhóm 0	Nhóm 1	Nhóm 2
0.82696435	0.02046618	0.15256947

Do biến mục tiêu ta đang xét có ba nhóm phân loại nên trong kết quả ước lượng LDA, ta chú ý sẽ có thông tin phần trăm khả năng phân loại của hai trục phân biệt LD1 và LD2. Đối với mô hình này, ta có 99.11% dữ liệu được phân loại bằng trục LD1 và 0.89% dữ liệu được phân loại từ trục LD2. Để đánh giá được khả năng phân loại của hai trục, ta sử dụng đồ thị histogram dành riêng cho LDA (histlda) thông qua hàm `ldahist` được hỗ trợ trong cùng thư viện mô hình.

Từ hình 17, ta nhận thấy rằng có sự trùng lặp đáng kể giữa các đồ thị histogram của 3 nhóm, do đó trục LD1 của mô hình 1 không có sự phân loại tốt giữa 3 nhóm. Nhận xét tương tự cho trục LD2.

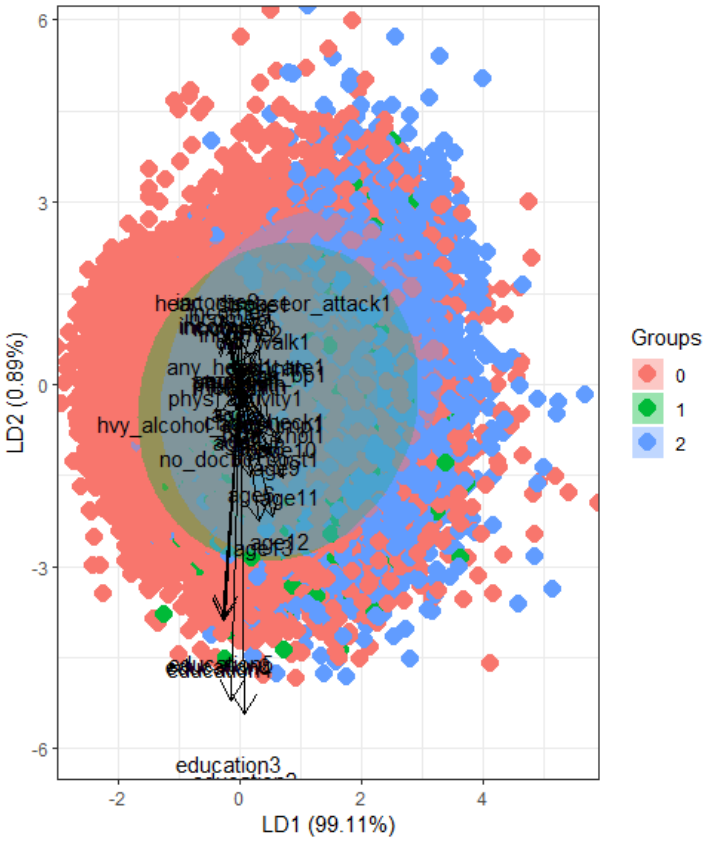


Hình 17: Đồ thị histlda tương ứng với trục LD1 của mô hình 1.



Hình 18: Đồ thị histlda tương ứng với trục LD2 của mô hình 1.

Một cách trực quan khác, ta có thể sử dụng đồ thị Biplot như hình 19. Từ hình vẽ, ta có thể thấy rằng không có sự phân biệt rõ giữa các nhóm. Bên cạnh đó ta khó có thể thấy được các biến có ảnh hưởng đến khả năng phân loại của một nhóm bất kỳ.



Hình 19: Đồ thị Biplot của mô hình 1.

Tiên đoán trên tập testing, ta thu được một số kết quả tính toán sau:

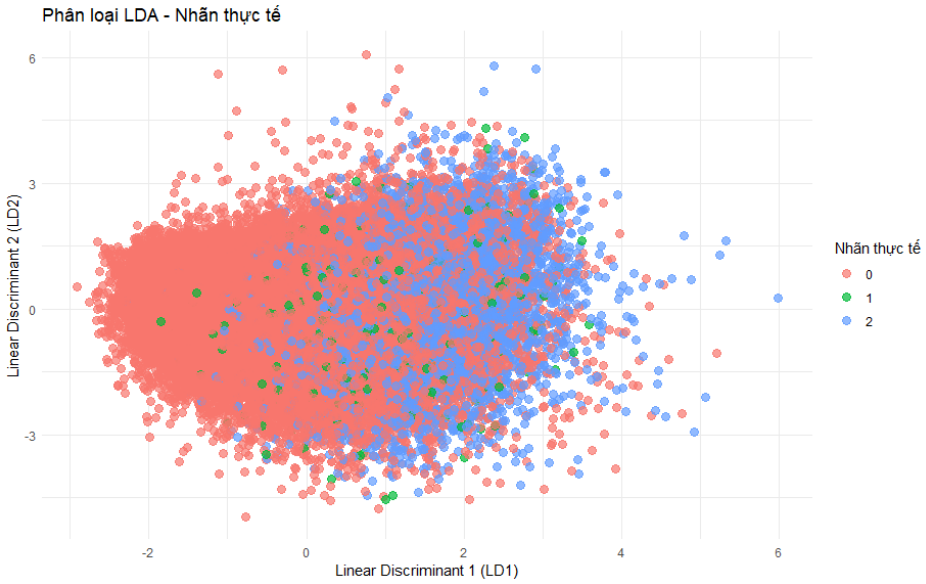
Bảng 56: Confusion Matrix giữa thực tế và dự đoán của mô hình 1.

Actual	**Predict**			**Tổng**
	Nhóm 0	**Nhóm 1**	**Nhóm 2**	
Nhóm 0	55182	1209	8351	64742
Nhóm 1	0	0	0	0
Nhóm 2	1936	130	2220	4286
Tổng	57118	1339	10571	69028

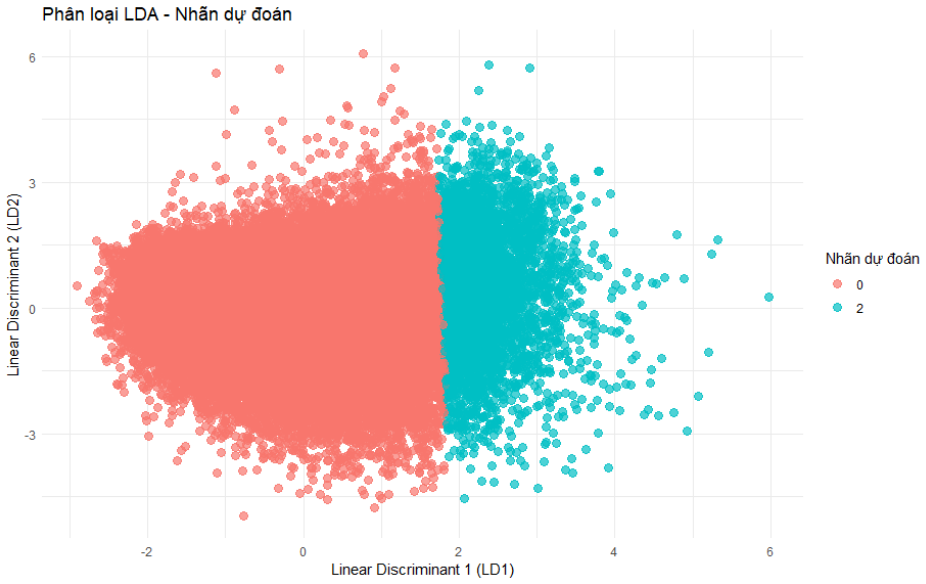
Do mô hình 1 không có gán bất kỳ nhãn 1 nào cho quan sát trong tập testing nên do đó, ta không có tính được chỉ số Precision (Bảng 57). Ta cũng tính được chỉ số dự đoán chính xác các lớp của mô hình 1 là khoảng 80.13%. Đây là một kết quả khá tốt. Việc mô hình không thể gán nhãn nhóm 1 có thể do việc ta chia phần trăm dữ liệu ban đầu làm cho dữ liệu của nhóm 1 không đủ nhiều cho mô hình huấn luyện nên việc tiên đoán có thể thiếu chính xác.

Bảng 57: Một số chỉ số đánh giá phân loại của mô hình 1.

	Nhóm 0	**Nhóm 1**	**Nhóm 2**
Precision	0.8523	–	0.518
Recall	0.9661	0	0.2100



Hình 20: Trực quan quan sát thực tế trên tập testing.



Hình 21: Trực quan quan sát dự đoán trên tập testing.

Mô hình 2: Mô hình LDA được xây dựng dựa trên 15 biến đã chọn lọc từ bộ dữ liệu gốc (không cân bằng) để phân loại 3 nhóm của biến mục tiêu `Diabetes_012`.

Từ kết quả ước lượng 58, ta thấy trong tập training:

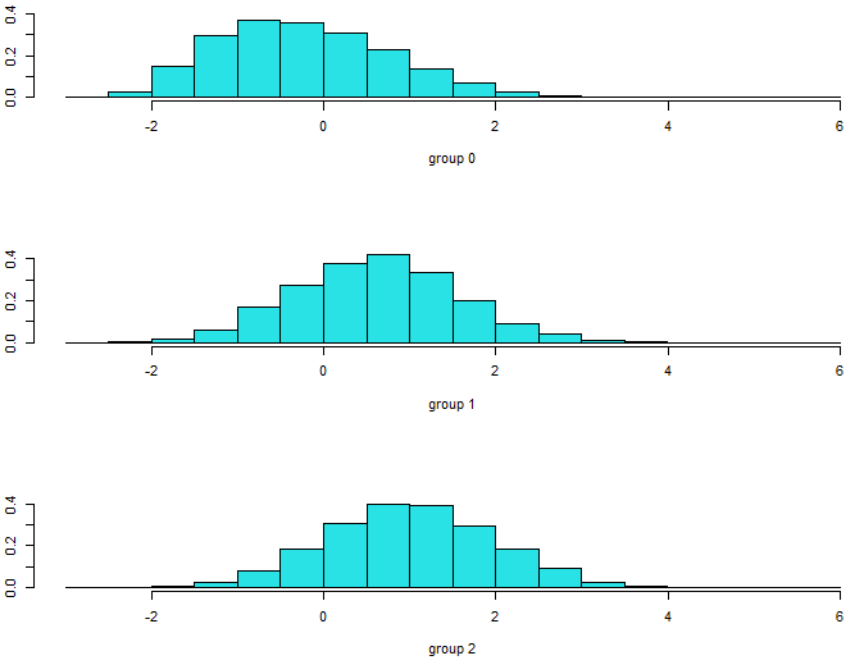
- Có khoảng 82.7% quan sát được phân loại vào nhóm 0, tức là không có mắc bệnh tiểu đường;
- Có khoảng 2.05% quan sát được phân loại vào nhóm 1, tức tiền tiểu đường;
- Có khoảng 15.26% quan sát được phân loại vào nhóm 2, tức là có mắc bệnh tiểu đường;

Bảng 58: Xác suất tiên nghiệm của mô hình 2.

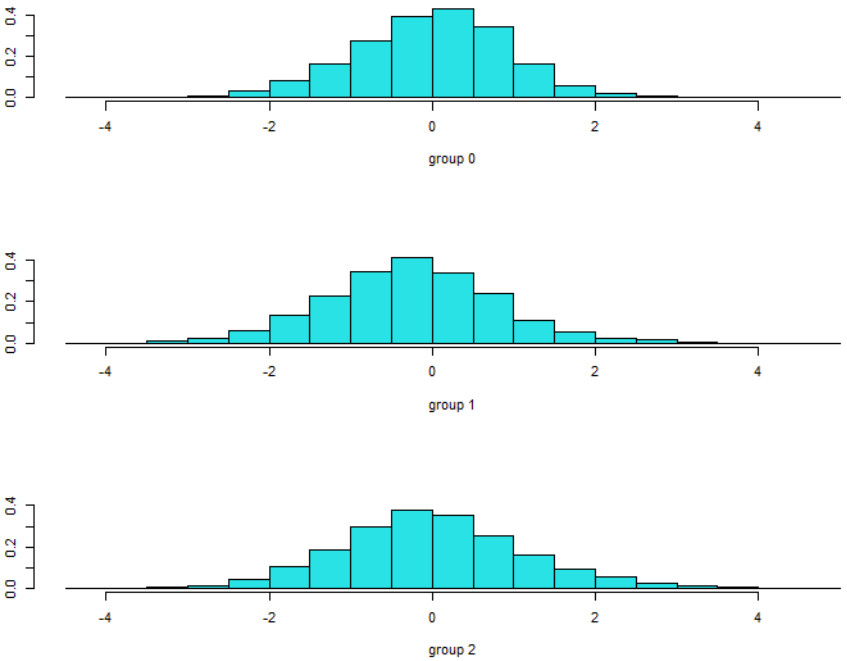
Nhóm 0	**Nhóm 1**	**Nhóm 2**
0.82696435	0.02046618	0.15256947

Do biến mục tiêu ta đang xét có ba nhóm phân loại nên trong kết quả ước lượng LDA, ta chú ý sẽ có thông tin phần trăm khả năng phân loại của hai trục phân biệt LD1 và LD2. Đối với mô hình này, ta có 99.3% dữ liệu được phân loại bằng trục LD1 và 0.7% dữ liệu được phân loại từ trục LD2. Để đánh giá được khả năng phân loại của hai trục, ta sử dụng đồ thị histogram dành riêng cho LDA (histlda) thông qua hàm `ldahist` được hỗ trợ trong cùng thư viện mô hình.

Từ hình 22, ta nhận thấy rằng có sự trùng lặp đáng kể giữa các đồ thị histogram của 3 nhóm, do đó trục LD1 của mô hình 2 không có sự phân loại tốt giữa 3 nhóm. Nhận xét tương tự cho trục LD2.

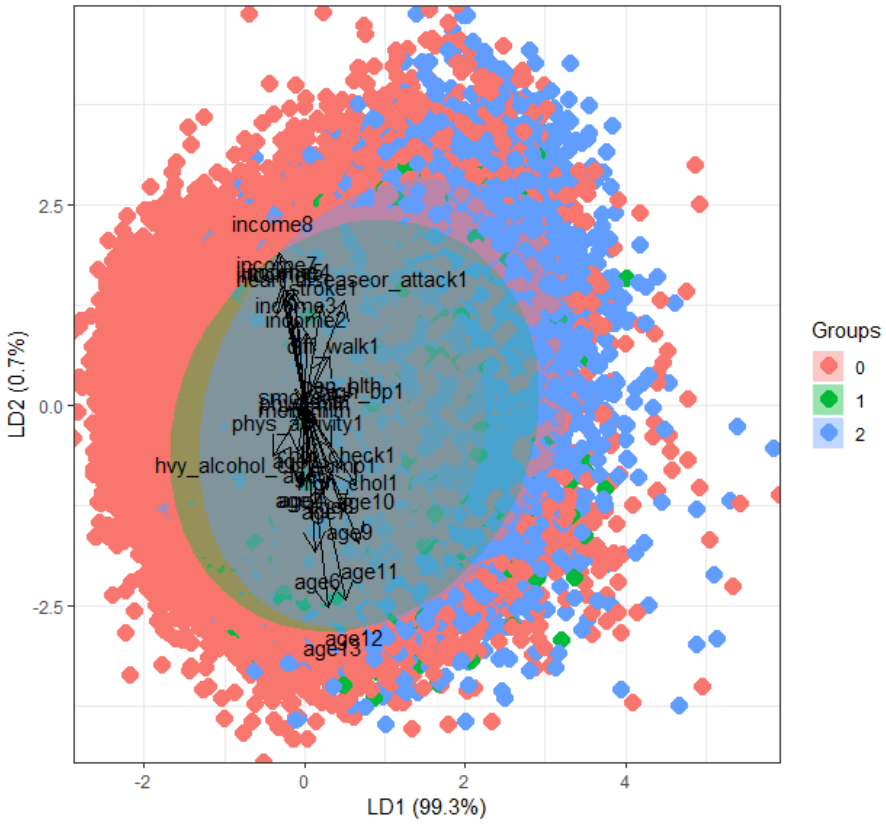


Hình 22: Đồ thị histlda tương ứng với trục LD1 của mô hình 2.



Hình 23: Đồ thị histlda tương ứng với trục LD2 của mô hình 2.

Một cách trực quan khác, ta có thể sử dụng đồ thị Biplot như hình 24. Từ hình vẽ, ta có thể thấy rằng không có sự phân biệt rõ giữa các nhóm. Bên cạnh đó ta khó có thể thấy được các biến có ảnh hưởng đến khả năng phân loại của một nhóm bất kỳ.



Hình 24: Đồ thị Biplot của mô hình 2.

Tiên đoán trên tập testing, ta thu được một số kết quả tính toán sau:

Bảng 59: Confusion Matrix giữa thực tế và dự đoán của mô hình 2.

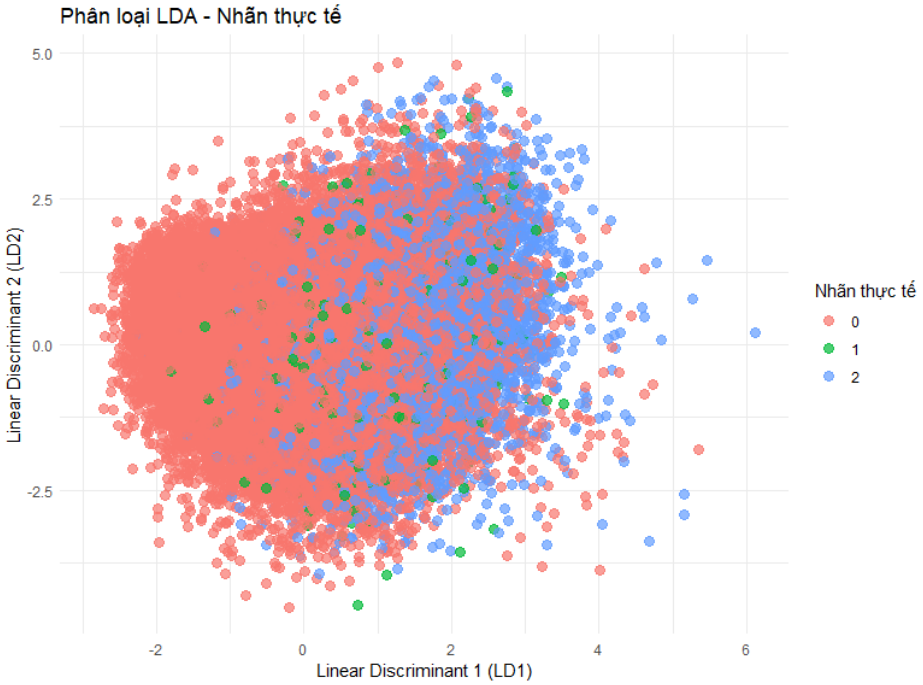
Actual	**Predict**			**Tổng**
	Nhóm 0	**Nhóm 1**	**Nhóm 2**	
Nhóm 0	55161	1209	8400	64770
Nhóm 1	0	0	0	0
Nhóm 2	1957	130	2171	4258
Tổng	57118	1339	10571	69028

Do mô hình 2 không có gán bất kỳ nhãn 1 nào cho quan sát trong tập testing nên do đó, ta không có tính được chỉ số Precision (Bảng 64). Ta cũng tính được chỉ số dự đoán chính xác các lớp của mô hình 2 là khoảng 80.1%. Đây là một kết quả khá tốt. Việc mô hình không

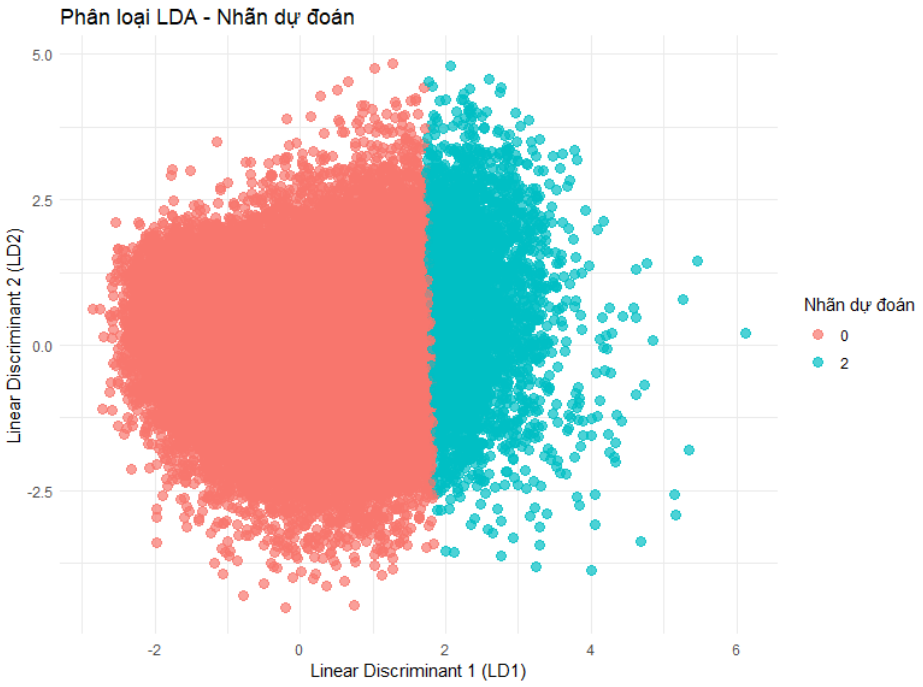
Bảng 60: Một số chỉ số đánh giá phân loại của mô hình 2.

	Nhóm 0	**Nhóm 1**	**Nhóm 2**
Precision	0.8516	–	0.5099
Recall	0.9657	0	0.2054

thể gán nhãn nhóm 1 có thể do việc ta chia phần trăm dữ liệu ban đầu hoặc do dữ liệu của nhóm 1 không đủ nhiều cho mô hình huấn luyện nên việc tiên đoán có thể thiếu chính xác.



Hình 25: Trực quan quan sát thực tế trên tập testing.



Hình 26: Trực quan quan sát dự đoán trên tập testing.

Biến đổi thông tin biến `Diabetes_012`: Ta ghép nhóm 1 và 2 của biến `Diabetes_012` của bộ dữ liệu gốc ban đầu vào chung 1 nhóm và ta có sự gán nhãn 2 nhóm mới:

- Nhóm 0 - nhóm người không bị bệnh tiểu đường;
- Nhóm 1 - nhóm người bị mắc bệnh tiểu đường.

Mô hình 3: Mô hình LDA được xây dựng dựa trên 15 biến đã chọn lọc từ bộ dữ liệu gốc (không cân bằng) để phân loại 2 nhóm của biến mục tiêu `Diabetes_012`.

Từ kết quả ước lượng 61, ta thấy trong tập training:

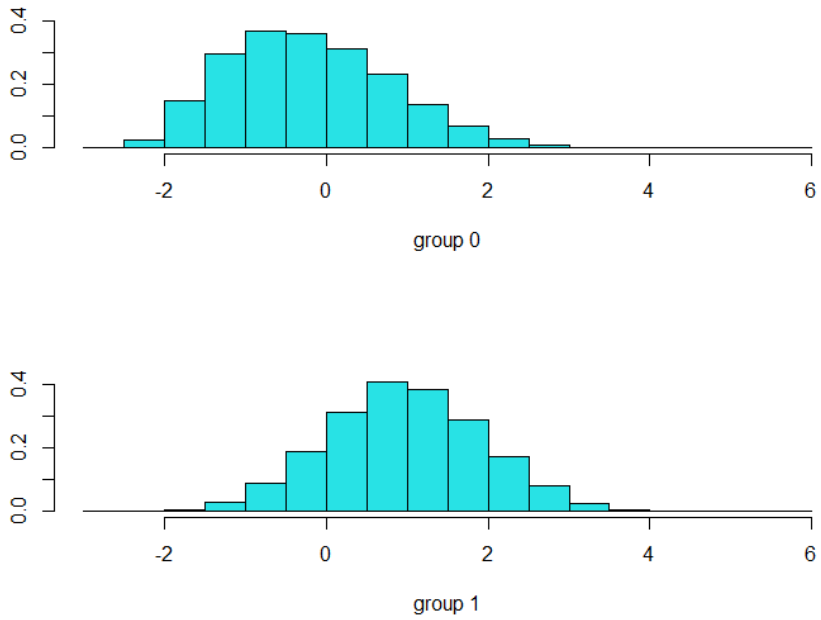
- Có khoảng 82.7% quan sát được phân loại vào nhóm 0, tức là không có mắc bệnh tiểu đường;
- Có khoảng 17.3% quan sát được phân loại vào nhóm 1, tức là có mắc bệnh tiểu đường;

Bảng 61: Xác suất tiên nghiệm của mô hình 3.

Nhóm 0	**Nhóm 1**
0.82696435	0.1730357

Do biến mục tiêu ta đang xét có hai nhóm phân loại nên trong kết quả ước lượng LDA, ta chú ý sẽ có thông tin phần trăm khả năng phân loại của một trục phân biệt duy nhất là LD1. Để đánh giá được khả năng phân loại của trục, ta sử dụng đồ thị histogram dành riêng cho LDA (histlda) thông qua hàm `ldahist` được hỗ trợ trong cùng thư viện mô hình.

Từ hình 27, ta nhận thấy rằng có sự trùng lặp đáng kể giữa các đồ thị histogram của 2 nhóm, do đó trục LD1 của mô hình 3 không có sự phân loại tốt giữa 2 nhóm.



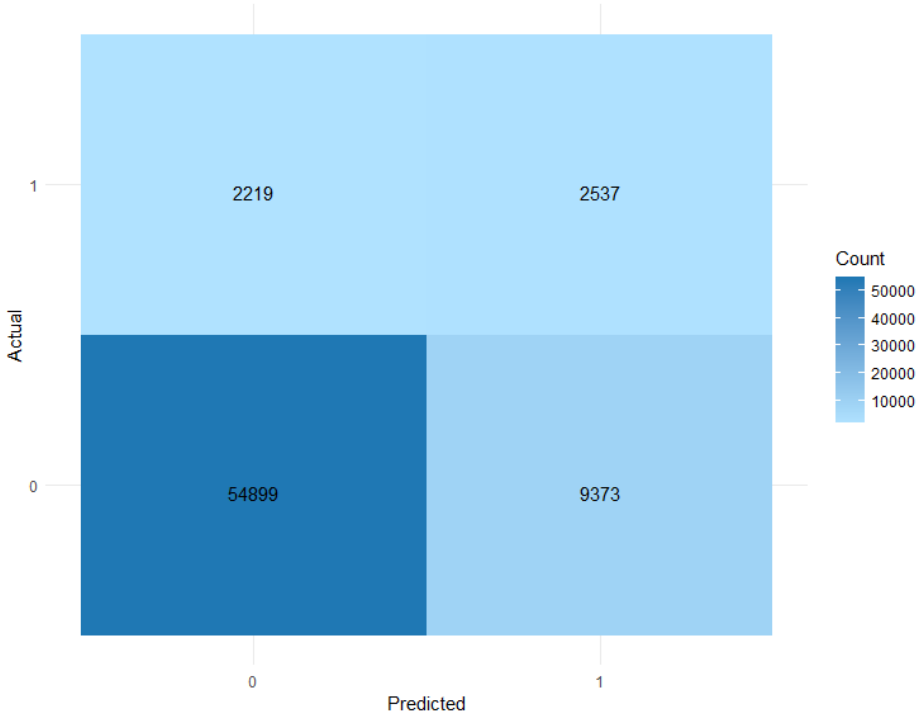
Hình 27: Đồ thị histlda tương ứng với trục LD1 của mô hình 3.

Tiên đoán trên tập testing, ta thu được một số kết quả tính toán như ma trận nhầm lẫn (Hình 28). Từ hình vẽ ta có thể thấy mô hình phân loại tốt nhóm 0 với mật độ khá cao, tuy nhiên đối với nhóm 1 mô hình có vẻ phân loại khá yếu. Một số chỉ số đánh giá từ Bảng 62 càng thêm củng cố cho phần nhận xét của ta.

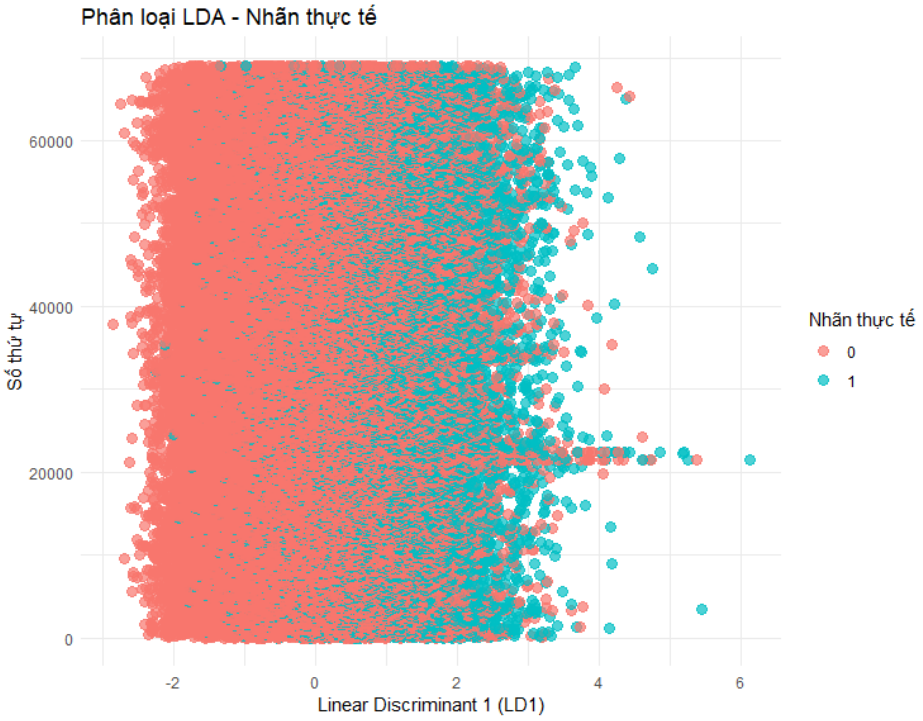
Bảng 62: Một số chỉ số đánh giá phân loại của mô hình 3.

	Nhóm 0	**Nhóm 1**
Precision	0.5812	0.5334
Recall	0.9612	0.2130

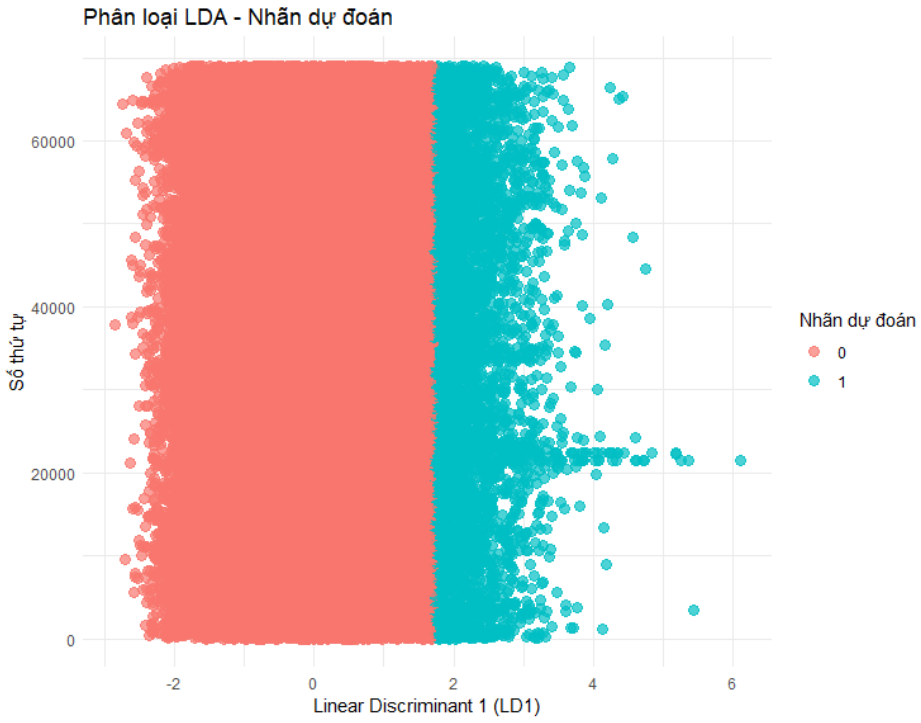
Ta tính được chỉ số dự đoán chính xác các lớp của mô hình 3 là khoảng 83.21%. Đây là một kết quả khá tốt đối với mô hình 3 nói riêng và tốt hơn so với hai mô hình trước đó. Điều đó cho ta thấy rằng việc phân loại trên hai nhóm có vẻ tốt hơn so với việc phân loại 3 nhóm.



Hình 28: Ma trận nhầm lẫn của mô hình 3.



Hình 29: Trực quan quan sát thực tế trên tập testing.



Hình 30: Trực quan quan sát dự đoán trên tập testing.

Cân bằng dữ liệu: Ta tiến hành cân bằng dữ liệu các nhóm của biến `Diabetes_012` trong cả hai trường hợp bằng phương pháp **SMOTE** được thi hành trong `R`:

- Biến `Diabetes_012` có 3 nhóm mức độ:

Bảng 63: Số lượng các nhóm (3 nhóm) trước và sau khi cân bằng.

	Nhóm 0	**Nhóm 1**	**Nhóm 2**
Trước khi cân bằng	190,055	4,629	35,097
Sau khi cân bằng	190,055	190,055	190,055

- Biến `Diabetes_012` có 2 nhóm mức độ:

Bảng 64: Số lượng các nhóm (2 nhóm) trước và sau khi cân bằng.

	Nhóm 0	**Nhóm 1**
Trước khi cân bằng	190,055	39,726
Sau khi cân bằng	190,055	190,055

Mô hình 4: Mô hình LDA được xây dựng dựa trên 15 biến đã chọn lọc từ bộ dữ liệu đã cân bằng để phân loại 3 nhóm của biến mục tiêu `Diabetes_012`.

Từ kết quả ước lượng 65, ta thấy trong tập training:

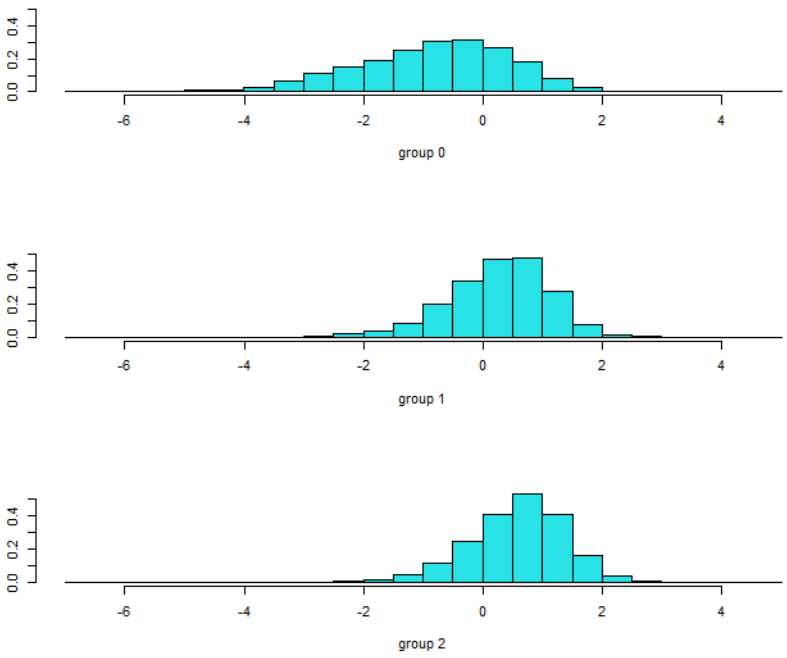
- Có khoảng 33.3% quan sát được phân loại vào nhóm 0, tức là không có mắc bệnh tiểu đường;
- Có khoảng 33.31% quan sát được phân loại vào nhóm 1, tức tiền tiểu đường;
- Có khoảng 33.4% quan sát được phân loại vào nhóm 2, tức là có mắc bệnh tiểu đường;

Bảng 65: Xác suất tiên nghiệm của mô hình 2.

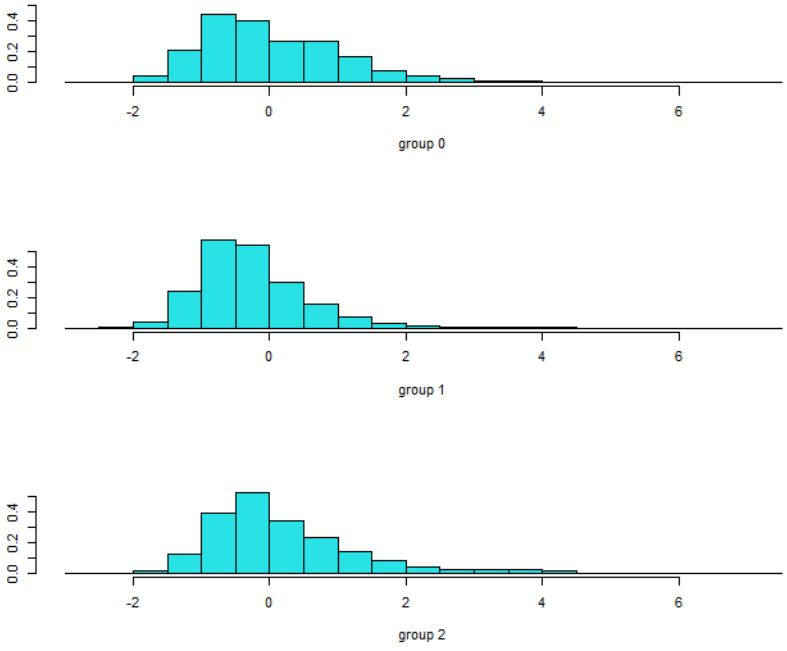
Nhóm 0	**Nhóm 1**	**Nhóm 2**
0.3329701	0.3330503	0.3339796

Do biến mục tiêu ta đang xét có ba nhóm phân loại nên trong kết quả ước lượng LDA, ta chú ý sẽ có thông tin phần trăm khả năng phân loại của hai trục phân biệt LD1 và LD2. Đối với mô hình này, ta có 91.23% dữ liệu được phân loại bằng trục LD1 và 8.77% dữ liệu được phân loại từ trục LD2. Để đánh giá được khả năng phân loại của hai trục, ta sử dụng đồ thị histogram dành riêng cho LDA (histlda) thông qua hàm `ldahist` được hỗ trợ trong cùng thư viện mô hình.

Từ hình 31, ta nhận thấy rằng có sự trùng lặp đáng kể giữa các đồ thị histogram của 3 nhóm, do đó trục LD1 của mô hình 4 không có sự phân loại tốt giữa 3 nhóm. Nhận xét tương tự cho trục LD2.

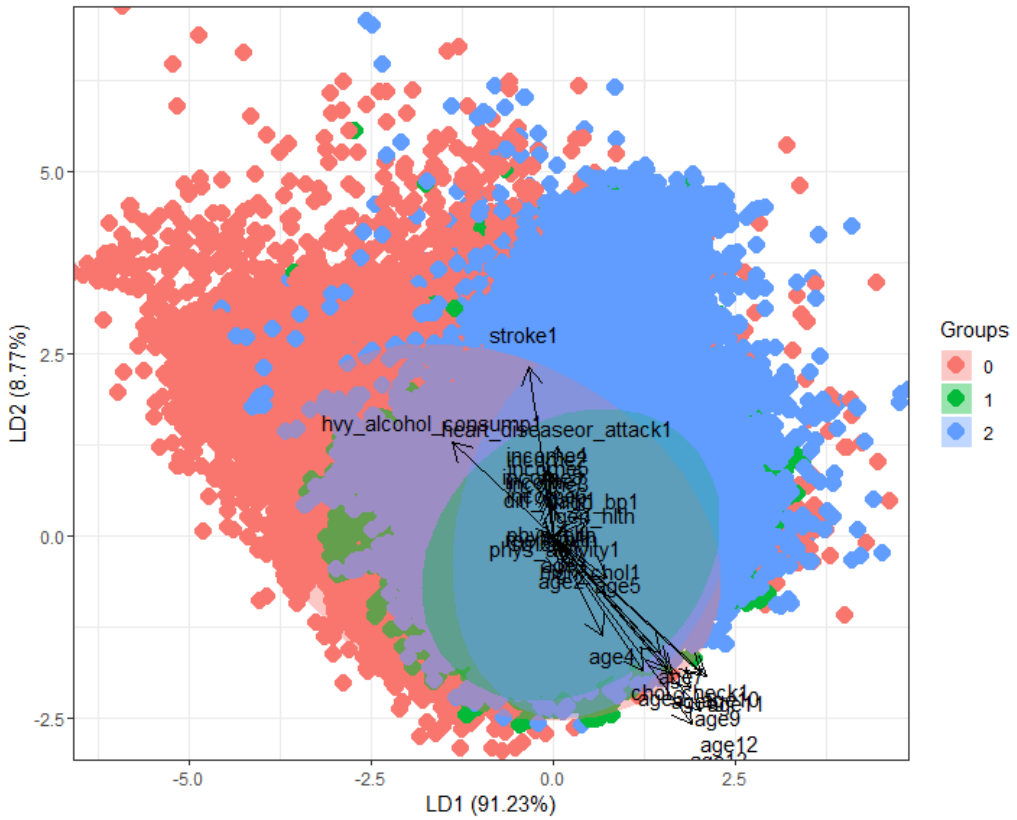


Hình 31: Đồ thị histlda tương ứng với trục LD1 của mô hình 4.



Hình 32: Đồ thị histlda tương ứng với trục LD2 của mô hình 4.

Một cách trực quan khác, ta có thể sử dụng đồ thị Biplot như hình 33. Từ hình vẽ, ta có thể thấy rằng không có sự phân biệt rõ giữa các nhóm. Bên cạnh đó ta khó có thể thấy được các biến có ảnh hưởng đến khả năng phân loại của một nhóm bất kỳ.



Hình 33: Đồ thị Biplot của mô hình 4.

Tiên đoán trên tập testing, ta thu được một số kết quả tính toán sau:

Bảng 66: Confusion Matrix giữa thực tế và dự đoán của mô hình 4.

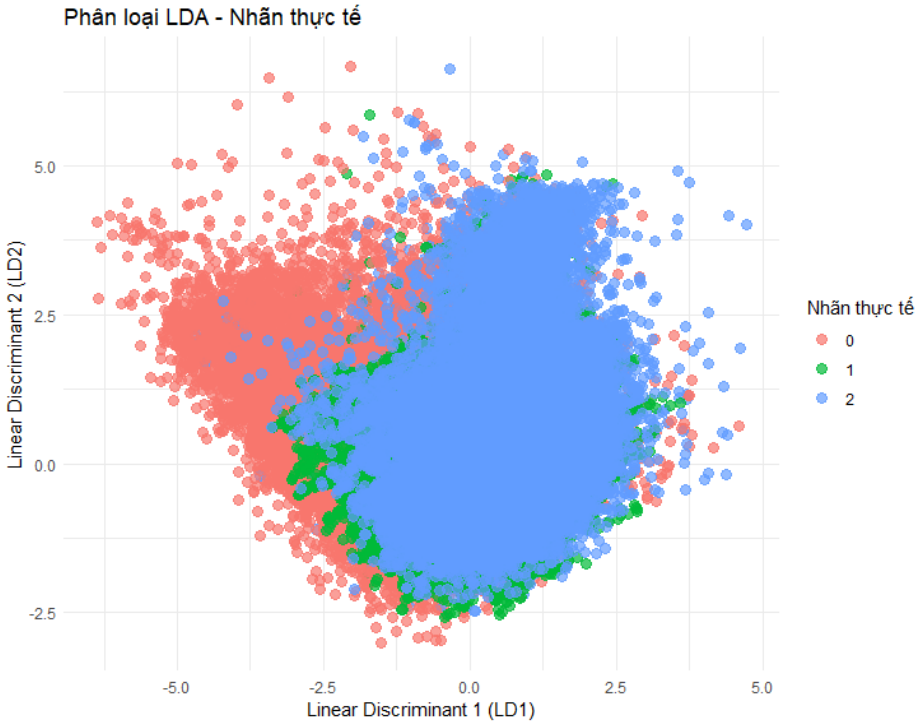
Actual	**Predict**			**Tổng**
	Nhóm 0	**Nhóm 1**	**Nhóm 2**	
Nhóm 0	33331	10930	6553	50814
Nhóm 1	14937	27760	19923	62620
Nhóm 2	8850	18396	30239	30239
Tổng	57118	57086	56715	170919

Ta cũng tính được chỉ số dự đoán chính xác các lớp của mô hình 4 là khoảng 53.4%. Đây là một kết quả tương đối, tức là mô hình không thể có sự phân biệt chính xác một nhóm nào

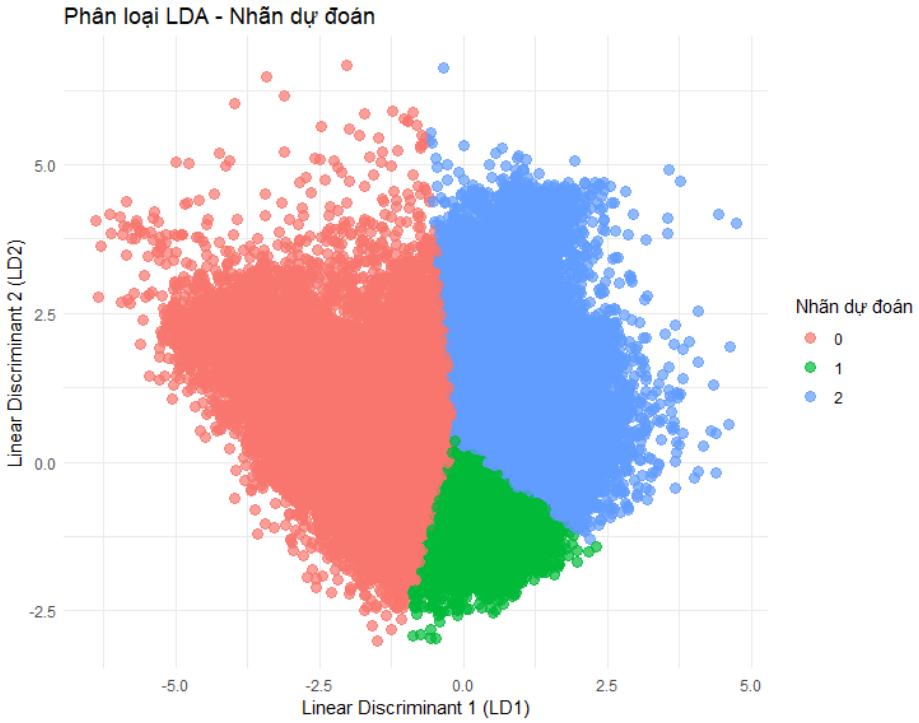
Bảng 67: Một số chỉ số đánh giá phân loại của mô hình 4.

	Nhóm 0	**Nhóm 1**	**Nhóm 2**
Precision	0.6559	0.4433	0.5260
Recall	0.5835	0.4863	0.5332

hết. Mỗi nhóm mà mô hình dự đoán đều có khả năng chính xác cũng như sai lầm như nhau. Chứng tỏ mô hình xây dựng trên bộ dữ liệu cân bằng không tốt.



Hình 34: Trực quan quan sát thực tế trên tập testing.



Hình 35: Trực quan quan sát dự đoán trên tập testing.

Mô hình 5: Mô hình LDA được xây dựng dựa trên 15 biến đã chọn lọc từ bộ dữ liệu đã cân bằng để phân loại 2 nhóm của biến mục tiêu `Diabetes_012`.

Từ kết quả ước lượng 68, ta thấy trong tập training:

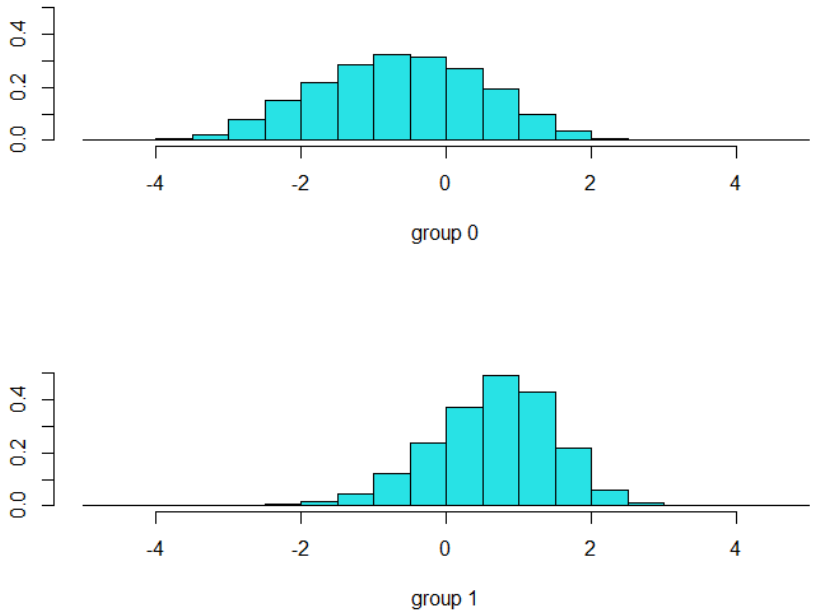
- Có khoảng 50.01% quan sát được phân loại vào nhóm 0, tức là không có mắc bệnh tiểu đường;
- Có khoảng 49.99% quan sát được phân loại vào nhóm 1, tức là có mắc bệnh tiểu đường;

Bảng 68: Xác suất tiên nghiệm của mô hình 5.

Nhóm 0	**Nhóm 1**
0.5000658	0.4999342

Do biến mục tiêu ta đang xét có hai nhóm phân loại nên trong kết quả ước lượng LDA, ta chú ý sẽ có thông tin phần trăm khả năng phân loại của một trục phân biệt duy nhất là LD1. Để đánh giá được khả năng phân loại của trục, ta sử dụng đồ thị histogram dành riêng cho LDA (histlda) thông qua hàm `ldahist` được hỗ trợ trong cùng thư viện mô hình.

Từ hình 36, ta nhận thấy rằng có sự trùng lặp đáng kể giữa các đồ thị histogram của 2 nhóm, do đó trục LD1 của mô hình 5 không có sự phân loại tốt giữa 2 nhóm.



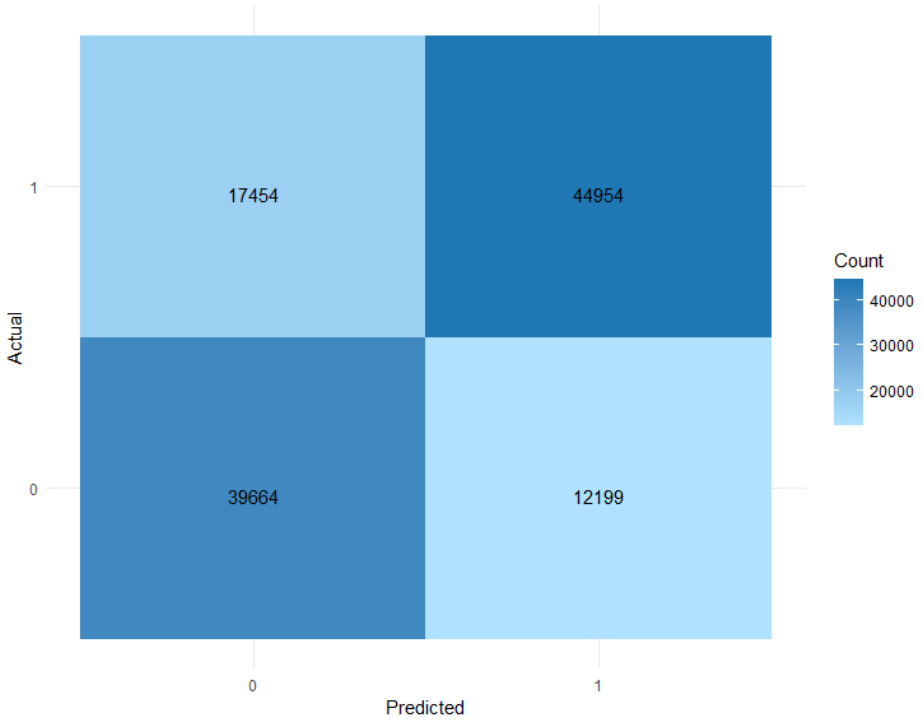
Hình 36: Đồ thị histlda tương ứng với trục LD1 của mô hình 5.

Tiên đoán trên tập testing, ta thu được một số kết quả tính toán như ma trận nhầm lẫn (Hình 37). Từ hình vẽ ta có thể thấy mô hình phân loại tốt nhóm 0 với mật độ khá cao, tuy nhiên đối với nhóm 1 mô hình có vẻ phân loại khá yếu. Một số chỉ số đánh giá từ bảng 69 càng thêm củng cố cho phần nhận xét của ta.

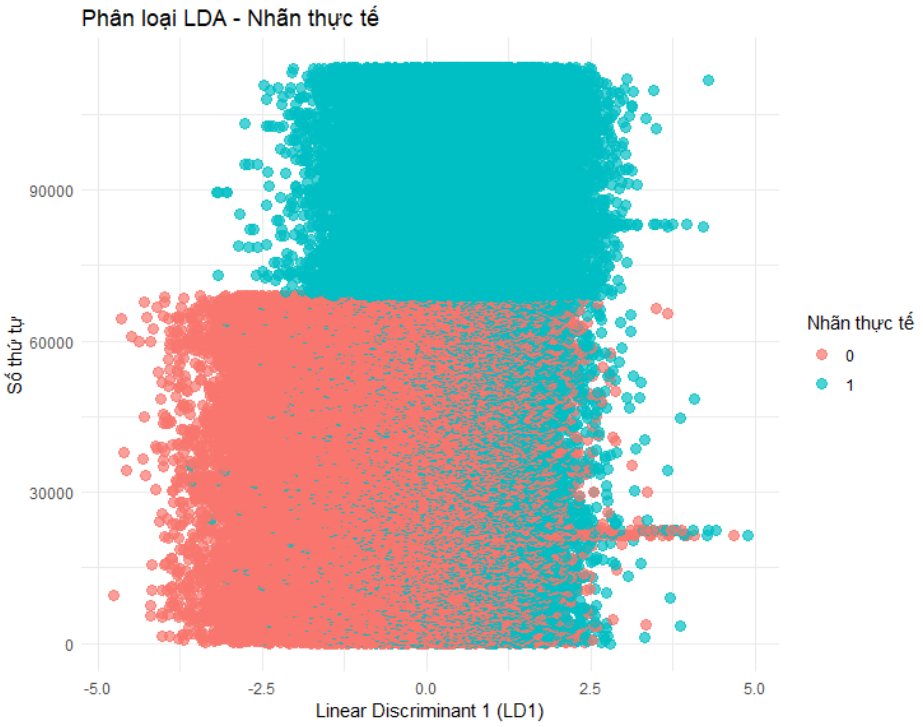
Bảng 69: Một số chỉ số đánh giá phân loại của mô hình 5.

	Nhóm 0	**Nhóm 1**
Precision	0.7648	0.7203
Recall	0.6944	0.7866

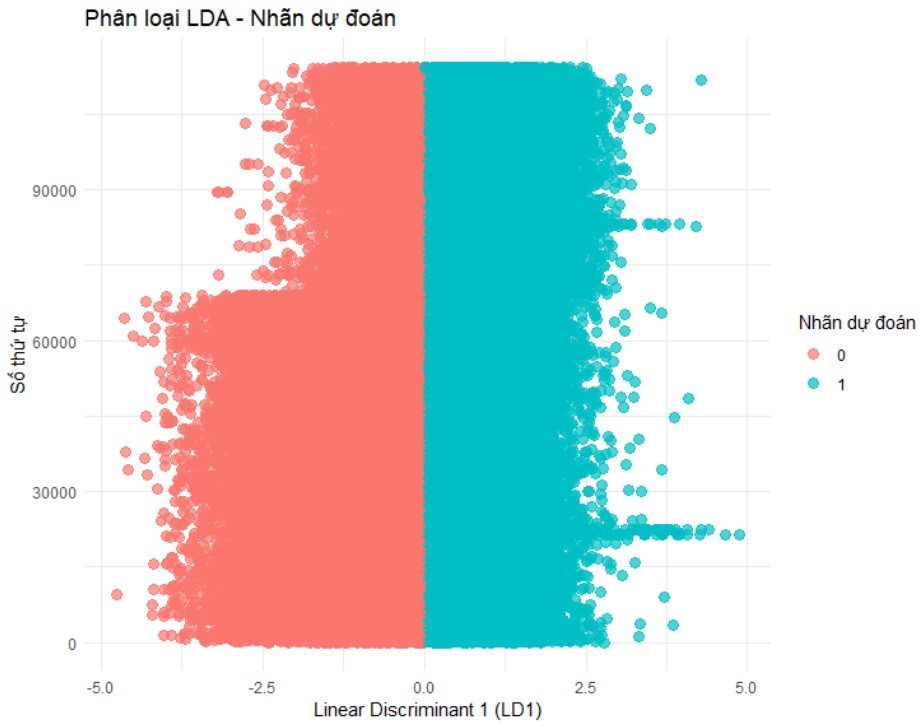
Ta tính được chỉ số dự đoán chính xác các lớp của mô hình 5 là khoảng 74.05%. Đây là một kết quả phân loại ở mức khá. Tuy nhiên mô hình 5 không tốt hơn so với mô hình 3. Do đó, ta có thể thấy việc áp dụng cân bằng mẫu có thể làm cho mô hình huấn luyện không tốt, dẫn đến kết quả tiên đoán giảm độ chính xác.



Hình 37: Ma trận nhầm lẫn của mô hình 5.



Hình 38: Trực quan quan sát thực tế trên tập testing.



Hình 39: Trực quan quan sát dự đoán trên tập testing.

9.3 QDA

Mô hình 1: Mô hình QDA được xây dựng dựa trên tất cả 22 biến từ bộ dữ liệu gốc (không cân bằng) để phân loại 3 nhóm của biến mục tiêu `Diabetes_012`.

Từ kết quả ước lượng 70, ta thấy trong tập training:

- Có khoảng 82.7% quan sát được phân loại vào nhóm 0, tức là không có mắc bệnh tiểu đường;
- Có khoảng 2.05% quan sát được phân loại vào nhóm 1, tức tiền tiểu đường;
- Có khoảng 15.26% quan sát được phân loại vào nhóm 2, tức là có mắc bệnh tiểu đường;

Bảng 70: Xác suất tiên nghiệm của mô hình 1.

Nhóm 0	**Nhóm 1**	**Nhóm 2**
0.82696435	0.02046618	0.15256947

Một điều chú ý rằng, kết quả ước lượng của mô hình 1 QDA giống với kết quả mô hình 1 của LDA. Tuy nhiên, ta không thể trực quan hóa các giá trị được phân loại như thế nào trong mô hình QDA vì ta không thể xác định được các trục phân loại tuyến tính như LDA thông thường. Do đó, ta chỉ có thể đánh giá mô hình thông qua việc tiên đoán các lớp của biến mục tiêu.

Tiên đoán trên tập testing, ta thu được một số kết quả tính toán sau:

Bảng 71: Confusion Matrix giữa thực tế và dự đoán của mô hình 1.

Actual	**Predict**			**Tổng**
	Nhóm 0	**Nhóm 1**	**Nhóm 2**	
Nhóm 0	32451	397	1807	34655
Nhóm 1	788	28	172	988
Nhóm 2	23879	914	8592	33385
Tổng	57118	1339	10571	69028

Bảng 72: Một số chỉ số đánh giá phân loại của mô hình 1.

	Nhóm 0	**Nhóm 1**	**Nhóm 2**
Precision	0.9364	0.0283	0.2574
Recall	0.5681	0.0209	0.8128

Ta cũng tính được chỉ số dự đoán chính xác các lớp của mô hình 1 là khoảng 59.5%. Đây là một kết quả phân loại trung bình. So với mô hình 1 của LDA, tuy đã có thể gán nhãn nhóm 1 nhưng khả năng dự đoán chính xác của mô hình đã giảm xuống hẳn.

Mô hình 2: Mô hình QDA được xây dựng dựa trên 15 biến đã chọn lọc từ bộ dữ liệu gốc (không cân bằng) để phân loại 3 nhóm của biến mục tiêu `Diabetes_012`.

Từ kết quả ước lượng 73, ta thấy trong tập training:

- Có khoảng 82.7% quan sát được phân loại vào nhóm 0, tức là không có mắc bệnh tiểu đường;
- Có khoảng 2.05% quan sát được phân loại vào nhóm 1, tức tiền tiểu đường;
- Có khoảng 15.26% quan sát được phân loại vào nhóm 2, tức là có mắc bệnh tiểu đường;

Bảng 73: Xác suất tiên nghiệm của mô hình 2.

Nhóm 0	**Nhóm 1**	**Nhóm 2**
0.82696435	0.02046618	0.15256947

Một điều chú ý rằng, kết quả ước lượng của mô hình 2 QDA giống với kết quả mô hình 2 của LDA. Tuy nhiên, ta không thể trực quan hóa các giá trị được phân loại như thế nào trong mô hình QDA vì ta không thể xác định được các trục phân loại tuyến tính như LDA

thông thường. Do đó, ta chỉ có thể đánh giá mô hình thông qua việc tiên đoán các lớp của biến mục tiêu.

Tiên đoán trên tập testing, ta thu được một số kết quả tính toán sau:

Bảng 74: Confusion Matrix giữa thực tế và dự đoán của mô hình 1.

Actual	**Predict**			**Tổng**
	Nhóm 0	**Nhóm 1**	**Nhóm 2**	
Nhóm 0	28500	316	1361	30177
Nhóm 1	286	11	60	357
Nhóm 2	28332	1012	9150	38494
Tổng	57118	1339	10571	69028

Bảng 75: Một số chỉ số đánh giá phân loại của mô hình 1.

	Nhóm 0	**Nhóm 1**	**Nhóm 2**
Precision	0.9444	0.0308	0.2377
Recall	0.499	0.0082	0.8656

Ta cũng tính được chỉ số dự đoán chính xác các lớp của mô hình 1 là khoảng 54.56%. Đây là một kết quả phân loại trung bình. So với mô hình 2 của LDA, tuy đã có thể gán nhãn nhóm 1 nhưng khả năng dự đoán chính xác của mô hình đã giảm xuống hẳn.

Mô hình 3: Mô hình QDA được xây dựng dựa trên 15 biến đã chọn lọc từ bộ dữ liệu gốc (không cân bằng) để phân loại 2 nhóm của biến mục tiêu `Diabetes_012`.

Từ kết quả ước lượng 76, ta thấy trong tập training:

- Có khoảng 82.7% quan sát được phân loại vào nhóm 0, tức là không có mắc bệnh tiểu đường;
- Có khoảng 17.3% quan sát được phân loại vào nhóm 1, tức là có mắc bệnh tiểu đường;

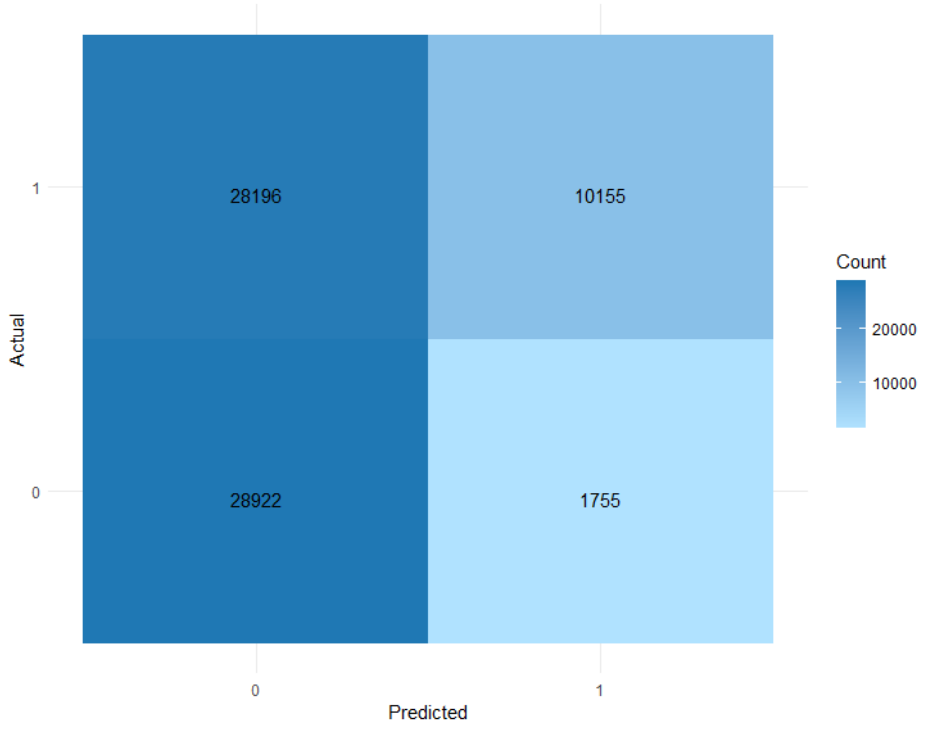
Bảng 76: Xác suất tiên nghiệm của mô hình 3.

Nhóm 0	**Nhóm 1**
0.8269643	0.1730357

Một điều chú ý rằng, kết quả ước lượng của mô hình 3 QDA giống với kết quả mô hình 3 của LDA. Tuy nhiên, ta không thể trực quan hóa các giá trị được phân loại như thế nào

trong mô hình QDA vì ta không thể xác định được các trục phân loại tuyến tính như LDA thông thường. Do đó, ta chỉ có thể đánh giá mô hình thông qua việc tiên đoán các lớp của biến mục tiêu.

Tiên đoán trên tập testing, ta thu được một số kết quả tính toán như ma trận nhầm lẫn (Hình 40). Từ hình vẽ ta có thể thấy mô hình phân loại tốt nhóm 0 với mật độ khá cao, tuy nhiên đối với nhóm 1 mô hình có vẻ phân loại khá yếu. Một số chỉ số đánh giá từ Bảng 77 càng thêm củng cố cho phần nhận xét của ta.



Hình 40: Ma trận nhầm lẫn của mô hình 3.

Bảng 77: Một số chỉ số đánh giá phân loại của mô hình 3.

	Nhóm 0	**Nhóm 1**
Precision	0.9428	0.2648
Recall	0.5064	0.8526

Ta cũng tính được chỉ số dự đoán chính xác các lớp của mô hình 3 là khoảng 56.61%. Đây là một kết quả phân loại trung bình. So với mô hình 3 của LDA, độ chính xác phân loại của mô hình đã giảm xuống hẳn.

Mô hình 4: Mô hình QDA được xây dựng dựa trên 15 biến đã chọn lọc từ bộ dữ liệu đã cân bằng để phân loại 3 nhóm của biến mục tiêu `Diabetes_012`.

Từ kết quả ước lượng 78, ta thấy trong tập training:

- Có khoảng 33.3% quan sát được phân loại vào nhóm 0, tức là không có mắc bệnh tiểu đường;
- Có khoảng 33.31% quan sát được phân loại vào nhóm 1, tức tiền tiểu đường;
- Có khoảng 33.4% quan sát được phân loại vào nhóm 2, tức là có mắc bệnh tiểu đường;

Bảng 78: Xác suất tiên nghiệm của mô hình 4.

Nhóm 0	**Nhóm 1**	**Nhóm 2**
0.3329701	0.3330503	0.3339796

Một điều chú ý rằng, kết quả ước lượng của mô hình 4 QDA giống với kết quả mô hình 4 của LDA. Tuy nhiên, ta không thể trực quan hóa các giá trị được phân loại như thế nào trong mô hình QDA vì ta không thể xác định được các trục phân loại tuyến tính như LDA thông thường. Do đó, ta chỉ có thể đánh giá mô hình thông qua việc tiên đoán các lớp của biến mục tiêu.

Tiên đoán trên tập testing, ta thu được một số kết quả tính toán sau:

Bảng 79: Confusion Matrix giữa thực tế và dự đoán của mô hình 4.

Actual	**Predict**			**Tổng**
	Nhóm 0	**Nhóm 1**	**Nhóm 2**	
Nhóm 0	18282	3411	2806	24499
Nhóm 1	27444	36931	25279	89654
Nhóm 2	11392	16744	28630	56766
Tổng	57118	57086	56715	170919

Bảng 80: Một số chỉ số đánh giá phân loại của mô hình 4.

	Nhóm 0	**Nhóm 1**	**Nhóm 2**
Precision	0.7462	0.4119	0.5044
Recall	0.3200	0.6469	0.5048

Ta cũng tính được chỉ số dự đoán chính xác các lớp của mô hình 4 là khoảng 49.05%. Đây là một kết quả phân loại trung bình. So với mô hình 4 của LDA, tuy đã có thể gán nhãn nhóm 1 nhưng khả năng dự đoán chính xác của mô hình đã giảm xuống hẳn.

Mô hình 5: Mô hình QDA được xây dựng dựa trên 15 biến đã chọn lọc từ bộ dữ liệu đã cân bằng để phân loại 2 nhóm của biến mục tiêu `Diabetes_012`.

Từ kết quả ước lượng 81, ta thấy trong tập training:

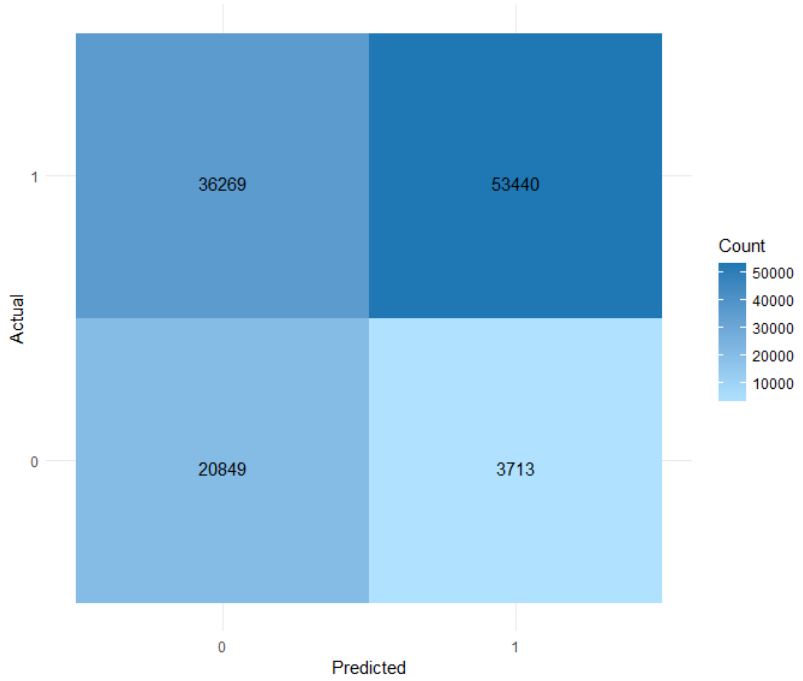
- Có khoảng 50.01% quan sát được phân loại vào nhóm 0, tức là không có mắc bệnh tiểu đường;
- Có khoảng 49.99% quan sát được phân loại vào nhóm 1, tức là có mắc bệnh tiểu đường;

Bảng 81: Xác suất tiên nghiệm của mô hình 5.

Nhóm 0	**Nhóm 1**
0.5000658	0.4999342

Một điều chú ý rằng, kết quả ước lượng của mô hình 5 QDA giống với kết quả mô hình 5 của LDA. Tuy nhiên, ta không thể trực quan hóa các giá trị được phân loại như thế nào trong mô hình QDA vì ta không thể xác định được các trục phân loại tuyến tính như LDA thông thường. Do đó, ta chỉ có thể đánh giá mô hình thông qua việc tiên đoán các lớp của biến mục tiêu.

Tiên đoán trên tập testing, ta thu được một số kết quả tính toán như ma trận nhầm lẫn (Hình 41). Từ hình vẽ ta có thể thấy mô hình phân loại tốt nhóm 1 với mật độ khá cao. Đối với nhóm 0 mô hình cũng phân loại khá. Tuy nhiên, ta có thể thấy khoảng hơn $36,269$ quan sát được gán nhãn 0 trong khi thực tế họ có nhãn là 1, tức gán nhãn người bệnh là không bị bệnh. Đó là một kết quả phân loại không tốt. Bảng 82 củng cố thêm cho phần nhận xét trên.



Hình 41: Ma trận nhầm lẫn của mô hình 5.

Bảng 82: Một số chỉ số đánh giá phân loại của mô hình 5.

	Nhóm 0	**Nhóm 1**
Precision	0.8488	0.5957
Recall	0.3650	0.9350

Ta cũng tính được chỉ số dự đoán chính xác các lớp của mô hình 5 là khoảng 65.01%. Đây là một kết quả phân loại trung bình. So với mô hình 5 của LDA, độ chính xác phân loại của mô hình đã giảm xuống hẳn.

9.4 Kết luận

Từ các kết quả ước lượng cũng như đánh giá nội bộ mô hình và so sánh giữa các mô hình khác nhau, ta thu được một số kết quả sau:

- Mô hình LDA tốt nhất là mô hình 3 - Mô hình LDA được xây dựng dựa trên 15 biến đã chọn lọc từ bộ dữ liệu gốc (không cân bằng) để phân loại 2 nhóm của biến mục tiêu `Diabetes_012` với độ chính xác đạt được trên tập kiểm tra là 83.21%;
- Mô hình QDA tốt nhất là mô hình 5 - Mô hình QDA được xây dựng dựa trên 15 biến đã chọn lọc từ bộ dữ liệu đã cân bằng để phân loại 2 nhóm của biến mục tiêu `Diabetes_012` với độ chính xác đạt được trên tập kiểm tra là 65.01%;

Trong suốt quá trình xây dựng, ta nhận thấy kết quả ước lượng xác suất tiên nghiệm của hai kiểu mô hình LDA và QDA giống nhau. Tuy nhiên, khi kiểm tra trên tập testing, ta thấy rằng các mô hình con của LDA cho ta kết quả tốt hơn so với mô hình QDA. Mặt khác, đối với mô hình LDA, ta có thể trực quan hóa các kết quả ước lượng để hỗ trợ trong việc đánh giá. Ngoài ra, mô hình QDA tốt hơn LDA ở chổ QDA có thể học và gán nhãn được nhóm 1 trong khi mô hình LDA không thể gán nhãn, trừ khi ta tăng kích thước số lượng các nhóm lên một lượng nhất định. Với các kết quả phân tích trên, đối với bộ dữ liệu này, ta khuyến khích sử dụng mô hình LDA trong việc áp dụng phương pháp phân tích phân biệt để có thể hỗ trợ cho một số công việc hoặc các hướng nghiên cứu tiếp theo.