ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH

TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN

KHOA TOÁN - TIN HỌC

XỬ LÍ SỐ LIỆU THỐNG KÊ

ĐỒ ÁN CUỐI KÌ: Body performance Analysis

Giảng viên: TS. Tô Đức Khánh

Nhóm sinh viên thực hiện:

Họ và tên	MSSV
Trần Duy An	22110007
Phạm Thái Thiên An	22110005
Phạm Xuân Bách	22110021
Nguyễn Tất Chiến	22110028
Vũ Xuân Hiệp	22110060

Ngày 14 tháng 1 năm 2025

1 Tổng Quan

1.1 Giới thiệu

Hiện nay, việc tập luyện thể thao không chỉ là một xu hướng mà còn trở thành một phần quan trọng trong lối sống lành mạnh của nhiều người. Các phong trào tập thể thao ngày càng phát triển mạnh mẽ, thu hút sự tham gia của nhiều nhóm tuổi và giới tính khác nhau. Để hiểu rõ hơn về hiệu quả của việc tập luyện thể dục và các yếu tố ảnh hưởng đến hiệu suất tập luyện, dữ liệu **bodyPerformance.csv** cung cấp một nguồn thông tin quý giá.

Tập dữ liệu này bao gồm thông tin của 13,393 người tham gia tập thể thao tại Hàn Quốc, với 12 biến đặc trưng:

- **age**: độ tuổi (từ 20 tới 64);
- **gender**: giới tính (**F**: nữ, **M**: nam);
- **height_cm**: chiều cao (đơn vị: cm);
- **weight_kg**: cân nặng (đơn vị: kg);
- **body fat_%**: phần trăm mỡ cơ thể (%);
- **diastolic**: huyết áp tâm trương (mmHg);
- **systolic**: huyết áp tâm thu (mmHg);
- **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);
- **class**: phân loại hiệu suất (**A**: tốt nhất, **B**, **C**, **D**).

Nhóm chúng tôi nghiên cứu và phân tích để đưa ra những đánh giá và khuyến nghị chính xác hơn nhằm nâng cao hiệu quả luyện tập.

1.2 Nhiệm vụ và Mục tiêu

1. **Sức khỏe tim mạch**:

   - Phân tích chỉ số diastolic, systolic để xác định sự khác biệt về chỉ số huyết áp của những người thuộc từng lớp hiệu suất.
   - Từ đó, đánh giá xem tập thể dục có thể cải thiện được sức khỏe tim mạch hay không.

2. **Thành phần cơ thể (Body Composition)**:

   - Phân tích chỉ số body_fat_
   - Từ đó, đánh giá liệu những người có tần suất tập thể dục cao hơn có cân nặng và tỉ lệ mỡ tối ưu hơn không.

2 Tìm Hiểu Dữ Liệu

2.1 Tổng quan

Đầu tiên ta sẽ hiển thị một số dòng dữ liệu :

	age	gender	height_cm	weight_kg	body_fat_percent	diastolic	systolic	grip_force	sit_and_bend_forward_cm	sit_ups_counts	broad_jump_cm	class
1	27.00	M	172.30	75.24	21.30	80.00	130.00	54.90	18.40	60.00	217.00	C
2	25.00	M	165.00	55.80	15.70	77.00	126.00	36.40	16.30	53.00	229.00	A
3	31.00	M	179.60	78.00	20.10	92.00	152.00	44.80	12.00	49.00	181.00	C
4	32.00	M	174.50	71.10	18.40	76.00	147.00	41.40	15.20	53.00	219.00	B
5	28.00	M	173.80	67.70	17.10	70.00	127.00	43.50	27.10	45.00	217.00	B
6	36.00	F	165.40	55.40	22.00	64.00	119.00	23.80	21.00	27.00	153.00	B
7	42.00	F	164.50	63.70	32.20	72.00	135.00	22.70	0.80	18.00	146.00	D
8	33.00	M	174.90	77.20	36.90	84.00	137.00	45.90	12.30	42.00	234.00	B
9	54.00	M	166.80	67.50	27.60	85.00	165.00	40.40	18.60	34.00	148.00	C
10	28.00	M	185.00	84.60	14.40	81.00	156.00	57.90	12.10	55.00	213.00	B

Tóm tắt về dữ liệu :

	age	gender	height_cm	weight_kg	body_fat_percent	diastolic	systolic	grip_force	sit_and_bend_forward_cm	sit_ups_counts	broad_jump_cm	class
X	Min. :21.00	Length:13391	Min. :125.0	Min. : 26.30	Min. : 3.00	Min. : 0.0	Min. : 0.0	Min. : 0.00	Min. :-25.00	Min. : 0.00	Min. : 0.0	Length:13391
X.1	1st Qu.:25.00	Class :character	1st Qu.:162.4	1st Qu.: 58.20	1st Qu.:18.00	1st Qu.: 71.0	1st Qu.:120.0	1st Qu.:27.50	1st Qu.: 10.90	1st Qu.:30.00	1st Qu.:162.0	Class :character
X.2	Median :32.00	Mode :character	Median :169.2	Median : 67.40	Median :22.80	Median : 79.0	Median :130.0	Median :37.90	Median : 16.20	Median :41.00	Median :193.0	Mode :character
X.3	Mean :36.78		Mean :168.6	Mean : 67.45	Mean :23.24	Mean : 78.8	Mean :130.2	Mean :36.96	Mean : 15.21	Mean :39.77	Mean :190.1	
X.4	3rd Qu.:48.00		3rd Qu.:174.8	3rd Qu.: 75.30	3rd Qu.:28.00	3rd Qu.: 86.0	3rd Qu.:141.0	3rd Qu.:45.20	3rd Qu.: 20.70	3rd Qu.:50.00	3rd Qu.:221.0	
X.5	Max. :64.00		Max. :193.8	Max. :138.10	Max. :78.40	Max. :156.2	Max. :201.0	Max. :70.50	Max. :213.00	Max. :80.00	Max. :303.0	

2.1.1 Tổng quan về dữ liệu

Số lượng biến Dữ liệu bao gồm 12 biến, trong đó có:

- 2 biến phân loại: *gender* và *class*.
- 10 biến số (bao gồm cả các chỉ số đo lường sức khỏe và hiệu suất).

Kích thước dữ liệu Dữ liệu bao gồm 13,393 dòng, đủ lớn để thực hiện các phân tích thống kê đáng tin cậy.

Thông tin chi tiết về các biến

- Các biến có giá trị đo lường thực tế và đa dạng, ví dụ:
  - *age* (độ tuổi): từ 20 đến 64 tuổi.
  - *height_cm* (chiều cao): từ 125 cm đến 193.8 cm.
  - *broad_jump_cm* (nhảy xa): từ 0 cm đến 303 cm.
- Có sự chênh lệch lớn giữa giá trị tối thiểu và tối đa ở một số biến, ví dụ: *sit_and_bend_forward_cm* từ -25 đến 213 cm, cho thấy cần kiểm tra dữ liệu bất thường hoặc ngoại lệ.

Các vấn đề cần 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]
- **Dữ liệu mất mát hoặc không hợp lệ:**
  - Chưa rõ mức độ mất dữ liệu, cần kiểm tra kỹ hơn.
- **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.

2.2 Làm sạch dữ liệu

Kiểm tra giá trị null trong bộ dữ liệu sau khi xóa các giá trị bất thường ở [1]:

Column_Name	DataType	Non_null_Values	Unique_Values	NaN_Values_Percentage
age	numeric	13391	44	0.00
gender	character	13391	2	0.00
height_cm	numeric	13391	467	0.00
weight_kg	numeric	13391	1398	0.00
body_fat_percent	numeric	13391	527	0.00
diastolic	numeric	13391	89	0.00
systolic	numeric	13391	102	0.00
grip_force	numeric	13391	550	0.00
sit_and_bend_forward_cm	numeric	13391	528	0.00
sit_ups_counts	numeric	13391	81	0.00
broad_jump_cm	numeric	13391	245	0.00
class	character	13391	4	0.00

- **Dữ liệu đầy đủ và không có giá trị thiếu (NaN):**
  - Tất cả các biến đều có số lượng giá trị không null (`Non_null_Values`) là 13,391, bằng với tổng số dòng dữ liệu. Điều này cho thấy không có giá trị thiếu (`NaN_Values`) trong bất kỳ cột nào.
  - Tỷ lệ giá trị thiếu (`NaN_Values_Percentage`) là 0% trên tất cả các cột.
- **Biến phân loại và số lượng giá trị duy nhất:**
  - Biến `gender` có 2 giá trị duy nhất, tương ứng với 2 giới tính (nam và nữ).

- Biến `class` có 4 giá trị duy nhất, đại diện cho các phân loại hiệu suất (A, B, C, D).

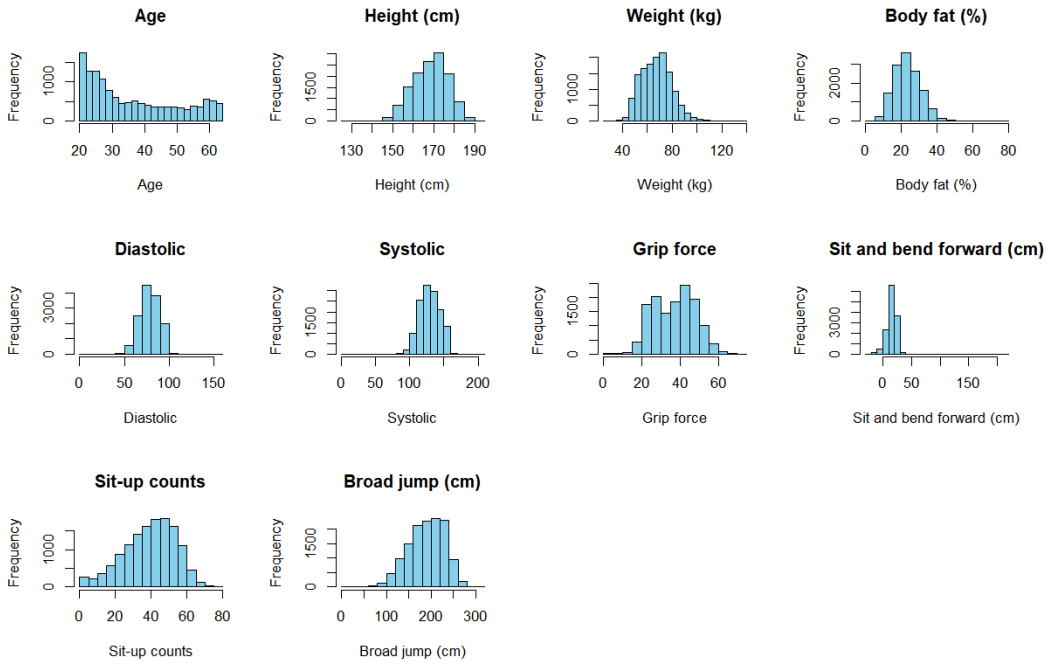
- **Biến số và độ phân giải:**
  - Các biến như `height_cm`, `weight_kg`, `body_fat_percent`, `grip_force`, và `broad_jump_cm` có số lượng giá trị duy nhất cao, thể hiện mức độ đa dạng dữ liệu.
  - Một số biến như `diastolic`, `systolic`, và `sit_ups_counts` có ít giá trị duy nhất hơn, có thể phù hợp với các nhóm giá trị cụ thể.

- **Cân Bằng Phân Lớp :**
  - Sau khi kiểm tra ta thấy các nhóm *class* (A, B, C, D) có được phân phối đồng đều :
    * Nhóm A: 3348 người.
    * Nhóm B: 3347 người.
    * Nhóm C: 3349 người.
    * Nhóm D: 3347 người.

2.3 Trực quan hóa dữ liệu

Mục tiêu của phần trực quan hóa dữ liệu này là so sánh và phân tích sâu hơn về các yếu tố ảnh hưởng đến hiệu suất thể thao của người tham gia, tập trung vào ba chủ đề chính: độ tuổi, giới tính và phân lớp hiệu suất. Bằng cách sử dụng các biểu đồ, chúng ta có thể nhận diện sự khác biệt giữa các nhóm tuổi và giới tính, cũng như khám phá mối quan hệ giữa các yếu tố này với phân lớp hiệu suất thể thao. Qua đó, việc trực quan hóa không chỉ giúp đánh giá sự phân bố các đặc tính cá nhân mà còn làm rõ các yếu tố có ảnh hưởng đến khả năng thể chất và kết quả tập luyện của từng nhóm người, từ đó đưa ra những nhận định chi tiết hơn về hiệu quả của các phong trào thể thao đối với từng đối tượng tham gia.

1. **Phân phối của các biến**:



(a) Phân Phối Các Biến

(a) **Age (Độ tuổi):** Phân phối lệch phải, tập trung ở nhóm tuổi từ 20-40, cho thấy đa phần người tham gia thuộc độ tuổi trẻ và trung niên. Số lượng giảm đáng kể ở nhóm từ 40 tuổi trở đi.

(b) **Height (Chiều cao):** Phân phối gần chuẩn, với chiều cao phổ biến trong khoảng 160-180 cm. Rất ít người có chiều cao dưới 150 cm hoặc trên 190 cm. Phân bố chiều cao tương đối bình thường và phù hợp với chiều cao trung bình của người Hàn Quốc.

(c) **Weight (Cân nặng):** Phân phối gần chuẩn nhưng hơi lệch phải, cho thấy cân nặng

thường nằm trong khoảng 50-80 kg. Số lượng người có cân nặng trên 100 kg rất ít, phản ánh nhóm đối tượng chủ yếu có thể trạng trung bình.

(d) **Body fat (Phần trăm mỡ cơ thể):** Phân phối lệch phải, tập trung phần lớn ở tỷ lệ mỡ cơ thể 10-30%. Số người có tỷ lệ mỡ trên 40% rất ít, phản ánh phần lớn người tham gia có mức độ mỡ cơ thể trong giới hạn lành mạnh.

(e) **Diastolic (Huyết áp tâm trương):** Phân phối gần chuẩn, với huyết áp tâm trương phổ biến từ 70-90 mmHg, cho thấy đa phần người tham gia có huyết áp trong khoảng bình thường.

(f) **Systolic (Huyết áp tâm thu):** Phân phối gần chuẩn, tập trung trong khoảng 110-140 mmHg. Điều này cho thấy huyết áp tâm thu của người tham gia chủ yếu nằm trong ngưỡng khỏe mạnh.

(g) **Sit-ups counts (Số lần gập bụng):** Phân phối gần chuẩn, số lần gập bụng thường nằm trong khoảng 30-60 lần. Điều này phản ánh thể trạng trung bình của nhóm người tập luyện.

(h) **Broad jump (Nhảy xa):** Phân phối gần chuẩn, phổ biến từ 150-250 cm. Số người nhảy xa trên 300 cm rất ít. Điều này cho thấy sức mạnh chân và khả năng bật nhảy của người tham gia khá đa dạng.

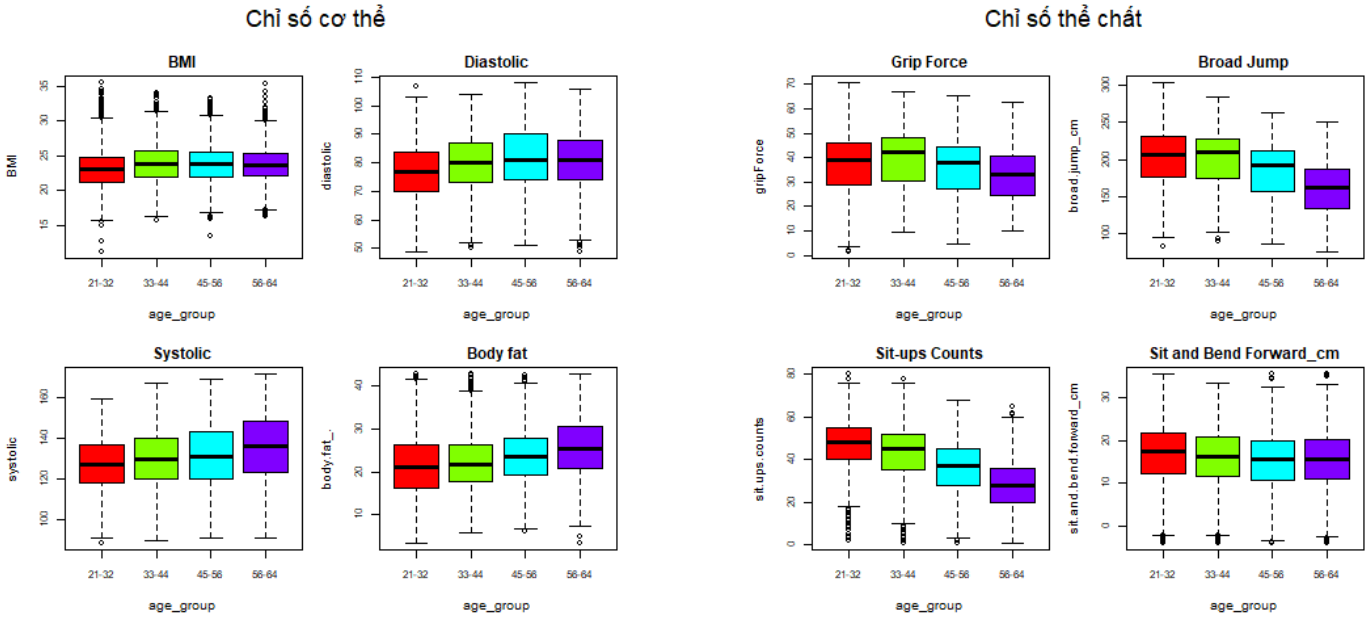
(i) **Nhận xét tổng quan:**

- Các biến sinh học như chiều cao, cân nặng, huyết áp, và lực kẹp đều có phân phối gần chuẩn, phản ánh tính đồng nhất về thể chất của nhóm người tham gia.
- Các biến như độ tuổi, phần trăm mỡ cơ thể và khả năng gập người lại có phân phối lệch, cho thấy sự khác biệt đáng kể giữa các cá nhân trong những đặc điểm này.
- Dữ liệu phản ánh một nhóm đối tượng tham gia đa dạng về độ tuổi, giới tính và khả năng thể chất. Đặc biệt, nhóm đối tượng trẻ tuổi và trung niên có xu hướng chiếm ưu thế, phù hợp với phong trào luyện tập thể thao hiện nay.

2. **So Sánh Giữa Độ Tuổi**:

Khám phá sự khác biệt và mối quan hệ giữa độ tuổi và các chỉ số như chiều cao, cân nặng, tỷ lệ mỡ cơ thể, huyết áp, và các yếu tố thể chất khác. Việc so sánh này giúp chúng ta nhận diện xu hướng thay đổi của các chỉ số cơ thể qua các độ tuổi, đồng thời đánh giá tác động của độ tuổi đến hiệu suất thể thao và tình trạng thể chất của người tham gia.

Ta sẽ bắt đầu kiểm tra các chỉ số thể chất và cơ thể :



(a) Chỉ số cơ thể (b) Chỉ số thể chất

Hình 1: Các chỉ số cơ thể và thể chất

(a) **BMI theo nhóm tuổi:**

- Xu hướng tăng nhẹ theo độ tuổi: Trung vị (đường kẻ đậm trong hộp) của BMI có xu hướng tăng dần từ nhóm tuổi 21-32 đến 56-64.
- Độ phân tán tương đối giống nhau giữa các nhóm tuổi: Kích thước hộp (biểu diễn khoảng tứ phân vị) tương đối giống nhau, cho thấy mức độ biến động của BMI trong các nhóm tuổi không khác biệt nhiều.
- Nhiều điểm ngoại lai (outliers): Xuất hiện nhiều điểm ngoại lai ở cả trên và dưới, đặc biệt ở nhóm tuổi 21-32, cho thấy có một số cá nhân có BMI cao hoặc thấp hơn đáng kể so với phần đông trong nhóm.

(b) **Huyết áp tâm trương theo nhóm tuổi:**

- Xu hướng tăng nhẹ theo độ tuổi: Trung vị của Diastolic cũng có xu hướng tăng dần

theo nhóm tuổi.

- Độ phân tán ít hơn so với BMI: Kích thước hộp nhỏ hơn so với BMI, cho thấy Diastolic ít biến động hơn.
- Ít điểm ngoại lai hơn: Số lượng điểm ngoại lai ít hơn so với BMI, chủ yếu xuất hiện ở phía dưới.

(c) **Huyết áp tâm thu theo nhóm tuổi:**

- Xu hướng tăng rõ rệt theo độ tuổi: Trung vị của Systolic tăng đáng kể theo từng nhóm tuổi, rõ rệt hơn so với BMI và Diastolic.
- Độ phân tán tương đối giống nhau: Mức độ biến động của Systolic tương đối giống nhau giữa các nhóm tuổi.
- Xuất hiện điểm ngoại lai ở cả trên và dưới: Có các điểm ngoại lai ở cả hai phía, cho thấy một số cá nhân có huyết áp tâm thu cao hoặc thấp hơn đáng kể so với phần đông trong nhóm.

(d) **Phần trăm mỡ cơ thể:**

- Xu hướng tăng nhẹ theo độ tuổi: Trung vị của Body fat có xu hướng tăng dần theo nhóm tuổi.
- Độ phân tán tương đối giống nhau: Mức độ biến động của Body fat tương đối giống nhau giữa các nhóm tuổi.
- Xuất hiện nhiều điểm ngoại lai ở phía dưới: Có nhiều điểm ngoại lai ở phía dưới, đặc biệt ở nhóm tuổi 21-32 và 56-64, cho thấy một số cá nhân có tỷ lệ mỡ cơ thể thấp hơn đáng kể so với phần đông trong nhóm.

(e) **Lực nắm tay theo nhóm tuổi:**

- Xu hướng giảm dần theo độ tuổi: Trung vị (đường kẻ đậm trong hộp) của Grip Force giảm dần từ nhóm tuổi 21-32 đến 56-64.
- Độ phân tán tương đối giống nhau giữa các nhóm tuổi: Kích thước hộp (biểu diễn khoảng tứ phân vị) tương đối giống nhau, cho thấy mức độ biến động của Grip Force trong các nhóm tuổi không khác biệt nhiều.
- Ít điểm ngoại lai: Chỉ có một vài điểm ngoại lai ở phía dưới, cho thấy hầu hết mọi người

đều có lực nắm tay trong khoảng giá trị bình thường.

(f) **Nhảy xa theo nhóm tuổi:**

- Xu hướng giảm dần theo độ tuổi: Trung vị của Broad Jump giảm dần theo nhóm tuổi, đặc biệt giảm mạnh ở nhóm 56-64.
- Độ phân tán tương đối giống nhau: Mức độ biến động của Broad Jump tương đối giống nhau giữa các nhóm tuổi.
- Xuất hiện một số điểm ngoại lai ở phía dưới: Có một số điểm ngoại lai ở phía dưới, cho thấy một số cá nhân có khả năng nhảy xa kém hơn đáng kể so với phần đông trong nhóm.

(g) **Độ linh hoạt ngồi và gập người theo nhóm tuổi:**

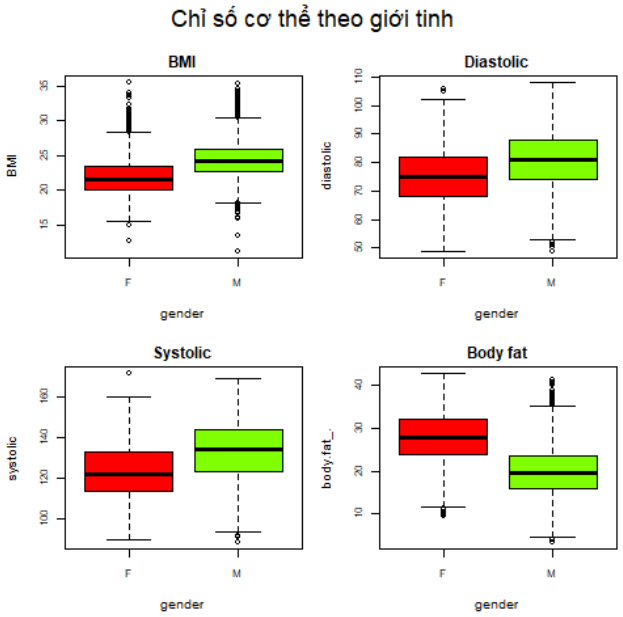
- Xu hướng giảm dần theo độ tuổi: Trung vị của Sit-ups Counts giảm dần theo nhóm tuổi.
- Độ phân tán tương đối giống nhau: Mức độ biến động của Sit-ups Counts tương đối giống nhau giữa các nhóm tuổi.
- Nhiều điểm ngoại lai ở nhóm tuổi 21-32: Có nhiều điểm ngoại lai ở phía dưới ở nhóm tuổi 21-32, cho thấy một số cá nhân có khả năng gập bụng kém hơn đáng kể. Nhóm tuổi này cũng có một số ít điểm ngoại lai ở phía trên, cho thấy một vài cá nhân có khả năng gập bụng tốt hơn đáng kể.

Nhận Xét Tổng Quan:

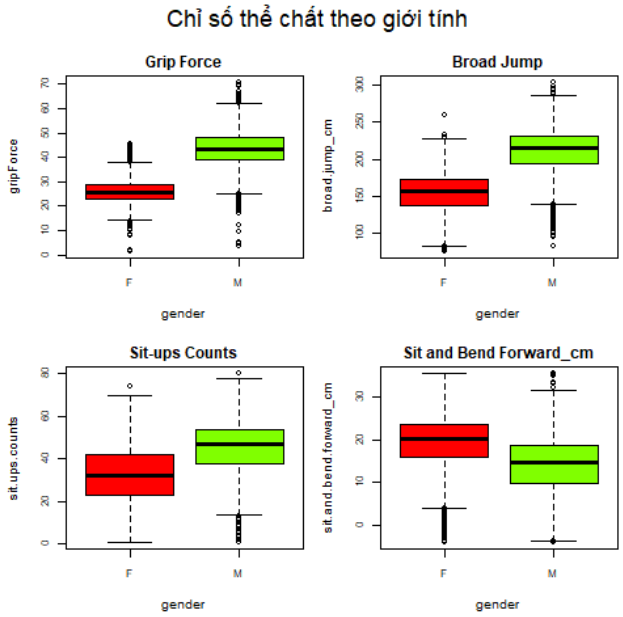
Các biểu đồ cho thấy xu hướng tăng các chỉ số cơ thể và giảm các chỉ số thể chất theo độ tuổi, phản ánh quá trình lão hóa tự nhiên. Tuy nhiên, tập luyện thể thao có thể giúp làm chậm quá trình này và duy trì sức khỏe tốt.

3. **So Sánh Giữa Giới Tính**

Tìm hiểu sự khác biệt giữa nam và nữ qua các chỉ số như chiều cao, cân nặng, tỷ lệ mỡ cơ thể, huyết áp và các yếu tố thể chất khác. Việc nghiên cứu này giúp chúng ta nhận diện các đặc điểm thể chất đặc trưng của từng giới, đồng thời đánh giá tác động của giới tính đến hiệu suất thể thao cũng như tình trạng sức khỏe của người tham gia:



(a) Chỉ số cơ thể



(b) Chỉ số thể chất

Hình 2: Các chỉ số cơ thể và thể chất

(a) **Chỉ số BMI (Body Mass Index):**

- Không có sự khác biệt lớn giữa nam và nữ: Trung vị (đường kẻ đậm trong hộp) của BMI ở nam và nữ gần như nhau.
- Độ phân tán tương tự nhau: Kích thước hộp (biểu diễn khoảng tứ phân vị) ở nam và nữ tương đối giống nhau, cho thấy mức độ biến động của BMI ở hai giới là tương tự.
- Xuất hiện nhiều điểm ngoại lai ở cả hai giới: Cho thấy có một số cá nhân có BMI cao hoặc thấp hơn đáng kể so với phần đông trong cùng giới.

(b) **Huyết áp tâm trương (Diastolic):**

- Nam giới có huyết áp tâm trương cao hơn: Trung vị của Diastolic ở nam cao hơn nữ.
- Độ phân tán ở nam lớn hơn: Hộp của nam dài hơn, cho thấy mức độ biến động của Diastolic ở nam cao hơn nữ.
- Ít điểm ngoại lai: Số lượng điểm ngoại lai ở cả hai giới đều ít.

(c) **Huyết áp tâm thu (Systolic):**

- Nam giới có huyết áp tâm thu cao hơn: Trung vị của Systolic ở nam cao hơn nữ.
- Độ phân tán ở nam lớn hơn: Hộp của nam dài hơn, cho thấy mức độ biến động của Systolic ở nam cao hơn nữ.
- Xuất hiện điểm ngoại lai ở cả hai giới: Có các điểm ngoại lai ở cả hai giới, nhưng số lượng ít.

(d) **Tỷ lệ mỡ cơ thể (Body Fat):**

- Nữ giới có tỷ lệ mỡ cơ thể cao hơn đáng kể: Trung vị của Body fat ở nữ cao hơn nam rất nhiều.
- Độ phân tán ở nữ lớn hơn: Hộp của nữ dài hơn, cho thấy mức độ biến động của Body fat ở nữ cao hơn nam.
- Xuất hiện một số điểm ngoại lai: Có một số điểm ngoại lai, chủ yếu ở phía dưới ở cả hai giới.

(e) **Grip Force (Lực nắm tay):**

- Nam giới có lực nắm tay mạnh hơn đáng kể: Trung vị (đường kẻ đậm trong hộp) của Grip Force ở nam cao hơn hẳn so với nữ.
- Độ phân tán ở nam lớn hơn: Kích thước hộp (biểu diễn khoảng tứ phân vị) ở nam lớn hơn, cho thấy mức độ biến động của Grip Force ở nam cao hơn nữ.
- Xuất hiện một số điểm ngoại lai ở cả hai giới: Cho thấy một số cá nhân có lực nắm tay yếu hơn đáng kể so với phần đông trong cùng giới.

(f) **Broad Jump (Bật xa):**

- Nam giới có khả năng nhảy xa tốt hơn: Trung vị của Broad Jump ở nam cao hơn nữ.
- Độ phân tán ở nam lớn hơn: Hộp của nam dài hơn, cho thấy mức độ biến động của Broad Jump ở nam cao hơn nữ.
- Xuất hiện một số điểm ngoại lai ở cả hai giới: Có một số điểm ngoại lai ở phía dưới ở cả hai giới, cho thấy một số cá nhân có khả năng nhảy xa kém hơn đáng kể so với phần đông.

(g) **Sit-ups Counts (Số lần gập bụng):**

- Nam giới có khả năng gập bụng tốt hơn: Trung vị của Sit-ups Counts ở nam cao hơn nữ.
- Độ phân tán ở nam lớn hơn: Hộp của nam dài hơn, cho thấy mức độ biến động của Sit-ups Counts ở nam cao hơn nữ.
- Xuất hiện nhiều điểm ngoại lai ở phía dưới ở cả hai giới: Cho thấy một số cá nhân có khả năng gập bụng kém hơn đáng kể so với phần đông.

(h) **Sit and Bend Forward (Độ dẻo dai):**

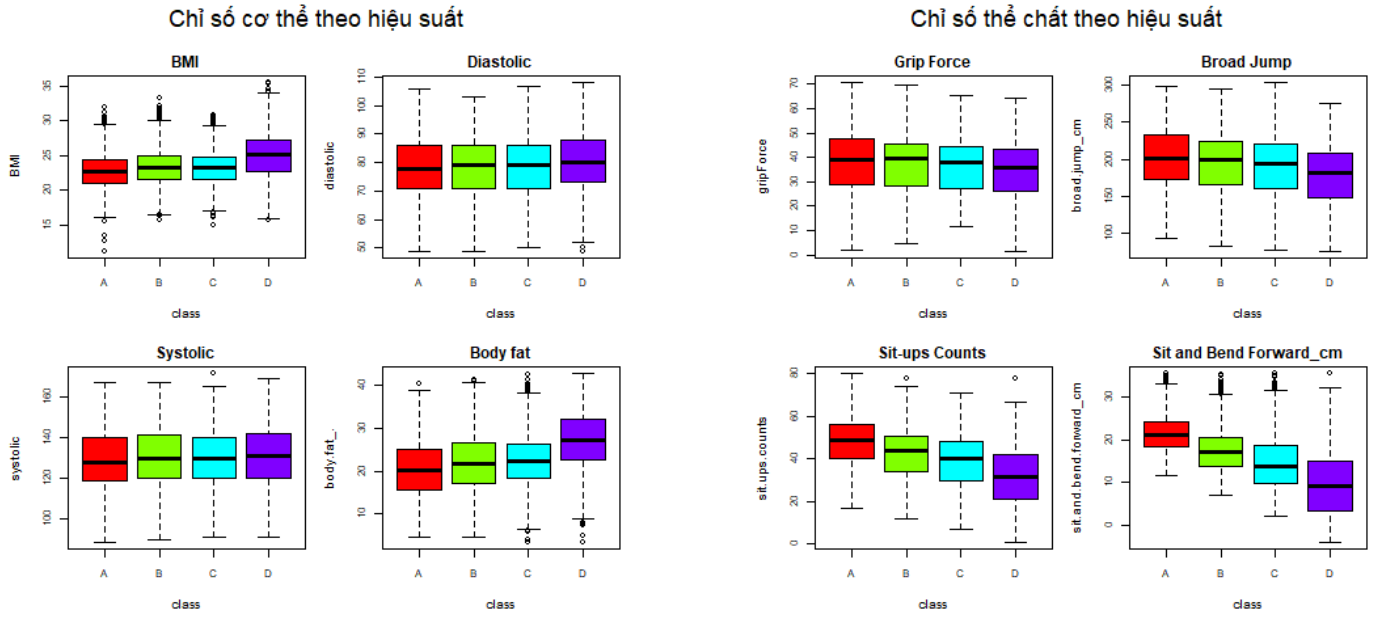
- Nữ giới có độ dẻo dai tốt hơn: Trung vị của Sit and Bend Forward ở nữ cao hơn nam.
- Độ phân tán tương đối giống nhau ở cả hai giới: Mức độ biến động của Sit and Bend Forward tương đối giống nhau ở nam và nữ.
- Xuất hiện nhiều điểm ngoại lai ở phía dưới ở cả hai giới: Cho thấy một số cá nhân có độ dẻo dai kém hơn đáng kể so với phần đông.

Nhận Xét Tổng Quan

- Có sự khác biệt rõ rệt giữa nam và nữ ở các chỉ số cơ thể (đặc biệt là huyết áp và tỷ lệ mỡ cơ thể) và các chỉ số thể chất (sức mạnh và độ dẻo dai).
- Những khác biệt này có thể liên quan đến các yếu tố sinh học, nội tiết tố, cấu trúc cơ thể và mức độ hoạt động thể chất khác nhau giữa hai giới.

4. **So Sánh Giữa hiệu suất**

Phân tích hiệu suất thể thao của các nhóm người tham gia, chia thành 4 lớp A, B, C, D. Mỗi lớp phản ánh mức độ khác nhau của khả năng thể chất, bao gồm sức bền, tốc độ, sức mạnh và độ dẻo dai, giúp chúng ta hiểu rõ hơn về sự khác biệt trong hiệu suất giữa các nhóm và đánh giá các yếu tố ảnh hưởng đến khả năng vận động của từng người.



(a) Chỉ số cơ thể

(b) Chỉ số thể chất

Hình 3: Các chỉ số cơ thể và thể chất

(a) **Chỉ số BMI**:

- Mối tương quan nghịch rõ ràng: Nhóm A (hiệu suất tốt nhất) có trung vị BMI thấp nhất, và tăng dần ở các nhóm B, C, và cao nhất ở nhóm D (hiệu suất kém nhất). Điều này khẳng định mối tương quan nghịch giữa BMI và hiệu suất tập luyện thể thao: hiệu suất càng cao, BMI có xu hướng càng thấp.
- Độ phân tán tương đối giống nhau: Kích thước các hộp tương đối giống nhau, cho thấy sự biến động của BMI trong mỗi nhóm hiệu suất không quá khác biệt.
- Nhiều điểm ngoại lai: Nhiều điểm ngoại lai ở cả phía trên và dưới cho thấy sự đa dạng về BMI trong từng nhóm, có thể do các yếu tố khác như gen, chế độ ăn uống, v.v.

(b) **Chỉ số huyết áp tâm trương**:

- Không có sự khác biệt đáng kể: Trung vị của Diastolic gần như nhau giữa các nhóm, cho thấy huyết áp tâm trương không có mối tương quan rõ ràng với hiệu suất tập luyện

thể thao trong mẫu nghiên cứu này.

- Độ phân tán tương đối giống nhau: Mức độ biến động của Diastolic tương đối giống nhau giữa các nhóm.
- Ít điểm ngoại lai: Số lượng điểm ngoại lai ít.

(c) **Chỉ số huyết áp tâm thu**:

- Mối tương quan nghịch: Nhóm A có trung vị Systolic thấp nhất, tăng dần ở các nhóm B, C, và cao nhất ở nhóm D. Điều này cho thấy hiệu suất tập luyện thể thao tốt hơn có liên quan đến huyết áp tâm thu thấp hơn.
- Độ phân tán tương đối giống nhau: Mức độ biến động của Systolic tương đối giống nhau giữa các nhóm.
- Xuất hiện điểm ngoại lai: Có các điểm ngoại lai ở cả hai phía, nhưng không nhiều.trong nhóm.

(d) **Tỷ lệ mỡ cơ thể (Body Fat):**

- Mối tương quan nghịch rõ ràng: Nhóm A có trung vị Body fat thấp nhất, tăng dần ở các nhóm B, C, và cao nhất ở nhóm D. Điều này khẳng định hiệu suất tập luyện thể thao tốt hơn có liên quan đến tỷ lệ mỡ cơ thể thấp hơn.
- Độ phân tán tương đối giống nhau: Mức độ biến động của Body fat tương đối giống nhau giữa các nhóm.
- Nhiều điểm ngoại lai ở phía dưới: Nhiều điểm ngoại lai ở phía dưới cho thấy một số cá nhân có tỷ lệ mỡ cơ thể thấp hơn đáng kể so với phần đông trong nhóm, có thể là do tập luyện chuyên sâu hoặc các yếu tố khác.

(e) **Grip Force (Lực nắm tay):**

- Xu hướng giảm: Nhóm A (hiệu suất tốt nhất) có trung vị Grip Force cao nhất, và có xu hướng giảm dần ở các nhóm B, C, và thấp nhất ở nhóm D (hiệu suất kém nhất). Điều này cho thấy mối tương quan thuận giữa Grip Force và hiệu suất tập luyện thể thao: hiệu suất càng cao, lực nắm tay có xu hướng càng mạnh.
- Độ phân tán: Nhóm A có độ phân tán lớn nhất (hộp dài nhất). Các nhóm còn lại có độ phân tán tương đối giống nhau.

– Ít điểm ngoại lai: Chỉ có một vài điểm ngoại lai ở phía dưới, chủ yếu ở nhóm D.

(f) **Broad Jump (Bật xa):**

– Xu hướng giảm: Nhóm A có trung vị Broad Jump cao nhất, tiếp theo là B, C và D. Điều này cho thấy hiệu suất tập luyện thể thao tốt hơn có liên quan đến khả năng nhảy xa tốt hơn.

– Độ phân tán: Các nhóm có độ phân tán tương đối giống nhau.

– Một số điểm ngoại lai: Có một số điểm ngoại lai ở phía dưới, cho thấy một số cá nhân có khả năng nhảy xa kém hơn đáng kể so với phần đông trong nhóm.

(g) **Sit-ups Counts (Số lần gập bụng):**

– Xu hướng giảm: Nhóm A có trung vị Sit-ups Counts cao nhất, và giảm dần ở các nhóm B, C, D. Khẳng định hiệu suất tập luyện thể thao tốt hơn liên quan đến khả năng gập bụng tốt hơn.

– Độ phân tán: Các nhóm có độ phân tán tương đối giống nhau.

– Điểm ngoại lai: Nhóm D có một điểm ngoại lai ở phía trên, các nhóm còn lại chủ yếu có điểm ngoại lai ở phía dưới.

(h) **Sit and Bend Forward (Độ dẻo dai):**

– Xu hướng giảm: Nhóm A có trung vị Sit and Bend Forward cao nhất, tiếp theo là B, C và D. Cho thấy hiệu suất tập luyện thể thao tốt hơn có liên quan đến độ dẻo dai tốt hơn.

– Độ phân tán: Các nhóm có độ phân tán tương đối giống nhau.

– Nhiều điểm ngoại lai: Có nhiều điểm ngoại lai ở phía dưới, đặc biệt là ở nhóm D, cho thấy một số cá nhân có độ dẻo dai kém hơn đáng kể so với phần đông.

(i) **Nhận Xét Tổng Quan**:

– Tập luyện thể thao thường xuyên và hiệu quả giúp cải thiện sức khỏe, nâng cao hiệu suất vận động, thể hiện qua việc:

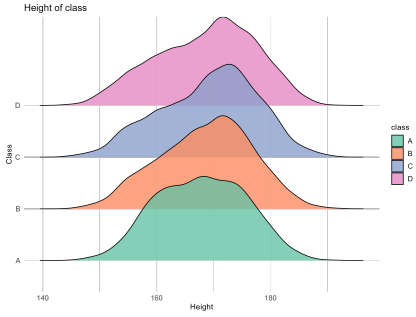
∗ Giảm BMI, huyết áp tâm thu và tỷ lệ mỡ cơ thể.

∗ Cải thiện các chỉ số thể chất (lực nắm tay, khả năng nhảy xa, số lần gập bụng và độ dẻo dai).

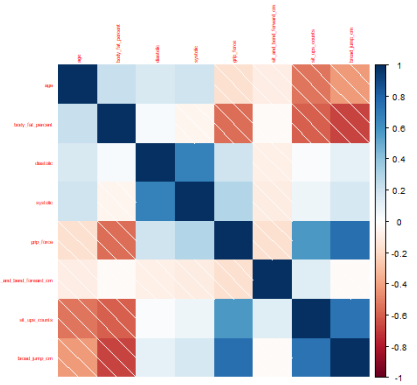
– Nhóm hiệu suất A (tốt nhất) có các chỉ số cơ thể và thể chất tốt nhất.

– Hiệu suất tập luyện càng tốt thì các chỉ số cơ thể (BMI, huyết áp tâm thu, tỷ lệ mỡ cơ thể) càng thấp và các chỉ số thể chất (lực nắm tay, nhảy xa, gập bụng, độ dẻo dai) càng cao. Tập luyện thể thao thường xuyên là yếu tố quan trọng để cải thiện sức khỏe và nâng cao hiệu suất.

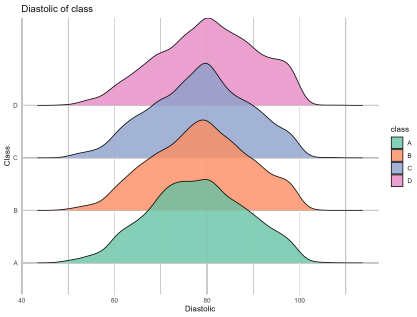
5. Một Số Biểu Đồ Khác



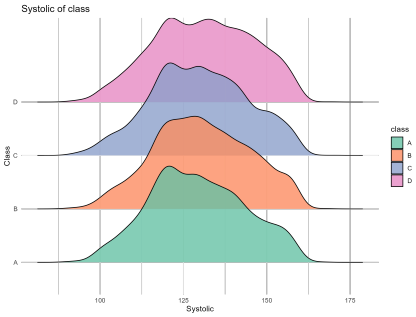
Hình 4: Biểu đồ chiều cao của lớp



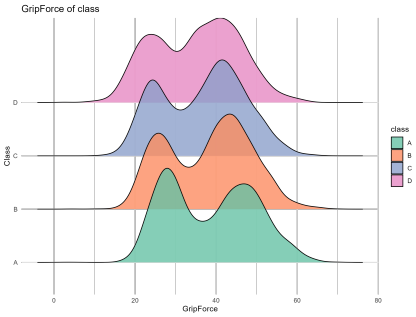
Hình 5: Biểu đồ tương quan các biến



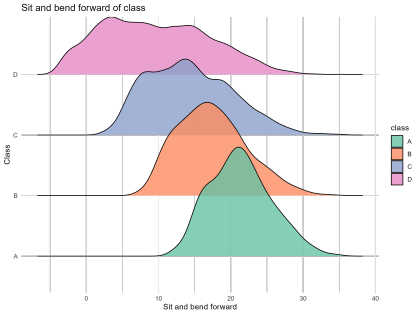
Hình 6: Biểu đồ huyết áp tâm trương của lớp



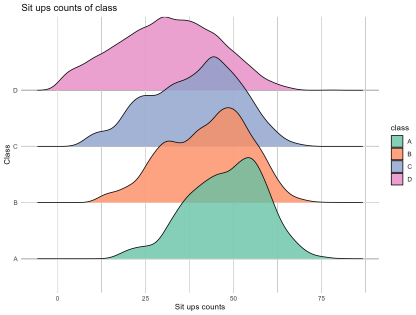
Hình 7: Biểu đồ huyết áp tâm thu của lớp



Hình 8: Biểu đồ lực nắm tay của lớp

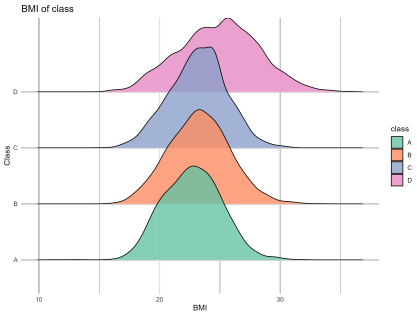


Hình 9: Biểu đồ độ dẻo dai của lớp

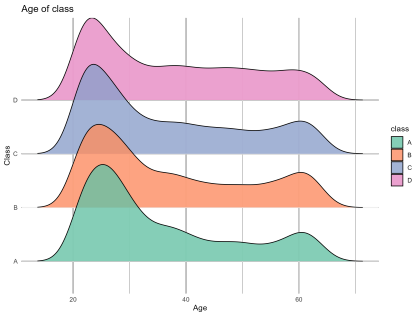


Hình 10: Biểu đồ số lần hít xà của lớp

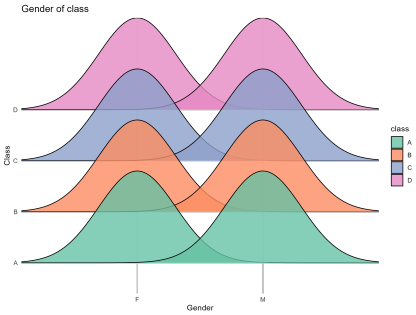
Hình 11: Biểu đồ nhảy xa của lớp



Hình 12: Biểu đồ chỉ số BMI của lớp

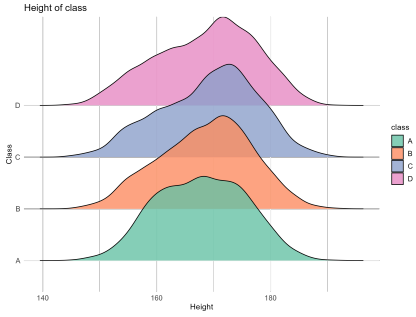


Hình 13: Biểu đồ độ tuổi của lớp

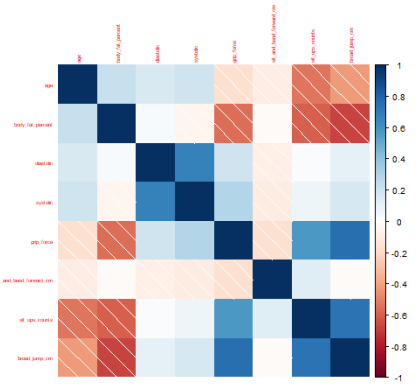


Hình 14: Biểu đồ giới tính của lớp

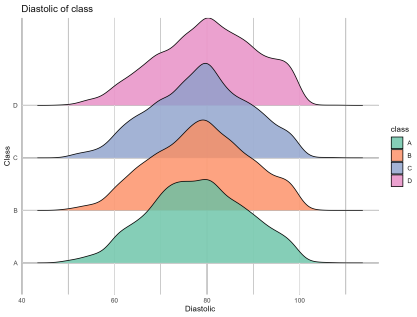
6. Một Số Biểu Đồ Khác



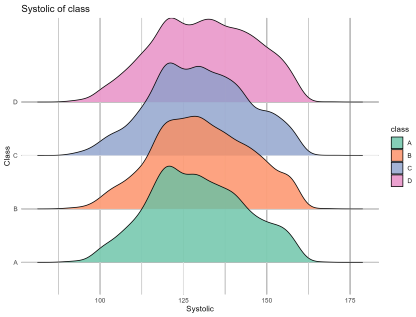
Hình 15: Biểu đồ chiều cao của lớp



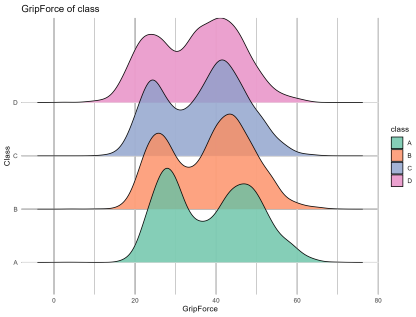
Hình 16: Biểu đồ tương quan các biến



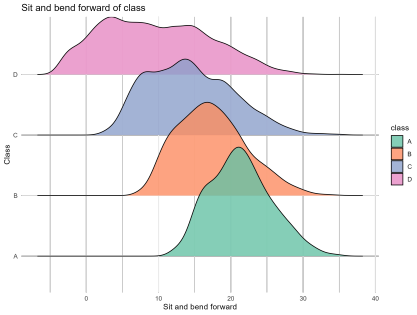
Hình 17: Biểu đồ huyết áp tâm trương của lớp



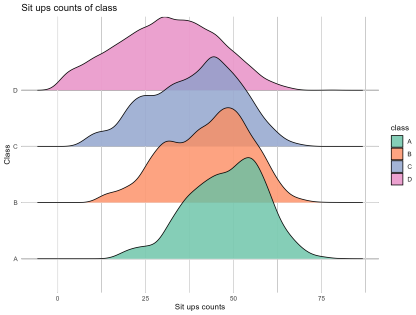
Hình 18: Biểu đồ huyết áp tâm thu của lớp



Hình 19: Biểu đồ lực nắm tay của lớp

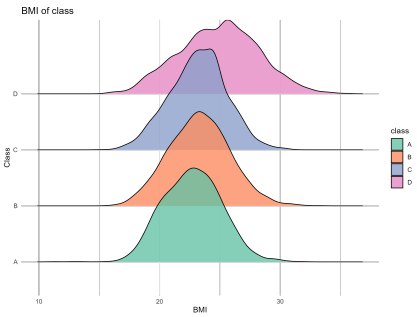


Hình 20: Biểu đồ độ dẻo dai của lớp

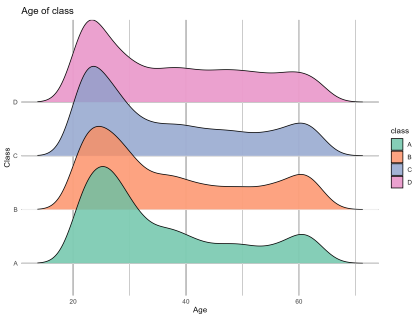


Hình 21: Biểu đồ số lần hít xà của lớp

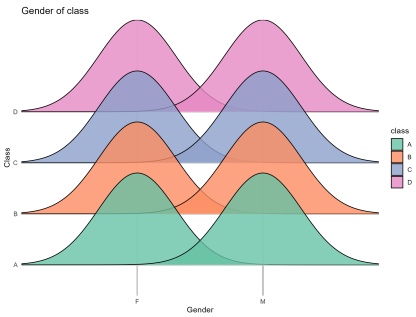
Hình 22: Biểu đồ nhảy xa của lớp



Hình 23: Biểu đồ chỉ số BMI của lớp



Hình 24: Biểu đồ độ tuổi của lớp



Hình 25: Biểu đồ giới tính của lớp

7. **Một Số Bảng Tóm Tắt Khác**

Gender	Total Grip Force	Mean Grip Force	Median Grip Force	Max Grip Force	Min Grip Force
F	127179.4	25.81799	25.6	45.5	0
M	367877.8	43.44842	43.3	70.5	0

Bảng 1: Tổng hợp theo giới tính

Class	Total Grip Force	Mean Grip Force	Median Grip Force	Max Grip Force	Min Grip Force
A	129285.4	38.61572	39.05	70.5	2.1
B	126886.2	37.91044	39.6	69.9	0
C	122515.3	36.58266	38	65.2	0
D	116370.2	34.74775	35.6	70.4	0

Bảng 2: Tổng hợp theo loại hiệu suất

Age Group	Total Grip Force	Mean Grip Force	Median Grip Force	Max Grip Force	Min Grip Force
1	217786	37.60766	38.2	70.5	0
2	108378.7	40.71326	42.6	70.4	9.1
3	68792.56	36.86632	38.3	65.4	0
4	56696.3	32.54667	31.4	62.9	10.4
NA	43403.68	32.58535	33.4	59.1	10.2

Bảng 3: Tổng hợp theo nhóm tuổi

	class	avg_age	avg_height_cm	avg_weight_kg	avg_body_fat_percent	avg_diastolic	avg_sys
1	A	35.27	167.87	64.42	20.54	77.90	12
2	B	37.07	168.58	66.61	22.04	78.66	13
3	C	36.70	169.16	66.76	22.64	78.55	12
4	D	38.06	168.63	72.00	27.74	80.08	13

Bảng 4: Tổng hợp theo nhóm lớp Hiệu suất

Những bảng tổng hợp và biểu đồ này sẽ cung cấp cái nhìn tổng quan và chi tiết về các yếu tố ảnh hưởng đến sức khỏe, giúp đưa ra các quyết định và nghiên cứu về con người hiệu quả hơn.

2.4 Phát Biểu và Kiểm Định Giả Thuyết

- Ở phần nhận xét về phân phối, những biến BMI, MAP và broad_jump_cm gần với phân phối chuẩn hơn những biến còn lại.
- Chúng tôi tiến hành kiểm định ANOVA cho những biến giả định là tuân theo phân phối chuẩn.

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

- Đặt giả thuyết:
  - H_0: Trung bình chỉ số 'BMI','MAP','broad_jump_cm' là như nhau.
  - H_1: 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.

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

a) Age

- Đặt giả thuyết:
  - H_0: Phân phối của 'age' trong 4 nhóm là giống nhau
  - H_1: 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.

b) Body_fat_percent

- Đặt giả thuyết:
  - H_0: Phân phối của 'body_fat_percent' trong 4 nhóm giống nhau.
  - H_1: 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 H_0.

Kết luận: Phân phối của 'body_fat_percent' trong 4 nhóm có sự khác biệt.

c) Grip_force

- Đặt giả thuyết:
  - H_0: Phân phối của 'grip_force' trong 4 nhóm giống nhau
  - H_1: 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 H_0.

Kết luận: Phân phối của 'Grip_force' trong 4 nhóm có sự khác biệt.

d) Sit_and_bend_forward_cm

- Đặt giả thuyết:
  - H_0: Phân phối của 'sit_and_bend_forward_cm' trong 4 nhóm giống nhau
  - H_1: Phân phối của 'sit_and_bend_forward_cm' trong 4 nhóm có sự khác biệt.

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

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 H_0.

Kết luận: Phân phối của 'Sit_and_bend_forward_cm' trong 4 nhóm có sự khác biệt.

e) Sit_ups_counts

- Đặt giả thuyết
  - H_0: Phân phối của 'sit_ups_counts' trong 4 nhóm giống nhau
  - H_1: Phân phối của 'sit_ups_counts' trong 4 nhóm có sự khác biệt.

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

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 H_0.

Kết luận: Phân phối của 'Sit_ups_counts' trong 4 nhóm có sự khác biệt.

2.4.3 Kiểm định Chi-square cho biến Gender

- Kiểm định Chi-square.
- H_0: Class và Gender độc lập.
- H_1: Class và Gender có sự phụ thuộc.

Tất cả các biến đều có chỉ số p-value < 0.05 nên chúng có ý nghĩa đối với biến phân loại "class".

Test	Chi-square test
Data	Gender by class
Chi-squared	112.77
df	3
p-value	$< 2.2 \times 10^{-16}$

Bảng 11: Chi-square test results for Gender by class.

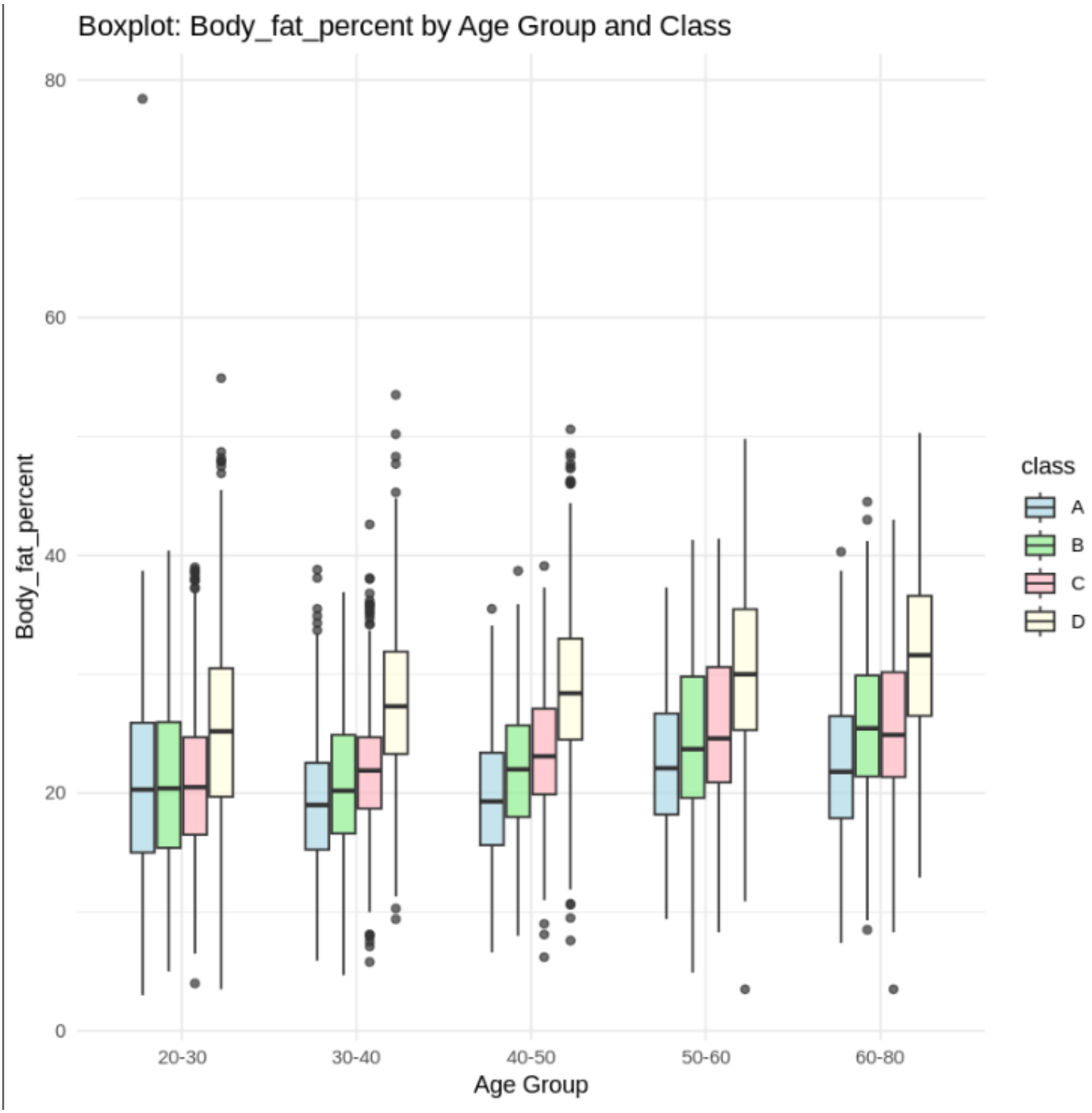
2.4.4 Hiệu quả của việc tập thể dục

Phân nhóm theo độ tuổi

Chúng tôi phân chia dữ liệu thành các nhóm tuổi sau:

- 20–30
- 30–40
- 40–50
- 50–60
- 60–70

2.4.4.1 So sánh chỉ số `body_fat_percent` giữa các nhóm tuổi



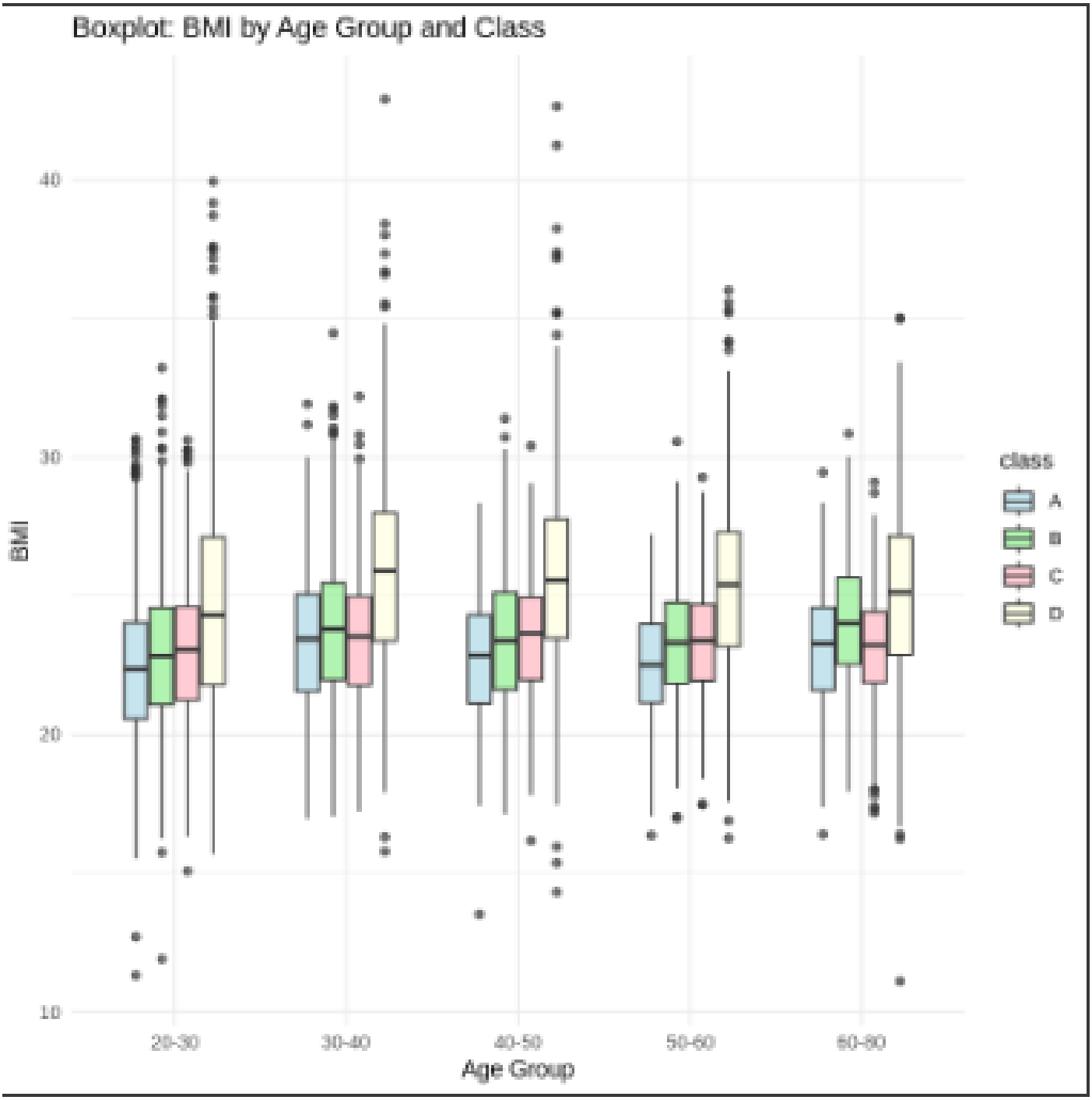
Hình 26: Chỉ số `body_fat_percent` giữa các nhóm tuổi theo từng class.

Nhận xét:

- **Sự khác biệt rõ rệt giữa các nhóm:**
  - Nhóm A (màu xanh nhạt) duy trì tỷ lệ mỡ cơ thể thấp nhất ở tất cả các độ tuổi.
  - Điều này chứng minh rằng việc tập luyện đều đặn và lối sống năng động giúp kiểm soát tỷ lệ mỡ cơ thể hiệu quả, bất kể độ tuổi.

- **Khoảng cách giữa các nhóm:**
  - 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 (60–80), nhóm A vẫn duy trì tỷ lệ mỡ cơ thể tốt hơn so với người trẻ tuổi hơn thuộc nhóm C và D.
- **Tính khả thi của việc cải thiện:**
  - Sự tồn tại của nhóm A ở mọi độ tuổi chứng minh rằng không bao giờ là quá muộn để bắt đầu tập luyện.
  - Việc duy trì tỷ lệ mỡ cơ thể hợp lý hoàn toàn có thể đạt được thông qua tập luyện, không phụ thuộc vào tuổi tác.
- **Lợi ích lâu dài:**
  - 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.

2.4.4.2 So sánh chỉ số BMI giữa các nhóm tuổi



Hình 27: Chỉ số BMI giữa các nhóm tuổi theo từng class.

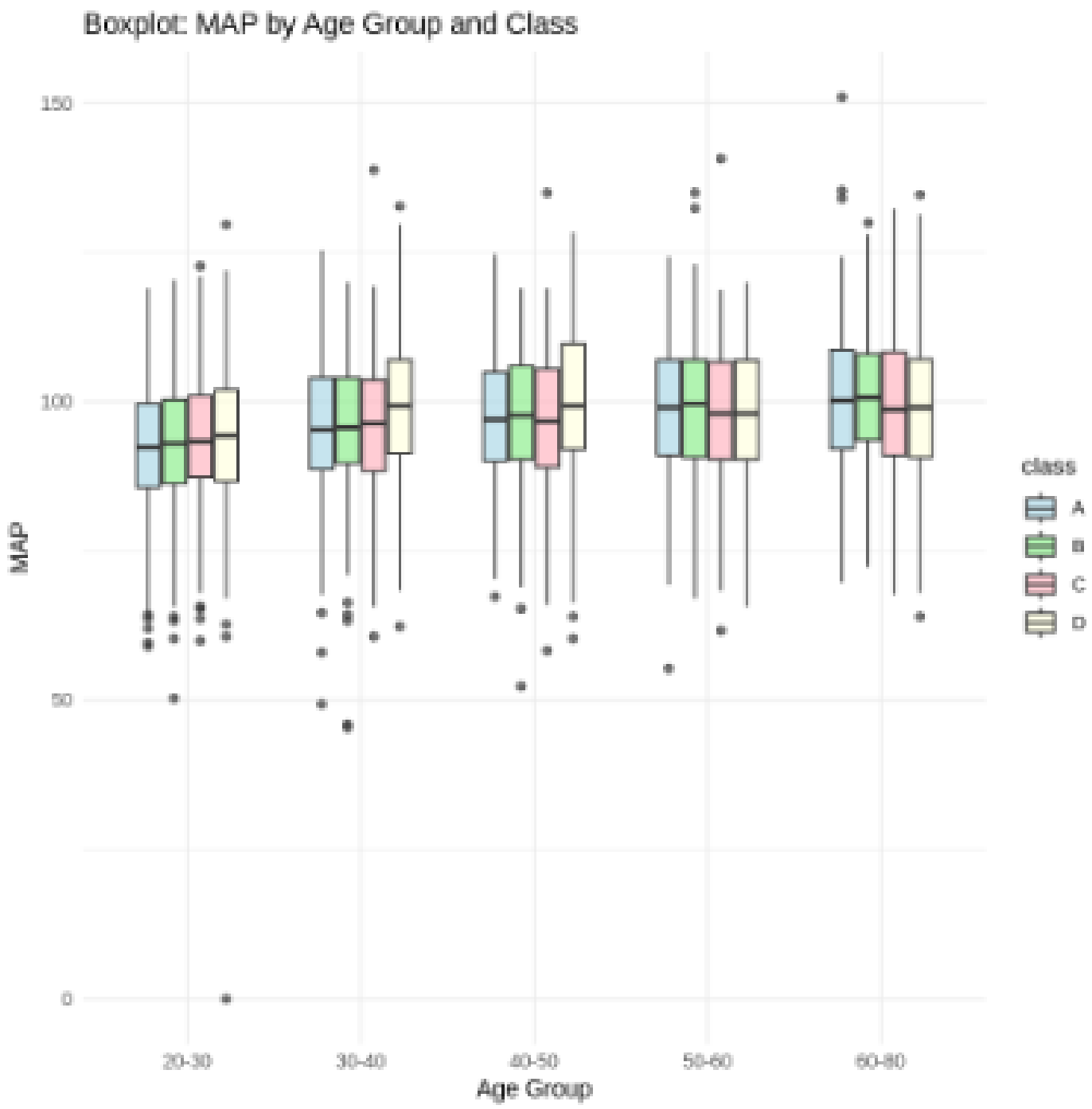
Nhận xét:

- **Sự khác biệt giữa các nhóm tập luyện:**
  - Nhóm A (màu xanh nhạt) duy trì chỉ số BMI ổn định và nằm trong vùng khỏe mạnh ở mọi độ tuổi.
  - Điều này cho thấy tập luyện thường xuyên giúp duy trì cân nặng lý tưởng, không phụ thuộc

vào tuổi tác.

- **Bằng chứng về hiệu quả tập luyện:**
  - Khoảng cách rõ rệt về BMI giữa nhóm A và nhóm D (không tập luyện) xuất hiện ở mọi nhóm tuổi.
  - Người lớn 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.
- **Động lực để bắt đầu tập luyện:**
  - Sự hiện diện của nhóm A ở mọi độ tuổi là minh chứng rằng không bao giờ là quá muộn để bắt đầu tập luyện.
  - Ngay cả ở nhóm tuổi cao (60–80), vẫn có thể duy trì BMI khỏe mạnh thông qua tập luyện đều đặn.
- **Lợi ích lâu dài:**
  - 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.

2.4.4.3 So sánh chỉ số MAP giữa các nhóm tuổi



Hình 28: Chỉ số MAP giữa các nhóm tuổi theo từng class.

Nhận xét:

- **Tác động tích cực của tập luyện đến 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 của tập luyện qua các độ 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.
- **Bằng chứng về lợi ích sức khỏe:**
  - Các điểm ngoại lai (outliers) ít xuất hiện hơn ở nhóm A.
  - Điều này gợi ý rằng tập luyện đều đặn giúp giảm nguy cơ huyết áp bất thường.
- **Động lực để duy trì tập luyện:**
  - 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 (60–80), tập luyện vẫn mang lại hiệu quả ổn định huyết áp.

***** Kết luận tổng thể:**

Biểu đồ `MAP` này, cùng với các biểu đồ về tỷ lệ mỡ cơ thể và `BMI` trước đó, khẳng định tầm quan trọng của việc tập luyện thể dục đối với sức khỏe tổng thể. Tập luyện không chỉ giúp kiểm soát cân nặng mà còn hỗ trợ duy trì huyết áp ổn định, một yếu tố quan trọng cho sức khỏe tim mạch ở mọi lứa tuổi.

3 Xây Dựng Model

Trong phần này, chúng tôi tiến hành xây dựng mô hình dựa trên dữ liệu `bodyPerformance.csv`, bao gồm các bước xử lý dữ liệu, xây dựng mô hình và đánh giá kết quả dự đoán.

3.1 Random Forest

3.1.1 Xử Lý Dữ Liệu

1. **Chuyển đổi trạng thái BMI:** Một hàm chuyển đổi trạng thái BMI được tạo ra để phân loại dữ liệu thành 4 nhóm chính: `Underweight`, `Healthy`, `Overweight`, và `Obese`.

```
convert_bmi <- function(y){
    if(y < 18.5){
      y <- "Underweight"
    }else
      if(y >= 18.5 & y < 25){
      y <- "Healthy"
    }else
      if(y >= 25 & y < 29.9){
      y <- "Overweight"
    }else{
      y <- "Obese"
    }
}
```

2. **Gắn trạng thái BMI vào dữ liệu:**

```
data_cleaned$BMI_status <- sapply(data_cleaned$BMI,
                                  FUN = convert_bmi)
```

3. **Kiểm tra kết quả phân loại BMI:**

```
data_cleaned %>%
  ggplot(aes(BMI_status, BMI, fill = BMI_status)) +
  geom_boxplot() +
  theme_minimal() +
  theme(legend.position = "none")
```

4. **Xóa các cột không cần thiết:**

```
data_cleaned <- data_cleaned %>% dplyr::select(-weight_kg, -height_cm, -BMI)
```

5. **Xử lý giá trị thiếu và chuyển đổi kiểu dữ liệu:**

```
colSums(is.na(data_cleaned))
data_cleaned$gender <- as.character(data_cleaned$gender)
data_cleaned$class <- as.character(data_cleaned$class)
data_cleaned$BMI_status <- as.character(data_cleaned$BMI_status)
data_cleaned$gender[data_cleaned$gender=="M"] <- "0"
data_cleaned$gender[data_cleaned$gender=="F"] <- "1"
data_cleaned$class[data_cleaned$class=="A"] <- "0"
data_cleaned$class[data_cleaned$class=="B"] <- "1"
data_cleaned$class[data_cleaned$class=="C"] <- "2"
data_cleaned$class[data_cleaned$class=="D"] <- "3"
data_cleaned$BMI_status[data_cleaned$BMI_status=="Underweight"] <- "0"
data_cleaned$BMI_status[data_cleaned$BMI_status=="Healthy"] <- "1"
data_cleaned$BMI_status[data_cleaned$BMI_status=="Overweight"] <- "2"
data_cleaned$BMI_status[data_cleaned$BMI_status=="Obese"] <- "3"
data_cleaned <- data_cleaned %>% mutate_if(is.character, as.factor)
```

6. **Đưa 4 phân lớp về 2 phân lớp:** Việc đưa 4 phân lớp về phân lớp nhằm mục đích giảm độ phức tạp và đơn giản hóa bài toán. Chúng ta sẽ đưa phân lớp A thành "0" và B,C,D thành "1".

```
df_accepted <-data_cleaneddf[data_cleaned$class == 0, ]
df_denied <- data_cleaned[!(data_cleaned$class == 0), ]
df_denied$class <- 1
data_cleaned_1 <- rbind(df_accepted, head(df_denied, 3318))
data_cleaned_1 <- data_cleaned_1[sample(nrow(data_cleaned_1)),]
data_cleaned_1 <- data_cleaned_1[, !(names(data_cleaned_1) %in% c("age", "gende
```

7. **Thêm cột MAP:** tổng hợp từ hai chỉ số diastolic và systolic

```
calculate_map <- function(data, diastolic, systolic) {
    if (!(diastolic %in% names(data)) || !(systolic %in% names(data))) {
    stop("Cột diastolic hoặc systolic không tồn tại trong dữ liệu.")}
data$MAP <- data[[diastolic]] + (data[[systolic]] - data[[diastolic]]) / 3
return(data)
}
data_cleaned_1 <- calculate_map(data_cleaned_1, "diastolic", "systolic")
```

8. **Kiểm tra dữ liệu:**

```
glimpse(data_cleaned_1)
prop.table(table(data_cleaned_1$class))
```

3.1.2 Xây dựng model

1. **Chia tập dữ liệu:** Chia dữ liệu thành tập huấn luyện (80%) và tập kiểm tra (20%) bằng cách sử dụng hàm `initial_split`.

```
set.seed(120)
index <- initial_split(data_cleaned_1, prop = 0.8, strata = "class")
data_train <- training(index)
```

```
data_test <- testing(index)
```

2. **Huấn luyện mô hình Random Forest:** Mô hình Random Forest được huấn luyện với 500 cây quyết định.

```
model_rf_1 <- randomForest(class~body_fat_percent+ grip_force +
                           sit_ups_counts +
                           broad_jump_cm +
                           BMI_status + MAP,
                           data = data_train,
                           importance = TRUE,
                           ntree = 500)
model_rf_1
```

3. **Dự đoán trên tập kiểm tra:** Tiến hành dự đoán trên tập kiểm tra.

```
data_mod1 <- predict(model_rf_1,
                     newdata = data_test)
```

4. **Đánh giá kết quả dự đoán:** Sử dụng ma trận nhầm lẫn để đánh giá hiệu suất của mô hình.

```
confusionMatrix(data = data_mod1,
                reference = data_test$class)
```

5. **Đánh giá tầm quan trọng của biến:** Vẽ biểu đồ để kiểm tra mức độ quan trọng của các biến trong mô hình.

```
plot(varImp(model_rf_1))
```

3.1.3 Độ chính xác tổng thể

- Độ chính xác tổng thể (**Accuracy**) đạt **82.53%** với khoảng tin cậy 95% CI: (80.38%, 84.54%).
- No Information Rate (NIR) là **50%**. So sánh với NIR, Mô hình phải đạt hiệu suất cao hơn 0.5 để chứng minh khả năng phân loại vượt trội hơn so với việc dự đoán ngẫu nhiên hoặc luôn dự đoán vào một lớp bất kỳ.(**P-value** $<$ **2.2e-16**).
- Kappa = **0.6506**, cho thấy mô hình có hiệu quả khá tốt. Kappa cao hơn 0.6 là một dấu hiệu tốt cho thấy mô hình không chỉ dự đoán ngẫu nhiên.

3.1.4 McNemar's Test

- P-value của McNemar's Test là nhỏ hơn **2.2e-16**, điều này cho thấy các lỗi của mô hình không phân bố ngẫu nhiên.

3.1.5 Nhận xét theo lớp

Bảng dưới đây trình bày các chỉ số chi tiết theo phân lớp:

Chỉ số	**Class 0 (A)**	**Class 1 (B,C,D)**
Sensitivity (%)	89.61	75.45
Specificity (%)	75.45	89.61
Pos Pred Value (%)	78.50	87.89
Neg Pred Value (%)	87.89	78.50
F1-score	0.837	0.812
Balanced Accuracy (%)	82.53	82.53

Bảng 12: Các chỉ số đánh giá theo từng lớp

3.1.6 Đánh giá tổng thể

- 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.

3.2 Logistic Regression

3.2.1 Xử Lý Dữ Liệu

Sử dụng dữ liệu đã được xử lý ở phía trên.

3.2.2 Xây dựng model

1. **Chia tập dữ liệu:** Chia dữ liệu thành tập huấn luyện (80%) và tập kiểm tra (20%) bằng cách sử dụng hàm `CreateDataPartition` của thư viện caret .

```
set.seed(123)
trainIndex <- createDataPartition(data_cleaned_1$class, p = 0.8, list = FALSE)
trainData <- data_cleaned_1[trainIndex, ]
testData <- data_cleaned_1[-trainIndex, ]
```

2. **Huấn luyện mô hình Logistic Regression:**

```
model <- multinom(class ~ body_fat_percent+ grip_force +
                  sit_and_bend_forward_cm +
                  sit_ups_counts +
                  broad_jump_cm + BMI_status + MAP, data = trainData)
```

3. **Dự đoán trên tập kiểm tra:**

```
predictions <- predict(model, newdata = testData)
```

4. **Ma trận nhầm lẫn và độ chính xác:** Sử dụng ma trận nhầm lẫn để đánh giá hiệu suất của mô hình.

```
conf_mat <- confusionMatrix(predictions, testData$class)
accuracy <- conf_mat$overall["Accuracy"]
```

3.2.3 Độ chính xác tổng thể

- Độ chính xác tổng thể (**Accuracy**) đạt **78.88%** với khoảng tin cậy 95% CI: (76.59%, 81.05%).
- No Information Rate (NIR) là **50%**. có thể có sự tập trung vào những yếu tố quan trọng, giúp cải thiện dự đoán. (**P-value** $<$ **2.2e-16**).
- Kappa = **0.5777** cho thấy mô hình đạt ở mức đồng thuận trung bình khá, Điều này có nghĩa là mô hình hoạt động ổn định và có thể sử dụng trong nhiều tình huống nhưng vẫn cần cải thiện.

3.2.4 McNemar's Test

- P-value của McNemar's Test là nhỏ hơn **2.2e-16**, cho thấy sự khác biệt không chỉ là ngẫu nhiên, mà có khả năng là do sự khác biệt thực sự giữa mô hình và nhãn thực tế.

3.2.5 Nhận xét theo lớp

Chỉ số	**Class 0 (A)**	**Class 1 (B,C,D)**
Sensitivity (%)	81.75	76.02
Specificity (%)	76.02	81.75
Pos Pred Value (%)	77.32	81.52
Neg Pred Value (%)	80.64	77.32
Prevalence	50.00	50.00
Detection Rate	40.87	38.01
Detection Prevalence	52.87	47.13
Balanced Accuracy (%)	78.88	78.88

Bảng 13: Các chỉ số đánh giá theo lớp

3.2.6 Đánh giá tổng thể

- 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.

4 Nhận Xét Và Đề Xuất

4.1 Nhận xét

- **Tình hình dữ liệu và phân tích:**
  - Bộ dữ liệu với 13,393 mẫu và các biến phong phú đã cung cấp cái nhìn toàn diện về các yếu tố sức khỏe và hiệu suất tập luyện.
  - Việc xáo trộn dữ liệu và chia làm 2 phân lớp theo biến class giúp làm đơn giản bài toán và dễ xử lý hơn.
  - Các chỉ số thể chất (BMI, lực nắm tay, khả năng nhảy xa) và các yếu tố sức khỏe (huyết áp, phần trăm mỡ cơ thể) cho thấy sự khác biệt rõ rệt giữa các nhóm hiệu suất - độ tuổi và các nhóm hiệu suất - giới tính.
  - Mô hình phân loại hiệu suất dựa trên Random Forest đạt độ chính xác tổng thể là 82.53%, cao hơn Logistic Regression 78.88%, cho thấy sự tốt hơn nhưng không quá vượt trội.
- **Kết quả phân tích:**
  - Nhóm hiệu suất D thường có BMI, huyết áp tâm thu và tâm trương 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.
- **Mô hình dự đoán:**
  - Random Forest so sự ổn định và khả năng phân loại hợp lý giữa các lớp, đặc biệt mô hình cho kết quả **accuracy** và **Kappa** cao, cho thấy một mô hình có tính khả thi cao trong việc ứng dụng thực tế.
  - Logistic Regression đạt được kết quả khá tốt và **accuracy** và **Kappa** ở mức vừa phải, tuy nhiên đồ đồng thuận ở mô hình này so với Random Forest thì kém hơn nên cần phải cải thiện.

4.2 Đề xuất

- **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.