

BODY PERFORMANCE ANALYSIS

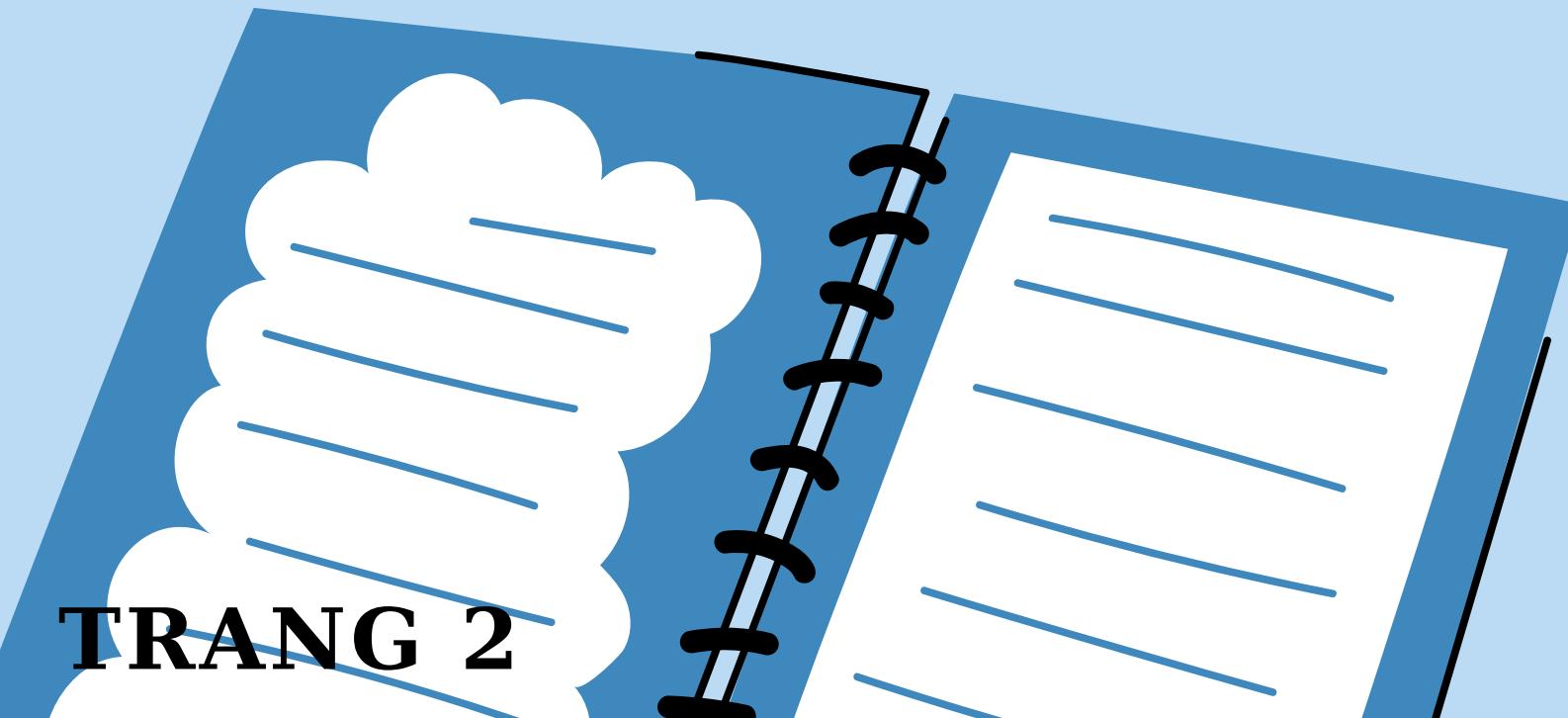
Môn: Xử Lý Số Liệu Thống Kê
Giảng viên: TS.Tô Đức Khanh

GROUP 22

Trần Duy An-22110007
Phạm Xuân Bách-22110021
Vũ Xuân Hiệp-22110060
Nguyễn Tất Chiến-22110028
Phạm Thái Thiên An-22110005



MỤC LỤC



TRANG 2

1

MỤC TIÊU DỰ ÁN

2

TỔNG QUAN VỀ DỮ LIỆU

3

QUY TRÌNH XỬ LÝ

4

KẾT QUẢ TỔNG KẾT

5

HƯỚNG PHÁT TRIỂN

I. MỤC TIÊU DỰ ÁN

1

Giúp các chuyên gia sức khoẻ biết
được hiệu quả của việc tập thể dục.

2

Các yếu tố ảnh hưởng tới hiệu quả của
việc tập thể dục là gì?

II. TỔNG QUAN DỮ LIỆU



SƠ LƯỢC VỀ DỮ LIỆU



CHI TIẾT CÁC BIỂN



TIỀN XỬ LÝ DỮ LIỆU

1. Sơ lược về dữ liệu



01

Bộ dữ liệu gồm 13,393 người tham gia thể thao tại Hàn Quốc, gồm có 12 biến đặc trưng.

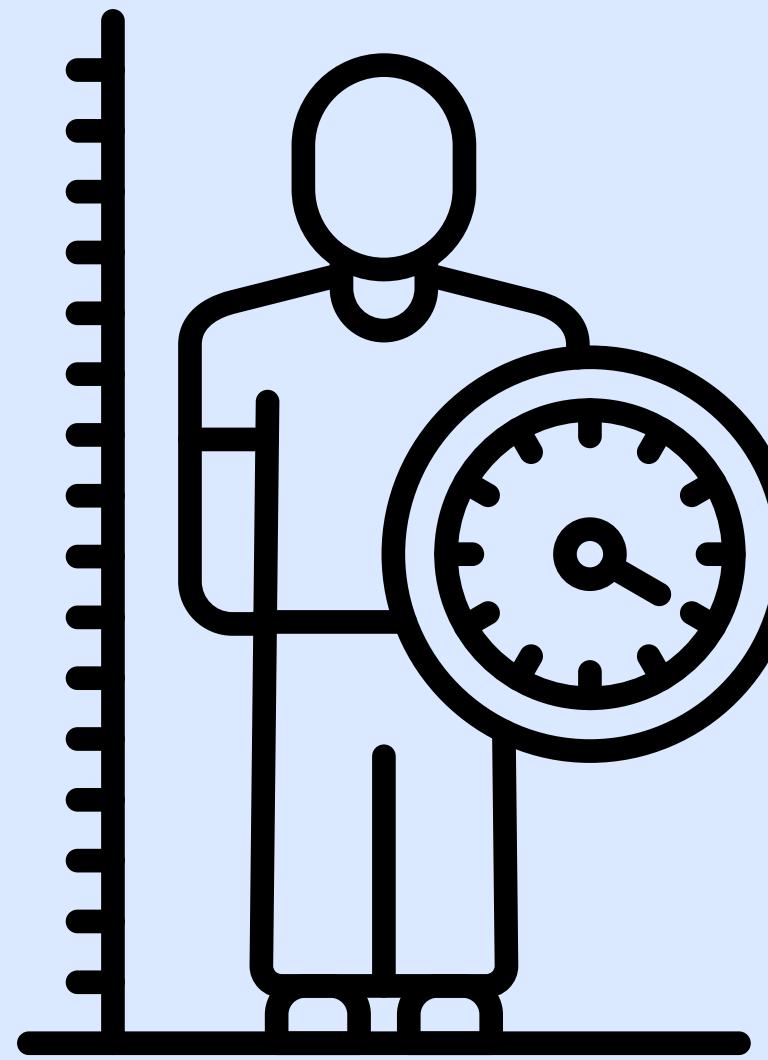
02

Mục tiêu: nghiên cứu
và đánh giá hiệu quả
luyện tập thể thao

2. Chi tiết các biến



class: phân loại hiệu suất (A: tốt nhất, B, C, D)



age: từ 20 tới 64

gender: giới tính (F: nữ, M: nam)

BMI: chỉ số giữa cân nặng và chiều cao

body fat_%: phần trăm mỡ cơ thể (%)

MAP: chỉ số biểu diễn diastolic & systolic



gripForce: lực kẹp tay (kg)



sit and bend

forward_cm: độ linh hoạt (ngồi và gập người về phía trước, cm)



sit-ups counts: số lần gập bụng



broad

jump_cm: nhảy xa (đơn vị: cm)



broad

jump_cm: nhảy xa (đơn vị: cm)

```

spc_tb1_ [13,393 x 12] (S3: spec_tb1_df/tb1_df/tb1/data.frame)
$ age                  : num [1:13393] 27 25 31 32 28 36 42 33 54 28 ...
$ gender               : chr [1:13393] "M" "M" "M" "M" ...
$ height_cm            : num [1:13393] 172 165 180 174 174 ...
$ weight_kg             : num [1:13393] 75.2 55.8 78 71.1 67.7 ...
$ body_fat_percent     : num [1:13393] 21.3 15.7 20.1 18.4 17.1 22 32.2 36.9 27.6 14.4 ...
$ diastolic             : num [1:13393] 80 77 92 76 70 64 72 84 85 81 ...
$ systolic              : num [1:13393] 130 126 152 147 127 119 135 137 165 156 ...
$ grip_force            : num [1:13393] 54.9 36.4 44.8 41.4 43.5 23.8 22.7 45.9 40.4 57.9 ...
$ sit_and_bend_forward_cm: num [1:13393] 18.4 16.3 12 15.2 27.1 21 0.8 12.3 18.6 12.1 ...
$ sit_ups_counts         : num [1:13393] 60 53 49 53 45 27 18 42 34 55 ...
$ broad_jump_cm          : num [1:13393] 217 229 181 219 217 153 146 234 148 213 ...
$ class                 : chr [1:13393] "C" "A" "C" "B" ...

```

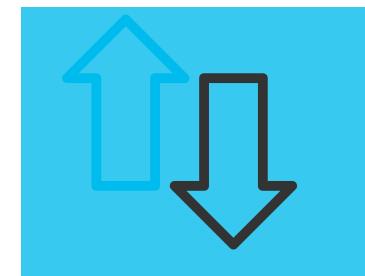
	age	gender	height_cm	weight_kg	body_fat_percent
Min.	:21.00	Length:13393	Min. :125.0	Min. : 26.30	Min. : 3.00
1st Qu.	:25.00	Class :character	1st Qu.:162.4	1st Qu.: 58.20	1st Qu.:18.00
Median	:32.00	Mode :character	Median :169.2	Median : 67.40	Median :22.80
Mean	:36.78		Mean :168.6	Mean : 67.45	Mean :23.24
3rd Qu.	:48.00		3rd Qu.:174.8	3rd Qu.: 75.30	3rd Qu.:28.00
Max.	:64.00		Max. :193.8	Max. :138.10	Max. :78.40
	diastolic	systolic	grip_force	sit_and_bend_forward_cm	sit_ups_counts
Min.	: 0.0	Min. : 0.0	Min. : 0.00	Min. :-25.00	Min. : 0.00
1st Qu.	:71.0	1st Qu.:120.0	1st Qu.:27.50	1st Qu.: 10.90	1st Qu.:30.00
Median	:79.0	Median :130.0	Median :37.90	Median : 16.20	Median :41.00
Mean	:78.8	Mean :130.2	Mean :36.96	Mean : 15.21	Mean :39.77
3rd Qu.	:86.0	3rd Qu.:141.0	3rd Qu.:45.20	3rd Qu.: 20.70	3rd Qu.:50.00
Max.	:156.2	Max. :201.0	Max. :70.50	Max. :213.00	Max. :80.00
	broad_jump_cm	class			
Min.	: 0.0	Length:13393			
1st Qu.	:162.0	Class :character			
Median	:193.0	Mode :character			
Mean	:190.1				
3rd Qu.	:221.0				
Max.	:201.0				

THỐNG KÊ MÔ TẢ

1. Tóm tắt dữ liệu

2. Summary

3. Tiền xử lý dữ liệu



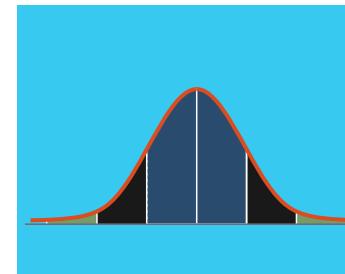
CHUYỂN ĐỔI DỮ
LIỆU



XỬ LÝ DỮ LIỆU
KHUYẾT



XỬ LÝ NGOẠI LAI



QUAN SÁT PHÂN
PHỐI



XỬ LÝ DỮ LIỆU
BÁT THƯỜNG

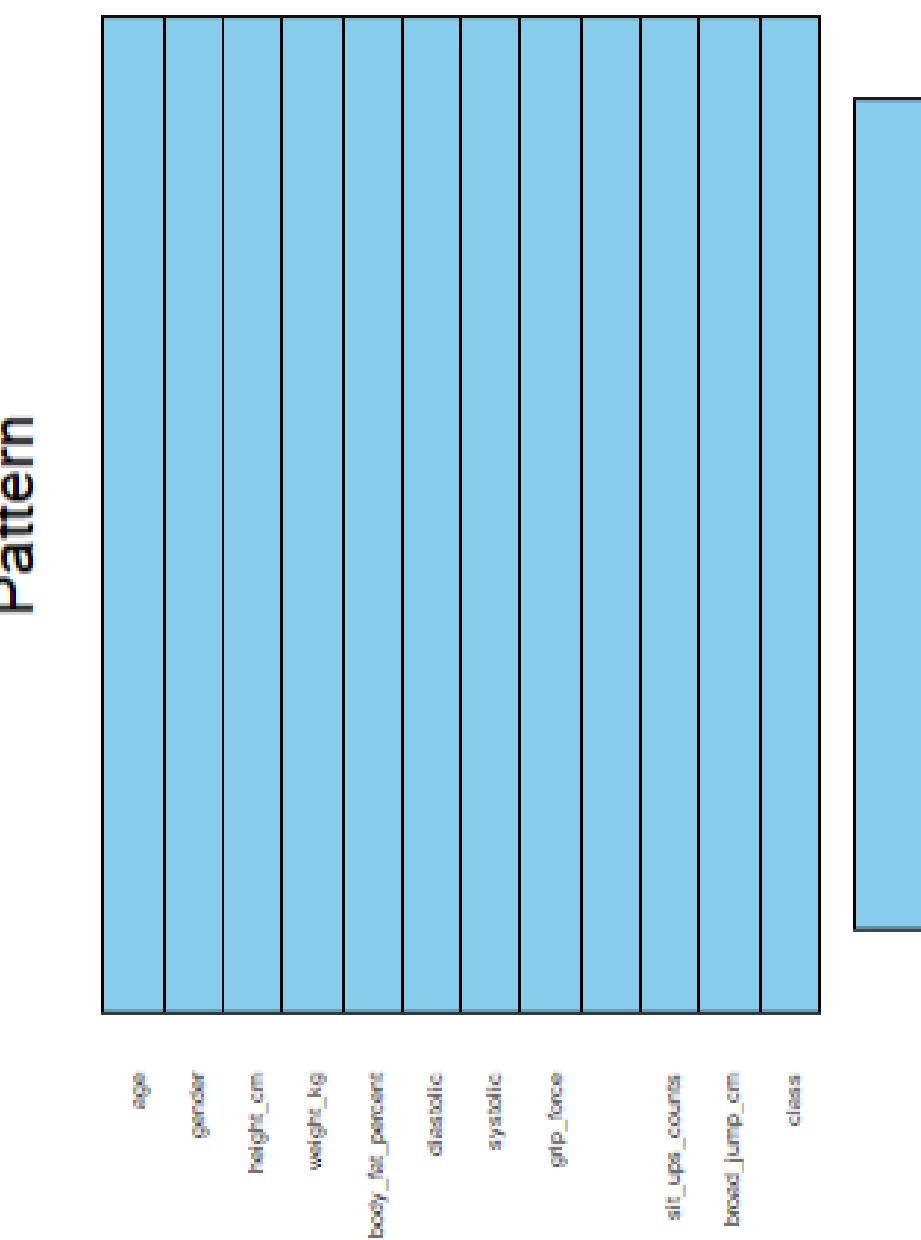
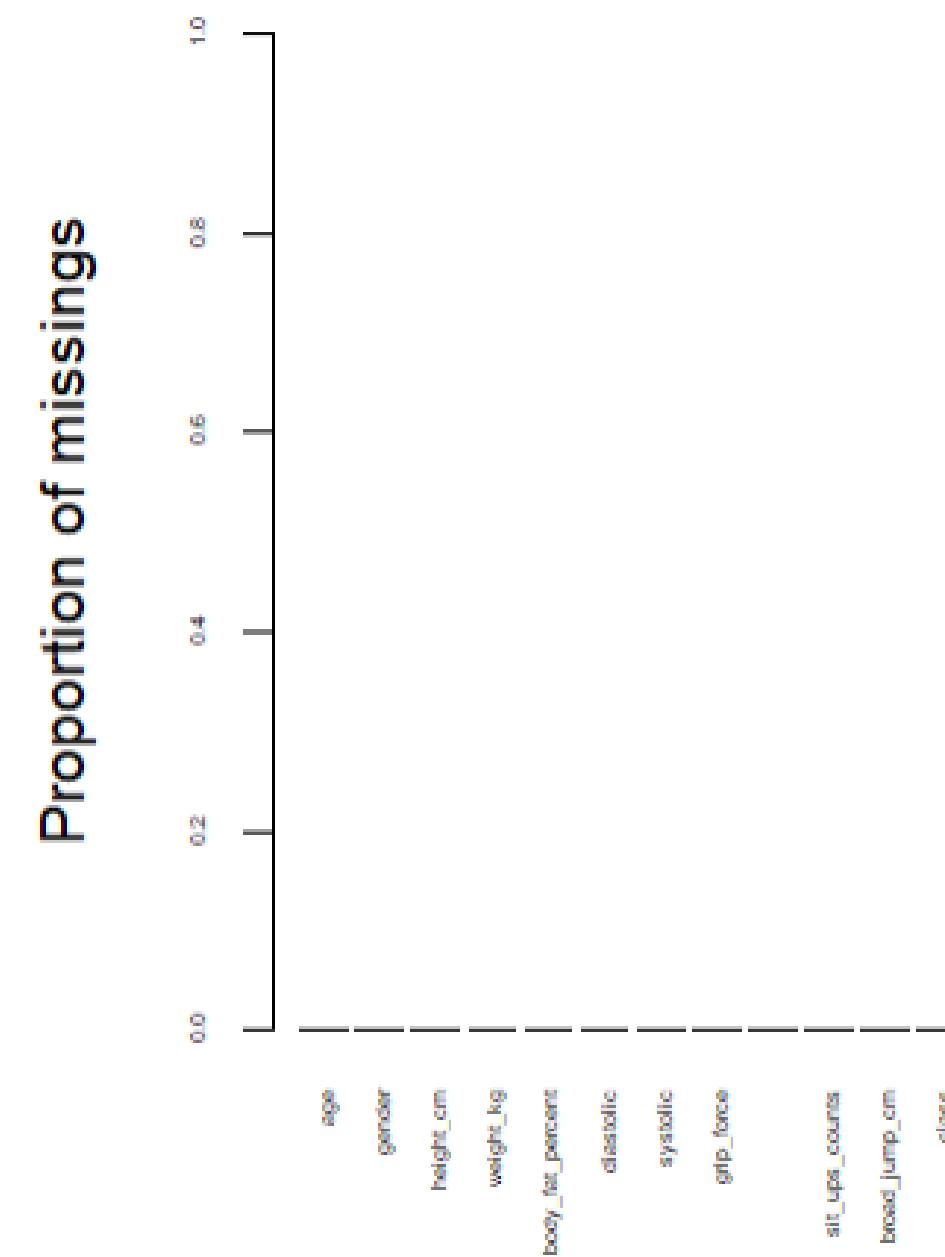
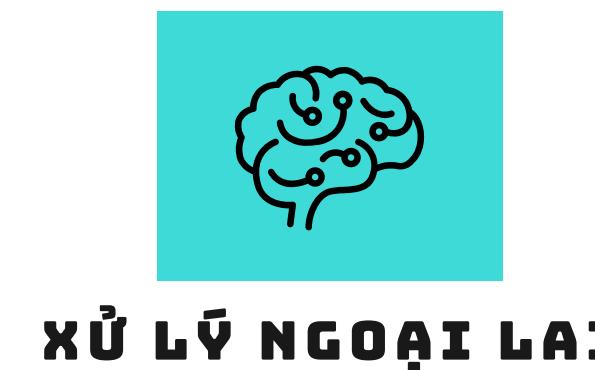
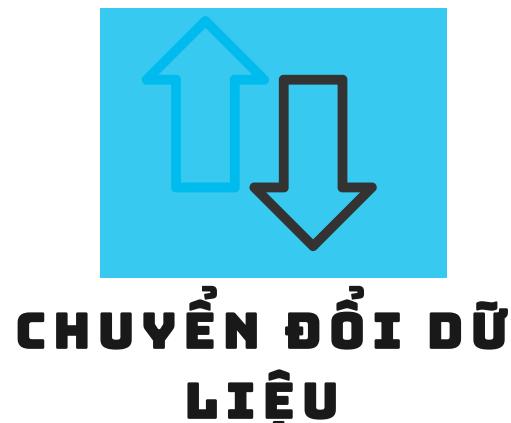


BMI: BODY MASS INDEX
(chỉ số khối cơ thể)

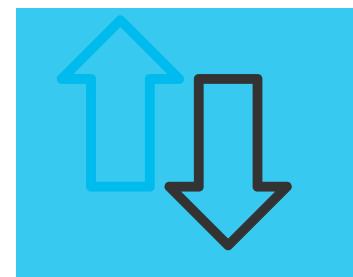
$$BMI = \frac{\text{CÂN NẶNG(KG)}}{[\text{CHIỀU CAO(M)}]^2}$$

MAP = diastolic + $\frac{1}{3}$ (systolic - diastolic)

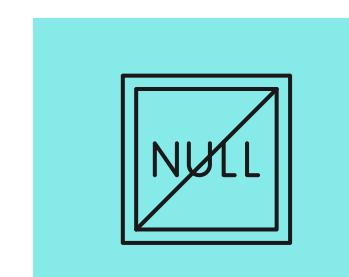
3. Tiên xử lý dữ liệu



3. Tiên xử lý dữ liệu



CHUYỂN ĐỔI DỮ LIỆU



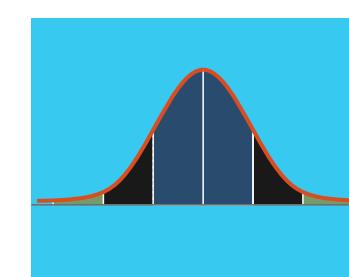
XỬ LÝ DỮ LIỆU KHUYẾT



XỬ LÝ NGOẠI LỆ



XỬ LÝ DỮ LIỆU BẤT THƯỜNG



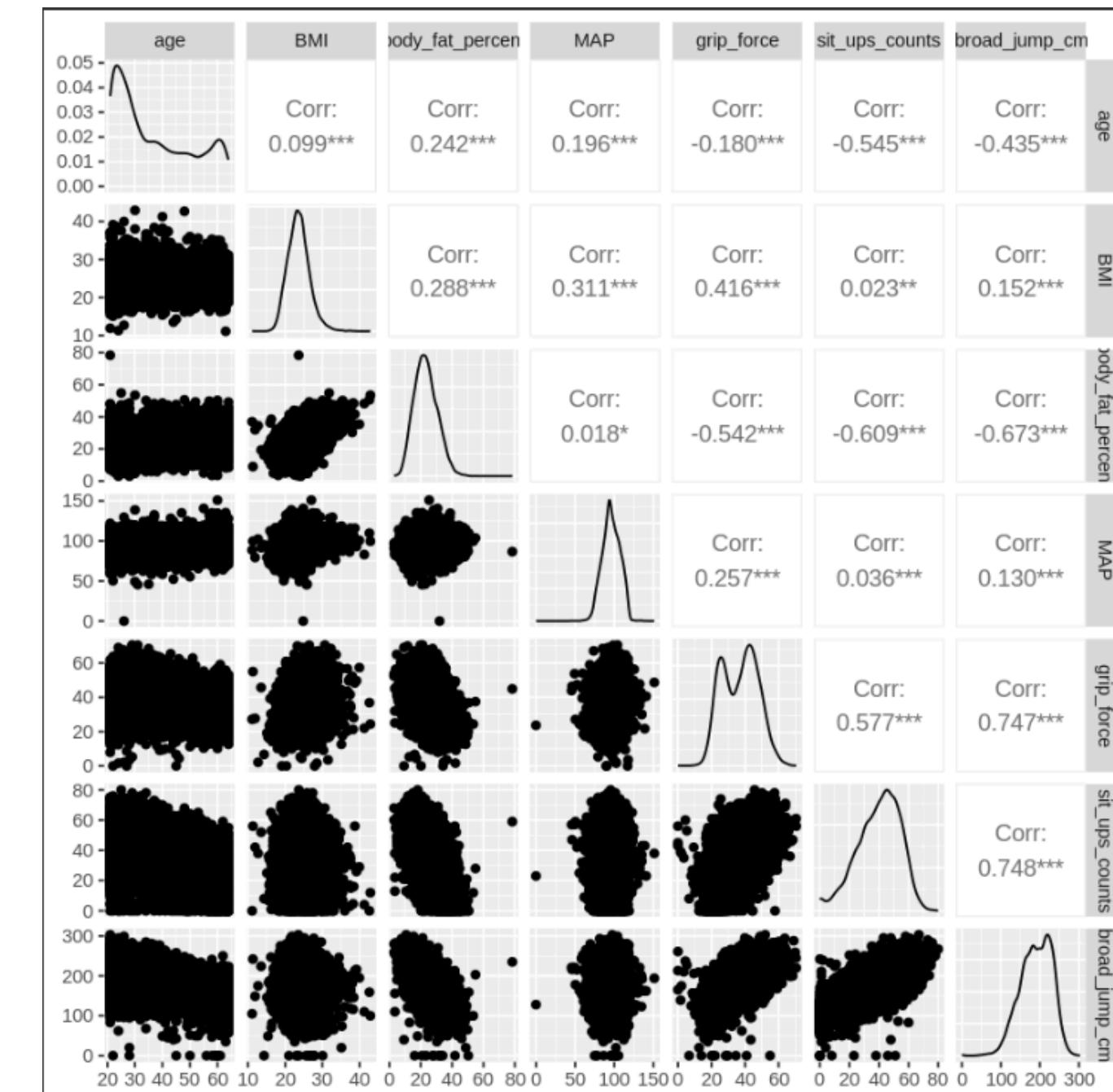
QUAN SÁT PHÂN PHỐI

Dữ liệu ngoại lệ:

Một số giá trị tối thiểu, ví dụ:

diastolic = 0,

sit_and_bend_forward_cm = - 25,
có thể là dữ liệu bất thường.



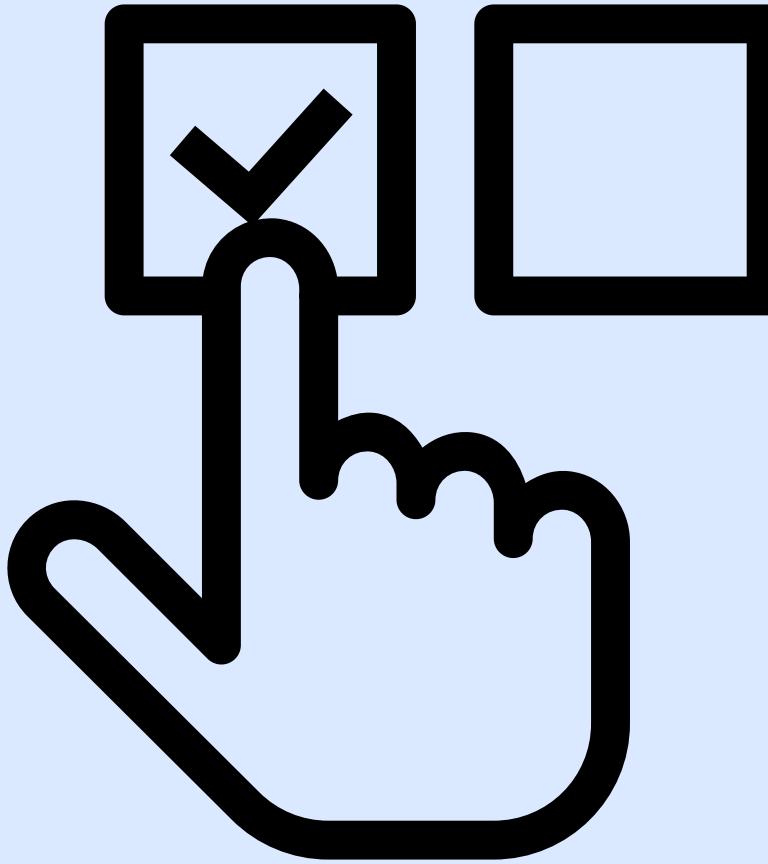
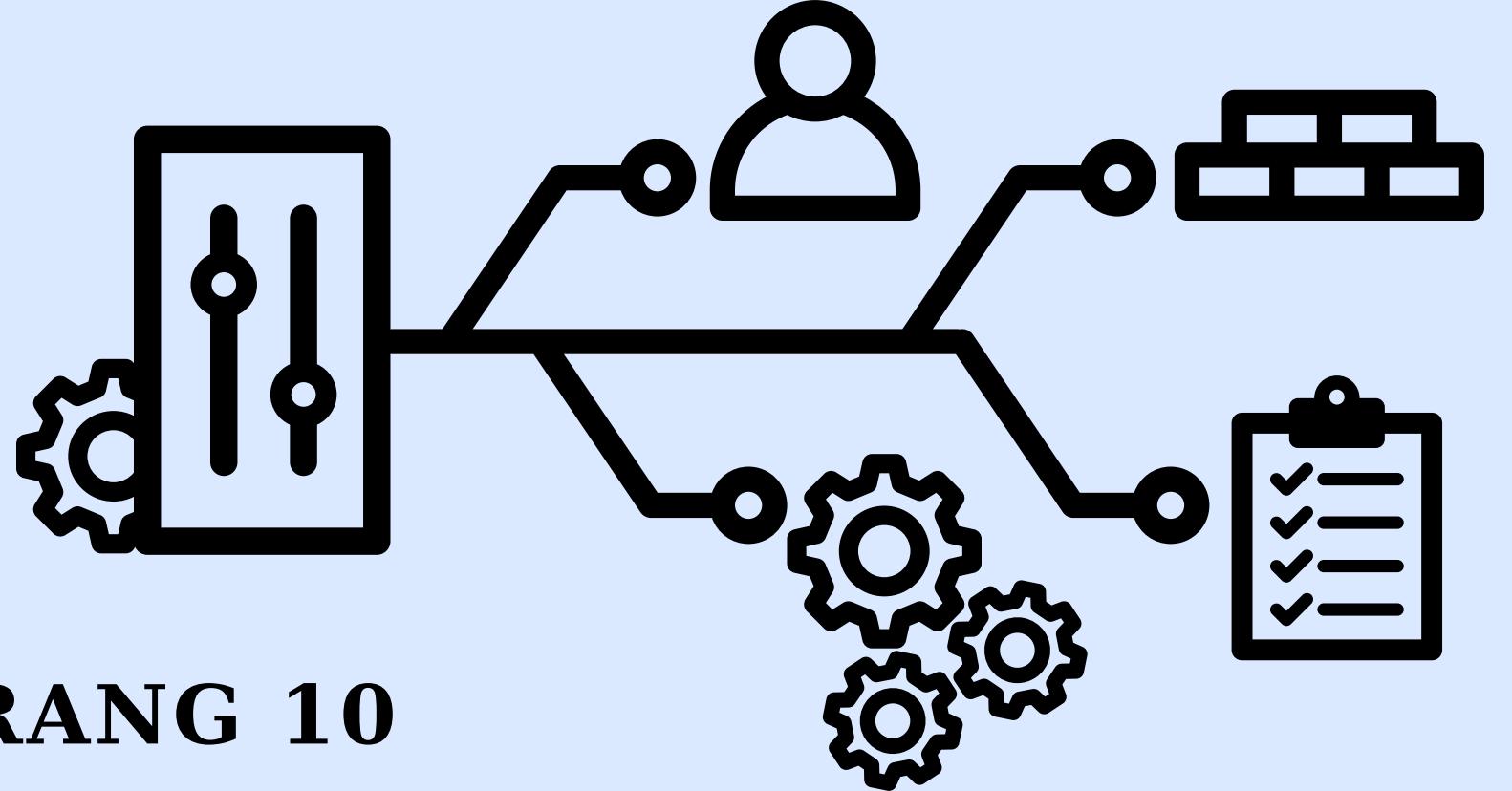
III. QUY TRÌNH XỬ LÝ

1. ĐÁNH GIÁ SƠ BỘ

2. A/B TESTING

3. LỰA CHỌN MÔ HÌNH

4. ĐÁNH GIÁ MÔ HÌNH



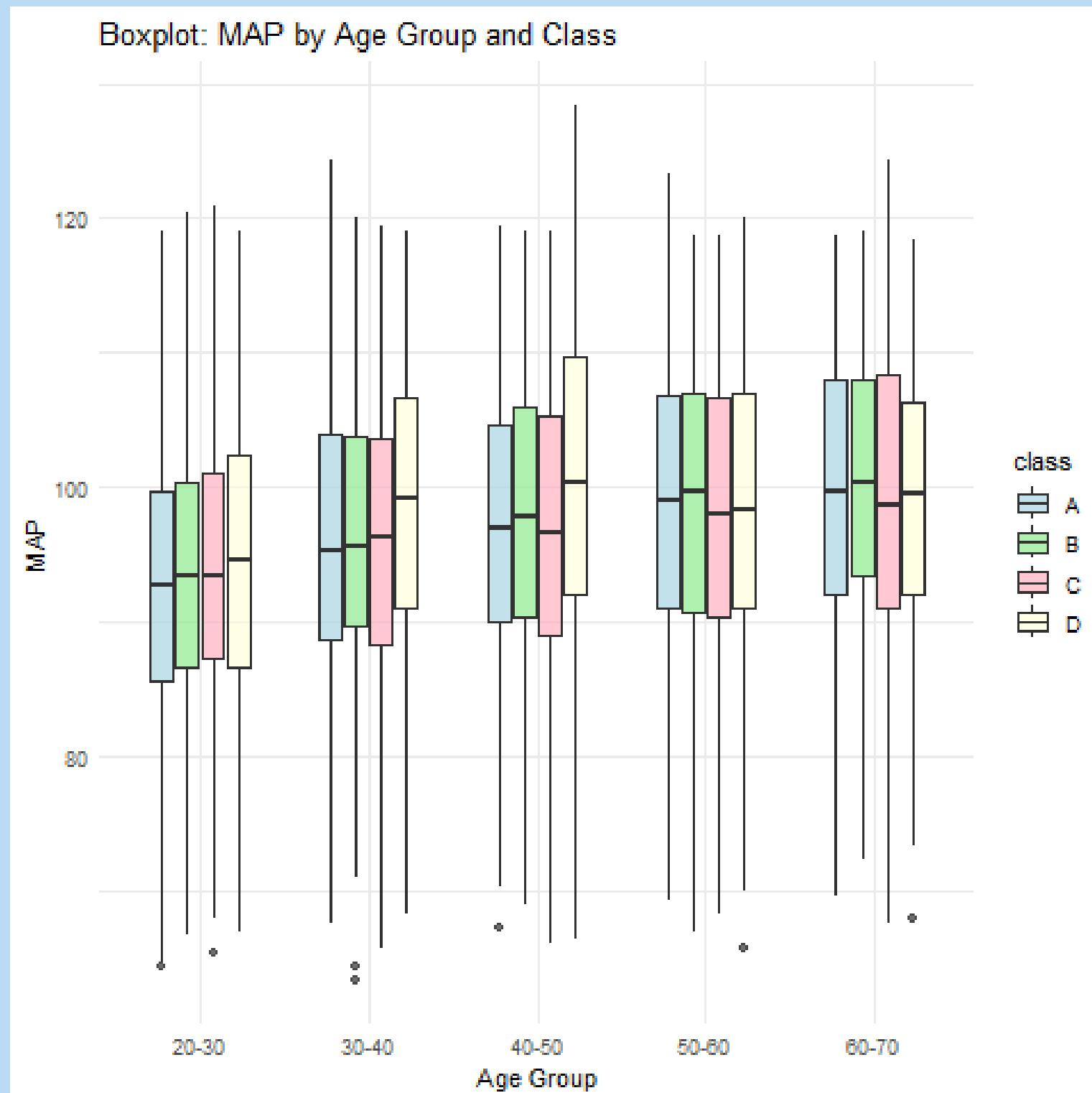
1. Đánh giá sơ bộ

body_fat_percent	diastolic	systolic	grip_force	
Min. : 4.50	Min. : 49.00	Min. : 89.0	Min. : 2.10	
1st Qu.:17.20	1st Qu.: 71.00	1st Qu.:120.0	1st Qu.:28.00	
Median :21.80	Median : 79.00	Median :130.0	Median :38.30	
Mean :22.17	Mean : 78.51	Mean :129.9	Mean :37.59	
3rd Qu.:26.80	3rd Qu.: 86.00	3rd Qu.:140.0	3rd Qu.:46.00	
Max. :42.60	Max. :107.00	Max. :167.0	Max. :70.50	
sit_and_bend_forward_cm	sit_ups_counts	broad_jump_cm	class	age_group
Min. :-3.70	Min. : 1.00	Min. : 77.0	0:3318	[21,32):3382
1st Qu.:14.10	1st Qu.:34.00	1st Qu.:168.0	1:3318	[32,44):1388
Median :18.60	Median :44.00	Median :197.0	2: 0	[44,56): 922
Mean :17.77	Mean :42.96	Mean :195.4	3: 0	[56,65): 944
3rd Qu.:22.30	3rd Qu.:53.00	3rd Qu.:226.0		
Max. :35.20	Max. :80.00	Max. :303.0		
body_fat_group	height_group	BMI_status	MAP	
<15% : 995	<160 cm :1128	0: 184	Min. : 63.33	
15–25%:3440	160–170 cm:2500	1:4881	1st Qu.: 88.00	
>25% :2201	>170 cm :3008	2:1479	Median : 95.33	
		3: 92	Mean : 95.65	
			3rd Qu.:103.67	
			Max. :124.33	

BMI

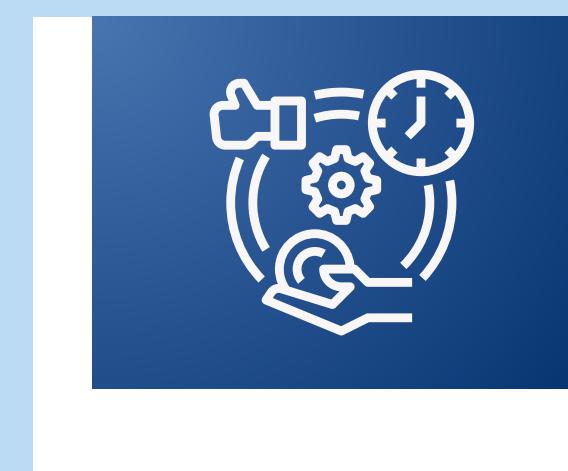
Min. :11.31
1st Qu.:21.56
Median :23.39
Mean :23.46
3rd Qu.:25.20
Max. :35.40

1. Đánh giá sơ bộ



TÁC ĐỘNG TỚI HUYẾT ÁP

- Nhóm A (màu xanh nhạt) duy trì huyết áp ổn định hơn các nhóm khác.
- Khoảng biến thiên của MAP trong nhóm A hẹp hơn, cho thấy sự ổn định trong kiểm soát huyết áp.



HIỆU QUẢ TẬP LUYỆN THEO TUỔI

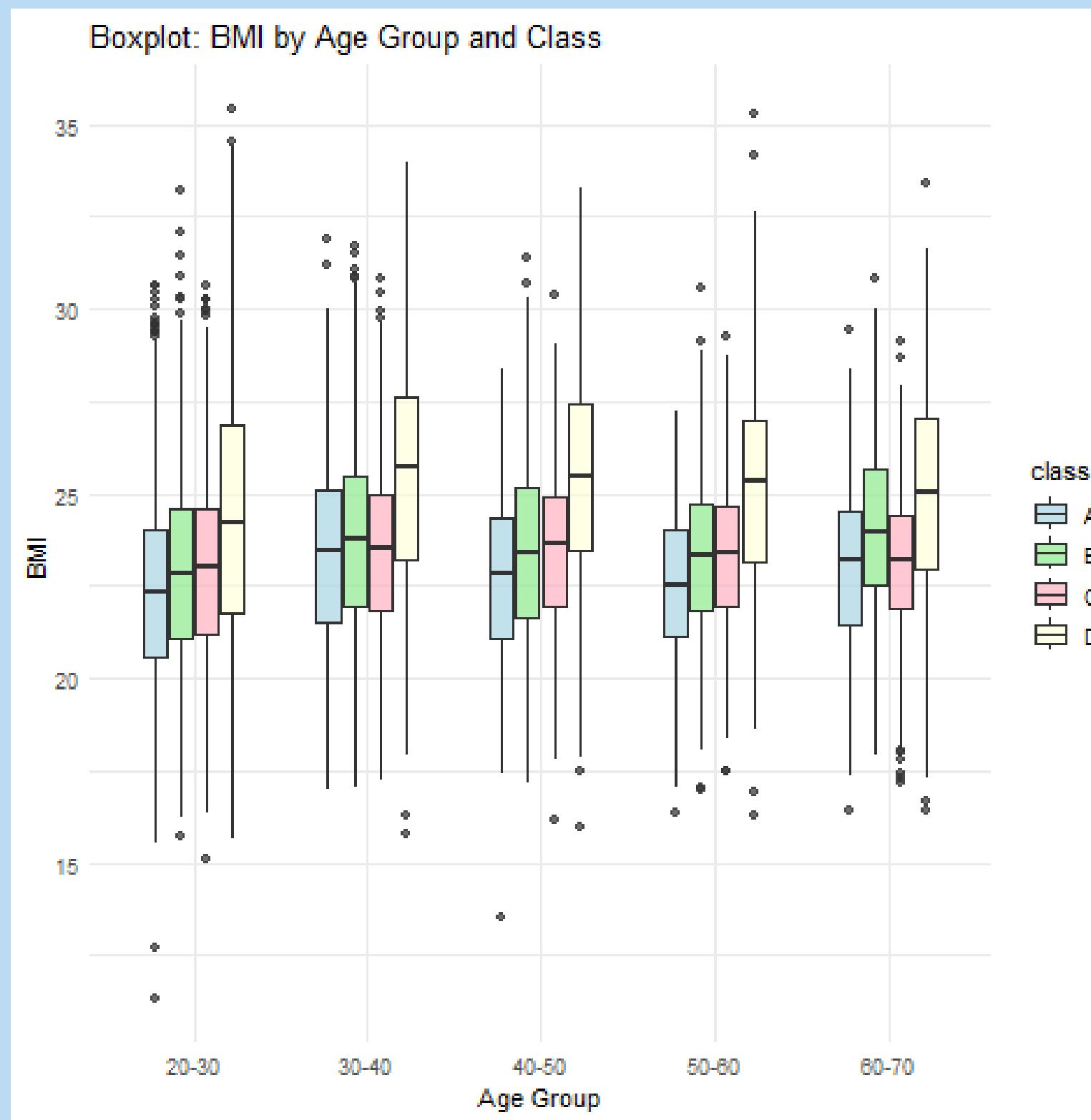
- Dù tuổi tác tăng lên, người thuộc nhóm tập luyện thường xuyên (nhóm A) vẫn duy trì được huyết áp ổn định.
- Sự chênh lệch về MAP giữa nhóm A và nhóm D không quá lớn, nhưng nhóm A thường có độ ổn định cao hơn.



NHẬN XÉT CHUNG

- Kết quả từ biểu đồ cho thấy tập luyện đều đặn giúp kiểm soát huyết áp ở mọi lứa tuổi.
- Ngay cả ở nhóm tuổi cao (60-70), tập luyện vẫn mang lại hiệu quả ổn định huyết áp.

1. Đánh giá sơ bộ



TÁC ĐỘNG TỚI BMI

- Trong cùng 1 nhóm tuổi, ở nhóm A duy trì BMI tốt hơn so với người trẻ tuổi thuộc nhóm C và D.
- Điều này chứng minh tập luyện có thể giúp kiểm soát cân nặng hiệu quả ở mọi lứa tuổi.



HIỆU QUẢ TẬP LUYỆN THEO TUỔI

- Khi độ tuổi càng tăng thì BMI ở nhóm A vẫn giữ mức ổn định và thấp hơn so với nhóm B, C, D.
- Ngay cả ở nhóm tuổi cao (60-70), vẫn có thể duy trì BMI khỏe mạnh thông qua tập luyện đều đặn

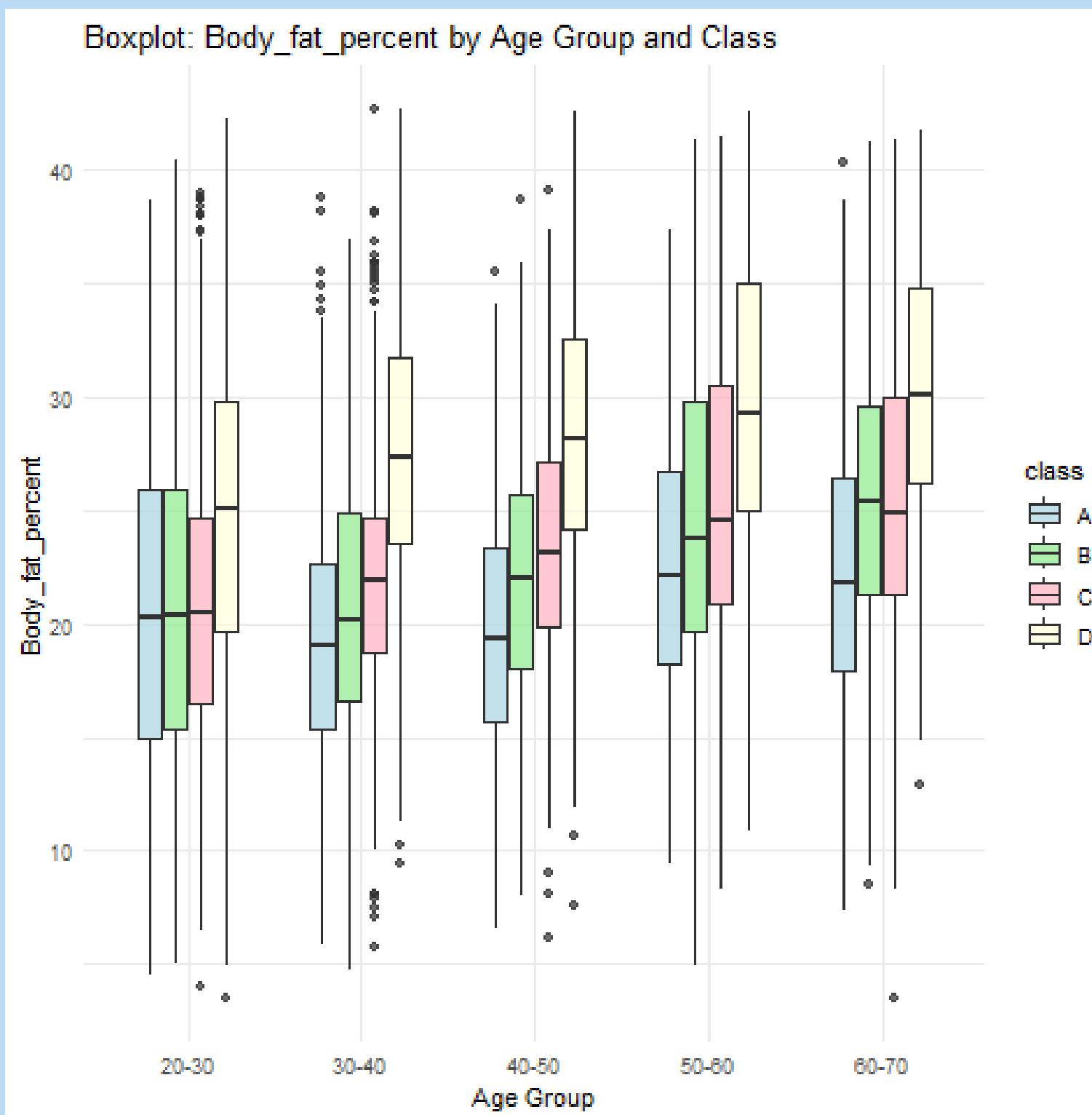


NHẬN XÉT CHUNG

- Nhóm A có xu hướng duy trì BMI ổn định theo thời gian.
- Điều này cho thấy tập luyện không chỉ giúp kiểm soát cân nặng mà còn hỗ trợ duy trì sức khỏe lâu dài.

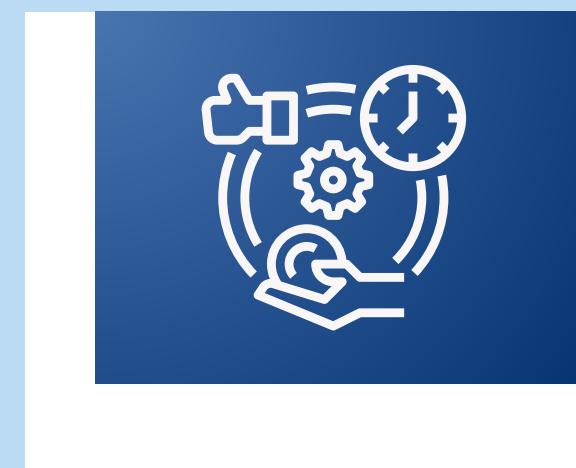


1. Đánh giá sơ bộ



TÁC ĐỘNG TỚI BODY_FAT_PERCENT

- Sự khác biệt lớn giữa nhóm A và nhóm D cho thấy tập luyện có tác động đáng kể.
- Ngay cả ở nhóm tuổi cao (60-70), nhóm A vẫn duy trì tỷ lệ mỡ cơ thể tốt hơn so với người trẻ tuổi hơn thuộc nhóm C và D.



HIỆU QUẢ TẬP LUYỆN THEO TUỔI

Theo độ tuổi tăng dần thì nhóm A vẫn giữ được tỉ lệ mỡ cơ thể ở mức bình thường.



NHẬN XÉT CHUNG

- Xu hướng tăng tỷ lệ mỡ theo tuổi ít rõ rệt hơn ở nhóm A.
- Điều này cho thấy tập luyện thường xuyên không chỉ giúp kiểm soát cân nặng mà còn làm chậm quá trình lão hóa tự nhiên của cơ thể.

2.A/B TESTING

Đặt giả thuyết:

- » H₀: Trung bình chỉ số 'BMI', 'MAP', 'broad_jump_cm' là như nhau.
- » H₁: Có ít nhất 1 trung bình trong các class khác với những cái còn lại.

a)

BMI,MAP,broad_jump_cm

Những biến BMI, MAP và broad_jump_cm gần với phân phối chuẩn hơn những biến còn lại.

KIỂM ĐỊNH ANOVA CHO TỪNG BIẾN TUÂN THEO PHÂN PHỐI CHUẨN.

Variable	BMI	MAP	broad_jump_cm
p-value	5.564086e-234	1.011128e-12	8.210931e-152

Bảng 1: Data table showing BMI, MAP, and broad jump p-values.

Nhận xét: Cả 3 chỉ số đều có giá trị p-value < 0.05.

Sự thay đổi trong biến phụ thuộc có thể được giải thích (ít nhất một phần) bởi 3 biến độc lập trên.

2. A/B TESTING

► PERMUTATION ANOVA

- phân phối dữ liệu không cần chuẩn.
- giá trị p_value trên toàn bộ nhóm
- tốn tài nguyên tính toán.

► DUNN TEST

- phân phối dữ liệu không cần chuẩn
- cho thấy được sự so sánh giữa từng nhóm.
- tiết kiệm tài nguyên tính toán nếu số lượng nhóm nhỏ.

```
```{r}
out_aov_1 <- aovp(formula = age ~ class, data = data_cleaned, perm = "Prob")
summary(out_aov_1)
```

[1] "Settings: unique SS"
Component 1 :
      Df R Sum Sq R Mean Sq Iter Pr(Prob)
class       3    9190   3063.43 5000 < 2.2e-16 ***
Residuals 12615  2295329    181.95
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

2.A/B TESTING

Đặt giả thuyết:

- ▶ H₀: Phân phối của 'body_fat_percent' trong 4 nhóm giống nhau.
- ▶ H₁: Phân phối của 'body_fat_percent' trong 4 nhóm có sự khác biệt.

b)

Body_fat_percent

Giả thuyết kiểm định cho từng cặp nhóm không theo phân phối chuẩn (Dunn Test).

| Comparison | Z | P.unadj | P.adj |
|------------|------------|---------------|---------------|
| A - B | -8.441419 | 3.135075e-17 | 1.881045e-16 |
| A - C | -12.093004 | 1.150034e-33 | 6.900204e-33 |
| B - C | -3.645395 | 2.669819e-04 | 1.601892e-03 |
| A - D | -35.456243 | 2.324352e-275 | 1.394611e-274 |
| B - D | -27.466581 | 4.403917e-166 | 2.642350e-165 |
| C - D | -24.034384 | 1.215925e-127 | 7.295547e-127 |

Bảng 3: Results of comparison between groups.

Kết luận: Ta thấy các giá trị p_value của từng cặp đều nhỏ hơn 0.05. Do đó, ta bác bỏ giả thuyết H₀ => Phân phối của 'body_fat_percent' trong 4 nhóm có sự khác biệt.

2.A/B TESTING

Đặt giả thuyết:

- ▶ H₀: Phân phối của 'age' trong 4 nhóm là giống nhau
- ▶ H₁: Phân phối của 'age' trong 4 nhóm có sự khác biệt

c)

Age

Giả thuyết kiểm định cho từng cặp nhóm không theo phân phối chuẩn (Dunn Test).

| Comparison | Z | P.unadj | P.adj |
|------------|------------|--------------|--------------|
| A - B | -4.7071460 | 2.512090e-06 | 1.507254e-05 |
| A - C | -2.8102360 | 4.950519e-03 | 2.970311e-02 |
| B - C | 1.8991749 | 5.754149e-02 | 3.452489e-01 |
| A - D | -5.2446258 | 1.565998e-07 | 9.395989e-07 |
| B - D | -0.7933069 | 4.275990e-01 | 1.000000e+00 |
| C - D | -2.5895481 | 9.610200e-03 | 5.766120e-02 |

Bảng 2: Results of comparison between groups.

Kết luận: Ta thấy các giá trị p_value của từng cặp đều cho thấy giá trị của chúng < 0,05 => Bác bỏ giả thuyết H₀.

2.A/B TESTING

Đặt giả thuyết:

- » H₀: Phân phối của 'grip_force' trong 4 nhóm giống nhau
- » H₁: Phân phối của 'grip_force' trong 4 nhóm có sự khác biệt.

d)

Grip_force

Giả thuyết kiểm định cho từng cặp nhóm không theo phân phối chuẩn (Dunn Test).

| Comparison | Z | P.unadj | P.adj |
|------------|-----------|--------------|--------------|
| A - B | 2.543697 | 1.096861e-02 | 6.581166e-02 |
| A - C | 7.777510 | 7.396564e-15 | 4.437938e-14 |
| B - C | 5.230700 | 1.688690e-07 | 1.013214e-06 |
| A - D | 12.735905 | 3.734896e-37 | 2.240937e-36 |
| B - D | 10.327785 | 5.276471e-25 | 3.165883e-24 |
| C - D | 5.387326 | 7.151348e-08 | 4.290809e-07 |

Bảng 4: Results of comparison between groups.

Kết luận: Ta thấy các giá trị p_value của từng cặp đều nhỏ hơn 0.05. Do đó, ta bác bỏ giả thuyết H₀ => Phân phối của 'Grip_force' trong 4 nhóm có sự khác biệt.

2.A/B TESTING

Đặt giả thuyết:

- » H₀: Phân phối của 'sit_and_bend_forward_cm' trong 4 nhóm giống nhau
- » H₁: Phân phối của 'sit_and_bend_forward_cm' trong 4 nhóm có sự khác biệt.

e)

Sit_and_bend_forward_cm

Giả thuyết kiểm định cho từng cặp nhóm không theo phân phối chuẩn (Dunn Test).

| Comparison | Z | P.unadj | P.adj |
|------------|----------|---------------|---------------|
| A - B | 26.66418 | 1.225592e-156 | 7.353552e-156 |
| A - C | 45.16472 | 0.000000e+00 | 0.000000e+00 |
| B - C | 18.47888 | 3.054746e-76 | 1.832847e-75 |
| A - D | 62.05603 | 0.000000e+00 | 0.000000e+00 |
| B - D | 36.83229 | 5.617808e-297 | 3.370685e-296 |
| C - D | 19.37896 | 1.161693e-83 | 6.970159e-83 |

Bảng 5: Results of comparison between groups.

Kết luận: Ta thấy các giá trị p_value của từng cặp đều nhỏ hơn 0.05. Do đó, ta bác bỏ giả thuyết H₀ => Phân phối của 'Sit_and_bend_forward_cm' trong 4 nhóm có sự khác biệt.

2.A/B TESTING

Đặt giả thuyết:

- » H₀: Phân phối của 'sit_ups_counts' trong 4 nhóm giống nhau
- » H₁: Phân phối của 'sit_ups_counts' trong 4 nhóm có sự khác biệt.

f)

Sit_ups_counts

Giả thuyết kiểm định cho từng cặp nhóm không theo phân phối chuẩn (Dunn Test).

| Comparison | Z | P.unadj | P.adj |
|------------|----------|---------------|---------------|
| A - B | 16.15672 | 1.018371e-58 | 6.110224e-58 |
| A - C | 27.75396 | 1.560943e-169 | 9.365660e-169 |
| B - C | 11.58400 | 4.967303e-31 | 2.980382e-30 |
| A - D | 44.93543 | 0.000000e+00 | 0.000000e+00 |
| B - D | 29.64955 | 3.437238e-193 | 2.062343e-192 |
| C - D | 18.71197 | 3.954474e-78 | 2.372684e-77 |

Bảng 6: Results of comparison between groups.

Kết luận: Ta thấy các giá trị p_value của từng cặp đều nhỏ hơn 0.05. Do đó, ta bác bỏ giả thuyết H₀ => Phân phối của 'Sit_ups_counts' trong 4 nhóm có sự khác biệt.

2.A/B TESTING

Đặt giả thuyết:

- » H₀: Phân phối của 'gender' trong 4 nhóm giống nhau
- » H₁: Phân phối của 'gender' trong 4 nhóm có sự khác biệt.

e)

gender

Giả thuyết kiểm định cho từng cặp nhóm không theo phân phối chuẩn (Chi-square).

Pearson's Chi-squared test

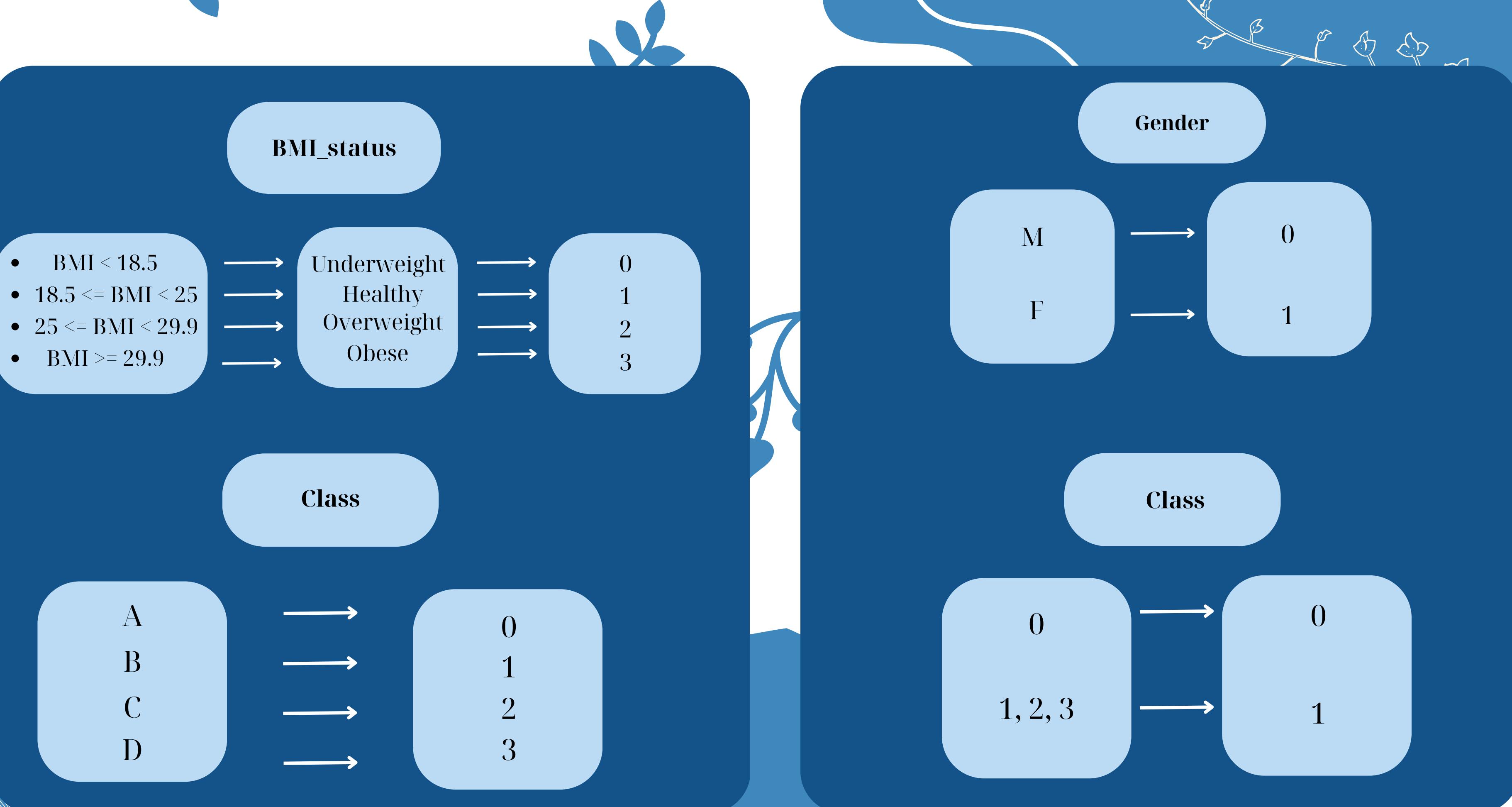
```
data: table_data  
X-squared = 121.2, df = 3, p-value < 2.2e-16
```

Kết luận: Ta thấy các giá trị p_value của từng cặp đều nhỏ hơn 0.05. Do đó, ta bác bỏ giả thuyết H₀ => Phân phối của 'gender' trong 4 nhóm có sự khác biệt.

3. LỰA CHỌN MÔ HÌNH

LOGISTIC REGRESSION

RANDOM FOREST



Logistic Regression

- Chia dữ liệu thành tập huấn luyện (80%) và tập kiểm tra (20%) bằng cách sử dụng hàm CreateDataPartition của thư viện caret.
- Huấn luyện mô hình logistic regression.
- Dự đoán trên tập kiểm tra: tiến hành dự đoán trên tập kiểm tra.
- Đánh giá kết quả dự đoán: sử dụng ma trận nhầm lẫn để đánh giá hiệu suất của mô hình.
- Đánh giá tầm quan trọng của biến: Vẽ biểu đồ để kiểm tra mức quan trọng của mô hình.

Logistic Regression

- Độ chính xác tổng thể
 - Accuracy: 78,88%, KTC 95% CI: (76.59%, 81.05%).
 - NIR: 50%.
 - Kappa=0.5777.
 - F1-score: 0.7947(class 0), 0.7867(class 1)
- Ma trận nhầm lẫn

| | | Reference | |
|------------|---|------------|----------|
| | | Prediction | 0 1 |
| Prediction | 0 | 556 | 157 |
| | 1 | 107 | 506 |

Random forest

- Chia tập dữ liệu: Chia dữ liệu thành tập huấn luyện (80%) và tập kiểm tra (20%) bằng cách sử dụng hàm initial_split.
- Huấn luyện mô hình Random Forest: Mô hình Random Forest được huấn luyện với 500 cây quyết định.
- Dự đoán trên tập kiểm tra: Tiến hành dự đoán trên tập kiểm tra.
- Đánh giá kết quả dự đoán: sử dụng ma trận nhầm lẫn để đánh giá hiệu suất của mô hình.
- Đánh giá tầm quan trọng của biến: Vẽ biểu đồ để kiểm tra mức quan trọng của mô hình.

Random forest

- Độ chính xác tổng thể
 - Accuracy: 82,53%, KTC 95% CI: (80.38%, 84.54%).
 - NIR: 50%.
 - Kappa=0.6506.
 - F1-score: 0.837(class 0), 0.812(class 1)
- Ma trận nhầm lẫn

| | | Reference | |
|------------|---|-----------|-----|
| Prediction | | 0 | 1 |
| 0 | 0 | 579 | 163 |
| | 1 | 85 | 501 |

Đánh giá mô hình

- Mô hình random forest
 - Mô hình đạt hiệu suất cao với các chỉ số chính Balanced Accuracy, F1-Score đều trên 80%.
- Mô hình logistic regression
 - Mô hình Logistic Regression đạt Balanced Accuracy là 78.88% cho cả hai lớp, một kết quả khá tốt. Điều này cho thấy mô hình có khả năng phân biệt tương đối tốt giữa hai lớp trong tập dữ liệu.
 - Độ nhạy của lớp 1 thấp hơn lớp 0, có thể gây ra bỏ sót các trường hợp quan trọng, Lớp 1 có Specificity cao hơn, nghĩa là mô hình ít bị nhầm lẫn hơn khi dự đoán không thuộc lớp 1.

→ Sử dụng Random Forest là công cụ chính

IV. KẾT QUẢ TỔNG KẾT

- Giúp các chuyên gia sức khoẻ biết được hiệu quả của việc tập thể dục.
- Nhóm hiệu suất D thường có BMI, MAP cao hơn, cho thấy nguy cơ sức khỏe như béo phì và các bệnh liên quan đến tim mạch.
- Nam giới vượt trội về sức mạnh cơ bắp, trong khi nữ giới có lợi thế về sự linh hoạt và độ dẻo dai.
- Hiệu suất tập luyện có xu hướng giảm dần theo độ tuổi, đặc biệt ở các chỉ số như lực nắm tay và khả năng nhảy xa.

VĨ HƯỚNG PHÁT TRIỂN

- Cải thiện hiệu suất tập luyện:
 - Nhóm chúng tôi khuyến khích các nhóm có hiệu suất thấp thực hiện các chương trình tập luyện cá nhân hóa, tập trung vào kiểm soát cân nặng và giảm mỡ cơ thể. Đồng thời, tăng cường giáo dục về dinh dưỡng và sức khỏe tim mạch, đặc biệt cho các nhóm tuổi trung niên và cao tuổi.

- Ứng dụng mô hình dự đoán:
 - Sử dụng Random Forest như công cụ chính để đánh giá hiệu suất tập luyện và dự đoán nguy cơ sức khỏe.
 - Tối ưu hóa mô hình bằng cách thêm các biến bổ sung như thói quen ăn uống, thời gian tập luyện, và lịch sử y tế.

- Phát triển chính sách hỗ trợ:
 - Đề xuất các chương trình khuyến khích hoạt động thể chất ở cộng đồng, đặc biệt tập trung vào nhóm tuổi lớn hơn và những người ít vận động.
 - Tăng cường hợp tác giữa các chuyên gia sức khỏe và tổ chức thể thao để xây dựng các chương trình tập luyện tối ưu.

Thank You

Nhóm 22 xin cảm ơn thầy và các bạn đã lắng nghe

2.A/B TESTING

Những biến BMI, MAP và broad_jump_cm gần với phân phối chuẩn hơn những biến còn lại.

Kiểm định ANOVA cho từng biến tuân theo phân phối chuẩn.

Đặt giả thuyết:

- H₀: Trung bình chỉ số 'BMI', 'MAP', 'broad_jump_cm' là như nhau.
- H₁: Có ít nhất 1 trung bình trong các class khác với những cái còn lại.

| Variable | BMI | MAP | broad_jump_cm |
|----------|---------------|--------------|---------------|
| p-value | 1.404499e-309 | 6.126191e-13 | 5.649599e-215 |

Bảng 5: Data table showing BMI, MAP, and broad jump p-values.

Nhận xét: Cả 3 chỉ số đều có giá trị p-value < 0.05.

Sự thay đổi trong biến phụ thuộc có thể được giải thích (ít nhất một phần) bởi 3 biến độc lập trên và đều là ngẫu nhiên.

3. Thống kê mô tả

Các vấn đề chú ý:

- **Dữ liệu ngoại lệ:**
 - Một số giá trị tối thiểu, ví dụ: diastolic = 0, sit_and_bend_forward_cm = - 25, có thể là dữ liệu bất thường.[1].
- **Cân bằng phân lớp:**
 - Cần kiểm tra xem các nhóm class (A, B, C, D) có được phân phối đồng đều không, vì điều này ảnh hưởng đến các phân tích tiếp theo.

Giả thuyết kiểm định cho từng cặp nhóm không theo phân phối chuẩn (Dunn Test).

a/ Age

Đặt giả thuyết:

- H0: Phân phối của 'age' trong 4 nhóm là giống nhau
- H1: Phân phối của 'age' trong 4 nhóm có sự khác biệt

| Comparison | Z | P.unadj | P.adj |
|------------|------------|---------------|---------------|
| A - B | -8.576304 | 9.797079e-18 | 5.878247e-17 |
| A - C | -12.036686 | 2.279297e-33 | 1.367578e-32 |
| B - C | -3.458843 | 5.425009e-04 | 3.255005e-03 |
| A - D | -38.958483 | 0.000000e+00 | 0.000000e+00 |
| B - D | -30.378629 | 1.052355e-202 | 6.314133e-202 |
| C - D | -26.923807 | 1.156226e-159 | 6.937355e-159 |

Bảng 6: Results of comparison between groups.

Kết luận: Ta thấy các giá trị p_value của từng cặp đều cho thấy giá trị của chúng $< 0,05 \Rightarrow$ Bác bỏ giả thuyết H0.

Giả thuyết kiểm định cho từng cặp nhóm không theo phân phối chuẩn (Dunn Test).

b/Body_fat_percent

Đặt giả thuyết:

- H0: Phân phối của 'body_fat_percent' trong 4 nhóm giống nhau.
- H1: Phân phối của 'body_fat_percent' trong 4 nhóm có sự khác biệt.

| Comparison | Z | P.unadj | P.adj |
|------------|------------|-----------------------------|-----------------------------|
| A - B | -8.576304 | 9.797079×10^{-18} | 5.878247×10^{-17} |
| A - C | -12.036686 | 2.279297×10^{-33} | 1.367578×10^{-32} |
| B - C | -3.458843 | 5.425009×10^{-4} | 3.255005×10^{-3} |
| A - D | -38.958483 | 0.000000×10^0 | 0.000000×10^0 |
| B - D | -30.378629 | $1.052355 \times 10^{-202}$ | $6.314133 \times 10^{-202}$ |
| C - D | -26.923807 | $1.156226 \times 10^{-159}$ | $6.937355 \times 10^{-159}$ |

Bảng 7: Kết quả so sánh giữa các nhóm.

- Ta thấy các giá trị p_value của từng cặp đều nhỏ hơn 0.05. Do đó, ta bác bỏ giả thuyết H0 => Phân phối của 'body_fat_percent' trong 4 nhóm có sự khác biệt.

Giả thuyết kiểm định cho từng cặp nhóm không theo phân phối chuẩn (Dunn Test).

c/Grip_force

Đặt giả thuyết:

- H0: Phân phối của ‘grip_force’ trong 4 nhóm giống nhau
- H1: Phân phối của ‘grip_force’ trong 4 nhóm có sự khác biệt.

| Comparison | Z | P.unadj | P.adj |
|------------|-----------|--------------|--------------|
| 1 (A - B) | 2.665996 | 7.676067e-03 | 4.605640e-02 |
| 2 (A - C) | 7.903816 | 2.704917e-15 | 1.622950e-14 |
| 3 (B - C) | 5.237031 | 1.631802e-07 | 9.790812e-07 |
| 4 (A - D) | 14.726286 | 4.370735e-49 | 2.622441e-48 |
| 5 (B - D) | 12.058992 | 1.738992e-33 | 1.043395e-32 |
| 6 (C - D) | 6.822979 | 8.917141e-12 | 5.350285e-11 |

Bảng 8: Comparison of groups with Z-scores and p-values.

- Ta thấy các giá trị p_value của từng cặp đều nhỏ hơn 0.05. Do đó, ta bác bỏ giả thuyết H0 => Phân phối của ‘Grip_force’ trong 4 nhóm có sự khác biệt.

Giả thuyết kiểm định cho từng cặp nhóm không theo phân phối chuẩn (Dunn Test).

d/ Sit_and_bend_forward_cm

Đặt giả thuyết:

- H0: Phân phối của 'sit_and_bend_forward_cm' trong 4 nhóm giống nhau
- H1: Phân phối của 'sit_and_bend_forward_cm' trong 4 nhóm có sự khác biệt.

| Comparison | Z | P.unadj | P.adj |
|--------------------|-------|----------|---------------|
| 1
3.082636e-146 | A - B | 25.82111 | 5.137726e-147 |
| 2
0.000000e+00 | A - C | 43.80581 | 0.000000e+00 |
| 3
1.692270e-71 | B - C | 17.97950 | 2.820450e-72 |
| 4
0.000000e+00 | A - D | 67.89141 | 0.000000e+00 |
| 5
0.000000e+00 | B - D | 42.06330 | 0.000000e+00 |
| 6
2.032995e-127 | C - D | 24.08740 | 3.388326e-128 |

Bảng 9: Comparison Results

- Ta thấy các giá trị p_value của từng cặp đều nhỏ hơn 0.05. Do đó, ta bác bỏ giả thuyết H0 => Phân phối của 'Sit_and_bend_forward_cm' trong 4 nhóm có sự khác biệt.

Giả thuyết kiểm định cho từng cặp nhóm không theo phân phối chuẩn (Dunn Test).

e/Sit_ups_counts

Đặt giả thuyết:

- H₀: Phân phối của ‘sit_ups_counts’ trong 4 nhóm giống nhau
- H₁: Phân phối của ‘sit_ups_counts’ trong 4 nhóm có sự khác biệt.

| Comparison | Z | P.unadj | P.adj |
|----------------------------------|-------|----------|-----------------------------|
| 1
1.137007×10^{-56} | A - B | 15.97544 | 1.895012×10^{-57} |
| 2
$7.424813 \times 10^{-165}$ | A - C | 27.42899 | $1.237469 \times 10^{-165}$ |
| 3
1.405912×10^{-29} | B - C | 11.45031 | 2.343187×10^{-30} |
| 4
0.000000×10^0 | A - D | 49.35439 | 0.000000×10^0 |
| 5
$1.956725 \times 10^{-243}$ | B - D | 33.37407 | $3.261208 \times 10^{-244}$ |
| 6
$8.608204 \times 10^{-106}$ | C - D | 21.92704 | $1.434701 \times 10^{-106}$ |

Bảng 10: Kết quả so sánh giữa các nhóm.

- Ta thấy các giá trị p_value của từng cặp đều nhỏ hơn 0.05. Do đó, ta bác bỏ giả thuyết H₀ => Phân phối của ‘Sit_ups_counts’ trong 4 nhóm có sự khác biệt.

3. Lựa chọn mô hình

1. Random forest

- Chia tập dữ liệu: Chia dữ liệu thành tập huấn luyện (80%) và tập kiểm tra (20%) bằng cách sử dụng hàm initial_split.
- Huấn luyện mô hình Random Forest: Mô hình Random Forest được huấn luyện với 500 cây quyết định.
- Dự đoán trên tập kiểm tra: Tiến hành dự đoán trên tập kiểm tra.
- Đánh giá kết quả dự đoán: sử dụng ma trận nhầm lẫn để đánh giá hiệu suất của mô hình.
- Đánh giá tầm quan trọng của biến: Vẽ biểu đồ để kiểm tra mức quan trọng của mô hình.

3. Lựa chọn mô hình

- Độ chính xác tổng thể
 - Accuracy: 82,53%, KTC 95% CI: (80.38%, 84.54%).
 - NIR: 50%.
 - Kappa=0.6506.
 - F1-score: 0.837(class 0), 0.812(class 1)

Ma trận nhầm lẫn

| | | Reference | |
|------------|---|-----------|-----|
| | | 0 | 1 |
| Prediction | 0 | 579 | 163 |
| | 1 | 85 | 501 |

3. Lựa chọn mô hình

2. Logistic Regression

- Chia dữ liệu thành tập huấn luyện (80%) và tập kiểm tra (20%) bằng cách sử dụng hàm CreateDataPartition của thư viện caret.
- Huấn luyện mô hình logistic regression.
- Dự đoán trên tập kiểm tra: tiến hành dự đoán trên tập kiểm tra.
- Đánh giá kết quả dự đoán: sử dụng ma trận nhầm lẫn để đánh giá hiệu suất của mô hình.
- Đánh giá tầm quan trọng của biến: Vẽ biểu đồ để kiểm tra mức quan trọng của mô hình.

3. Lựa chọn mô hình

- Độ chính xác tổng thể
 - Accuracy: 78,88%, KTC 95% CI: (76.59%, 81.05%).
 - NIR: 50%.
 - Kappa=0.5777.
 - F1-score: 0.7947(class 0), 0.7867(class 1)

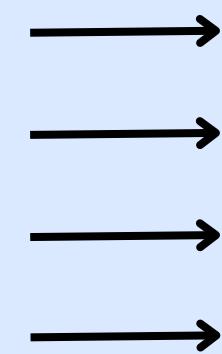
Ma trận nhầm lẫn

| | | Reference | |
|------------|---|-----------|-----|
| | | 0 | 1 |
| Prediction | 0 | 556 | 157 |
| | 1 | 107 | 506 |

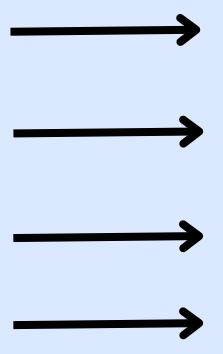
3. Lựa chọn mô hình

BMI_status

- $\text{BMI} < 18.5$
- $18.5 \leq \text{BMI} < 25$
- $25 \leq \text{BMI} < 29.9$
- $\text{BMI} \geq 29.9$



Underweight
Healthy
Overweight
Obese



0
1
2
3

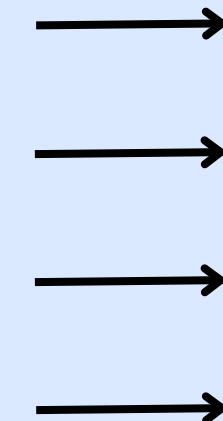
Class

A

B

C

D

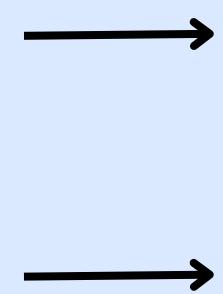


0
1
2
3

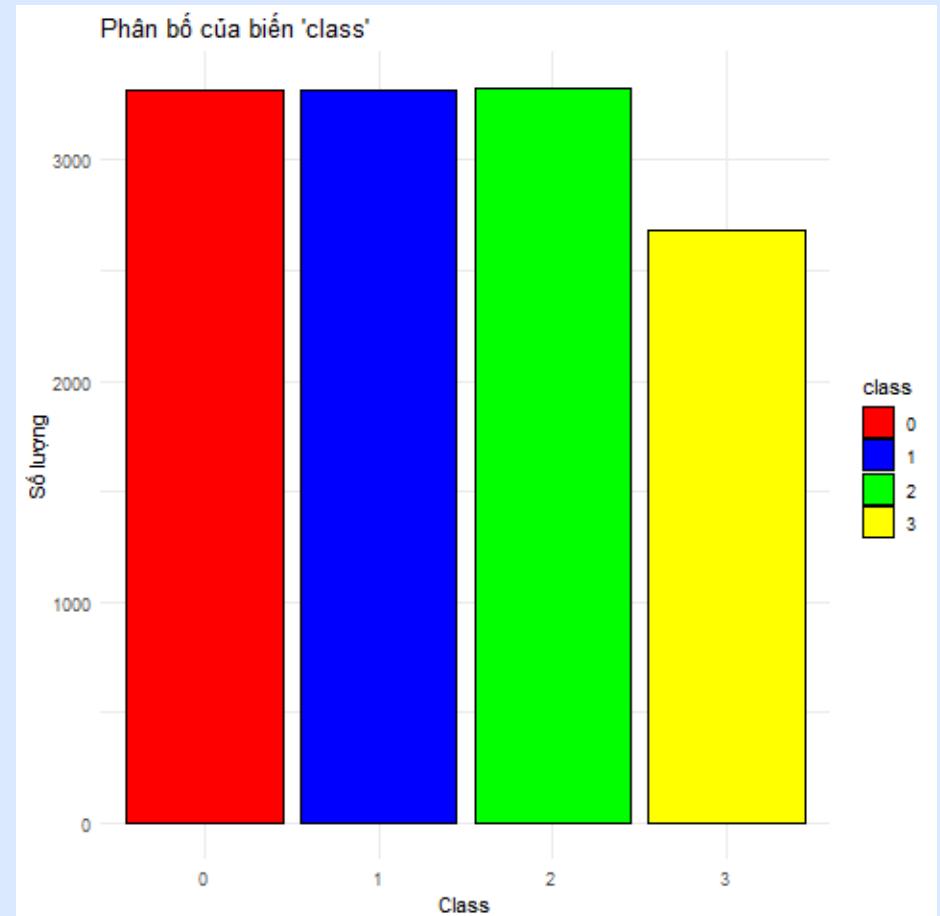
Gender

M

F



0
1



4. Đánh giá mô hình

- Mô hình random forest
 - Mô hình đạt hiệu suất cao với các chỉ số chính Balanced Accuracy, F1-Score đều trên 80%.
 - Mô hình hoạt động tốt và cân bằng, với hiệu suất cao cho cả hai lớp.
- Mô hình logistic regression
 - Mô hình Logistic Regression đạt Balanced Accuracy là 78.88% cho cả hai lớp, một kết quả khá tốt. Điều này cho thấy mô hình có khả năng phân biệt tương đối tốt giữa hai lớp trong tập dữ liệu.
 - Độ nhạy của lớp 1 thấp hơn lớp 0, có thể gây ra bỏ sót các trường hợp quan trọng, Lớp 1 có Specificity cao hơn, nghĩa là mô hình ít bị nhầm lẫn hơn khi dự đoán không thuộc lớp 1.

→ Sử dụng Random Forest là công cụ chính

IV. KẾT QUẢ TỔNG KẾT

...