BỘ GIÁO DỤC VÀ ĐÀO TẠO
TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN
KHOA TOÁN-TIN HỌC
BỘ MÔN XỬ LÝ SỐ LIỆU THỐNG KÊ

------ଓ✧ଌ------

BÁO CÁO ĐỒ ÁN

BODY PERFORMANCE DATA

NHÓM 13

Họ và tên	MSSV
Phạm Bá Hoàng Anh	**22280003**
Nguyễn Minh Đạt	**22280009**
Lư Xuân Dương	**22280015**
Nguyễn Đức Hiệp	**22280022**
Lê Trọng Nghĩa	**22280059**

Tháng 01/2025

MỤC LỤC

ĐỀ TÀI

Body performance Data

Hiện nay các phong trào tập thể thao đang ngày một phát triển, thu hút nhiều nhóm tuổi và giới tính. Dữ liệu bodyPerformance.csv chứa thông tin của 13,393 người tham gia tập thể thao tại Hàn Quốc, với 12 biến như sau:

- 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 (phút);
- systolic - huyết áp tâm thu (phút);
- gripForce - lực kẹp;
- sit and bend forward_cm - ngồi và gập người về phía trước;
- sit-ups counts - số lần gập bụng;
- broad jump_cm - nhảy xa (đơn vị: cm);
- class - phân lớp hiệu suất (A: tốt nhất, B, C, D).

Hãy xử lý dữ liệu này để giúp cho các chuyên gia sức khỏe biết được hiệu quả của việc tập thể dục, và các yếu tố ảnh hưởng tới hiệu quả.

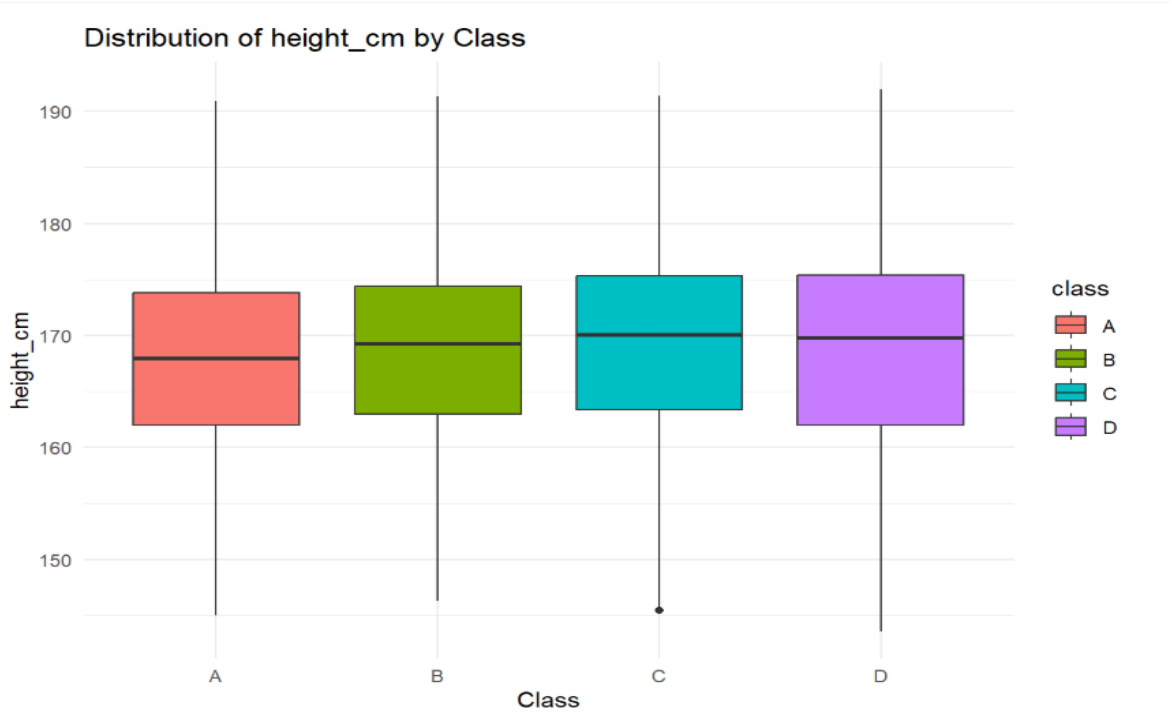
1. ĐỀ XUẤT PHÂN TÍCH VÀ XỬ LÝ DỮ LIỆU

1.1. Bảng đề xuất phân tích

- Các biến định tính: gender và class.
- Các biến định lượng: age, height_cm, weight_kg, grip_force, diastolic, systolic, sit_and_bend_forward_cm, sit_up_counts và broad_jump_cm.

Biểu đồ box plot thể hiện số liệu các biến định lượng so với phân lớp class (sau khi lọc outlier)

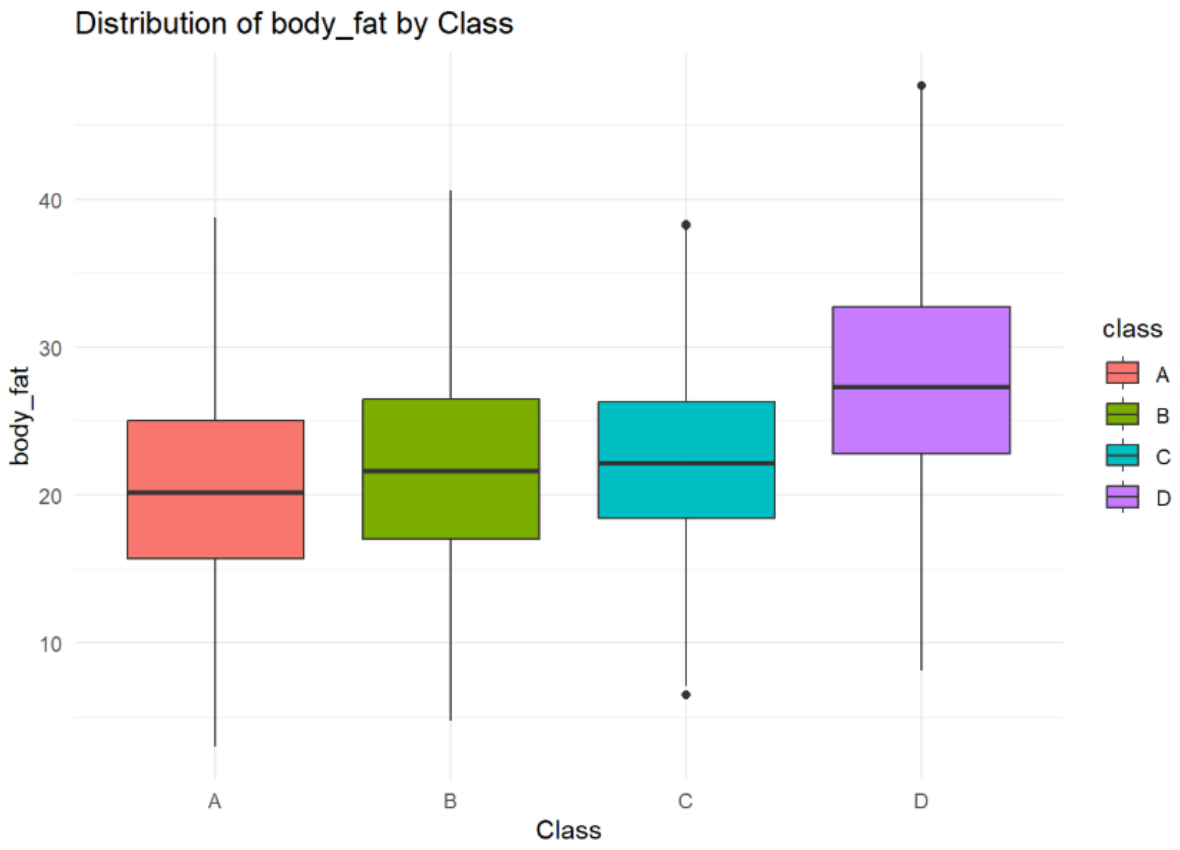
- Biến height_cm



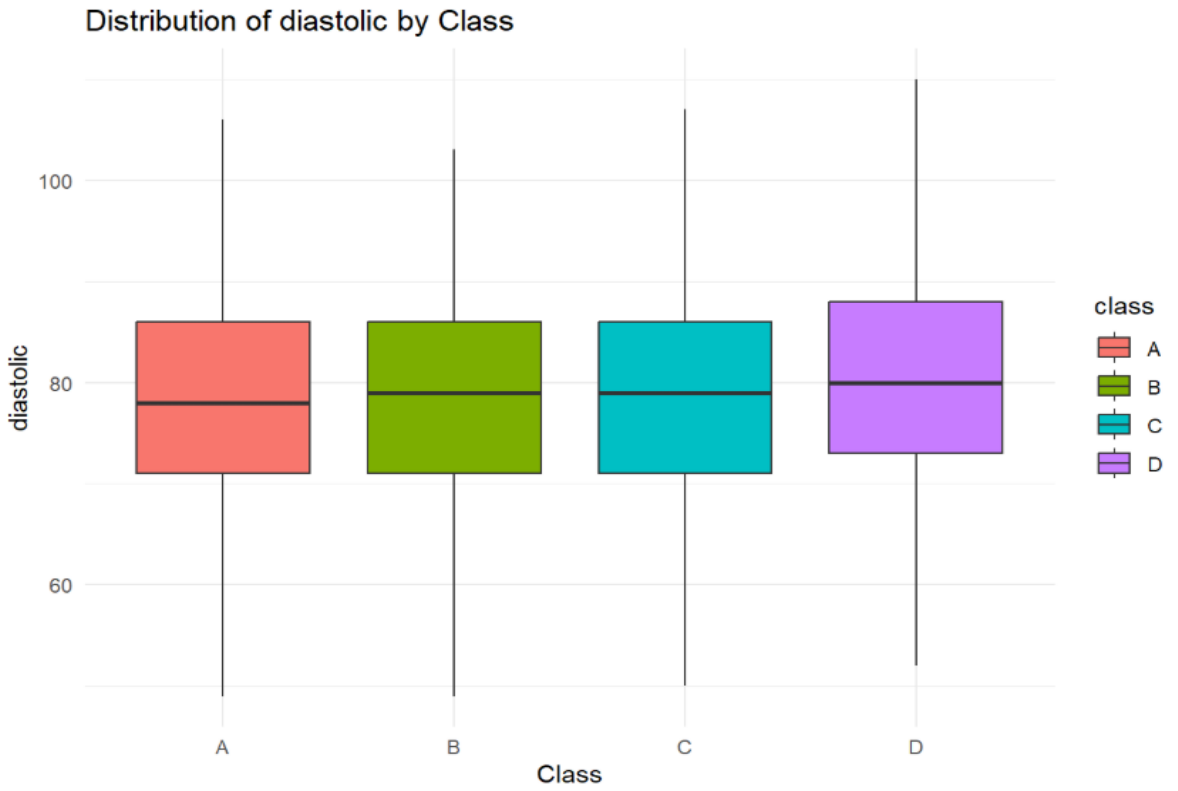
- Biến weight_kg

Distribution of body_fat by Class

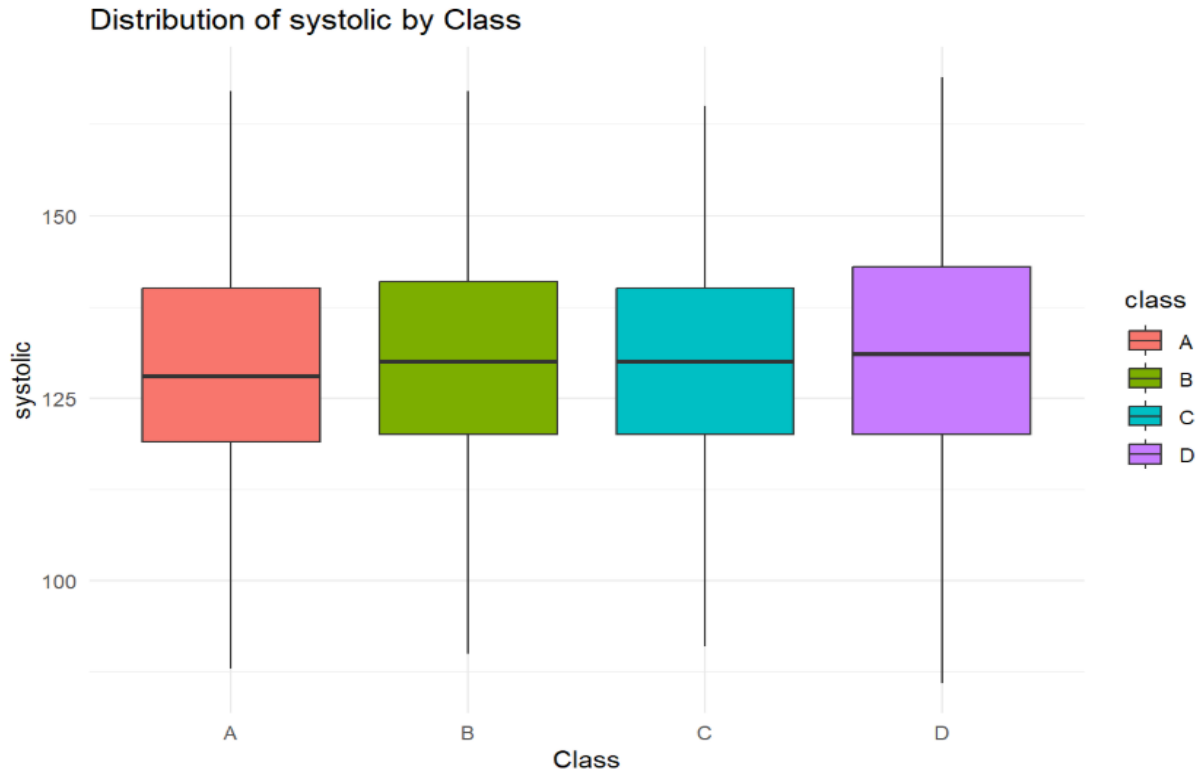
- Biến body_fat:



- Biến diastolic:

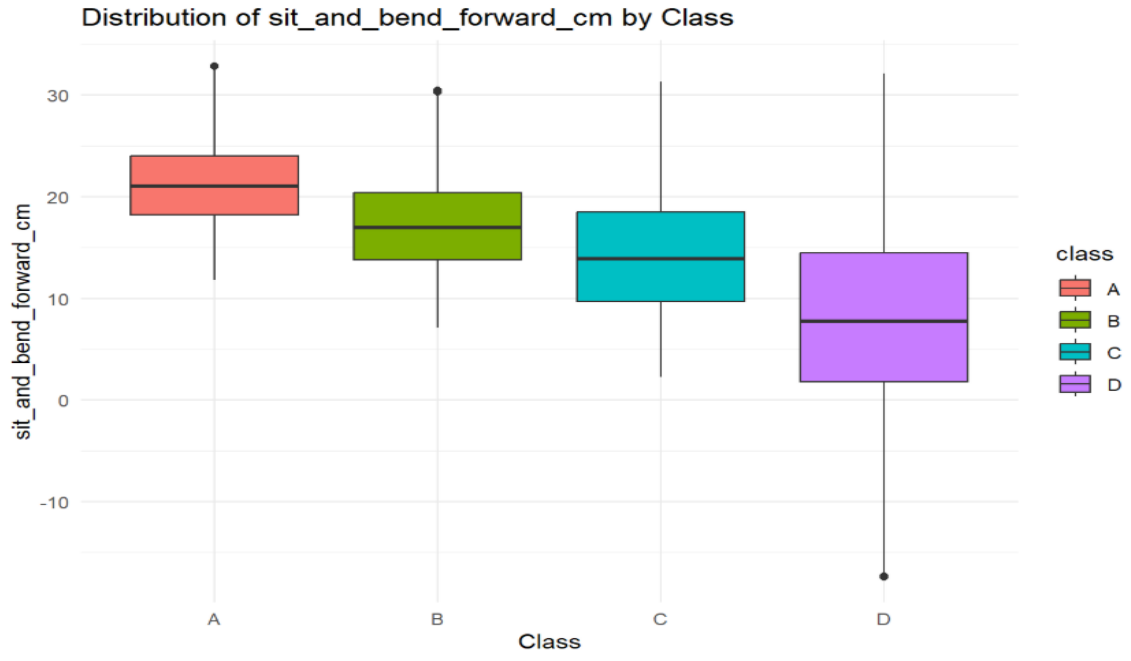


- Biến systolic:

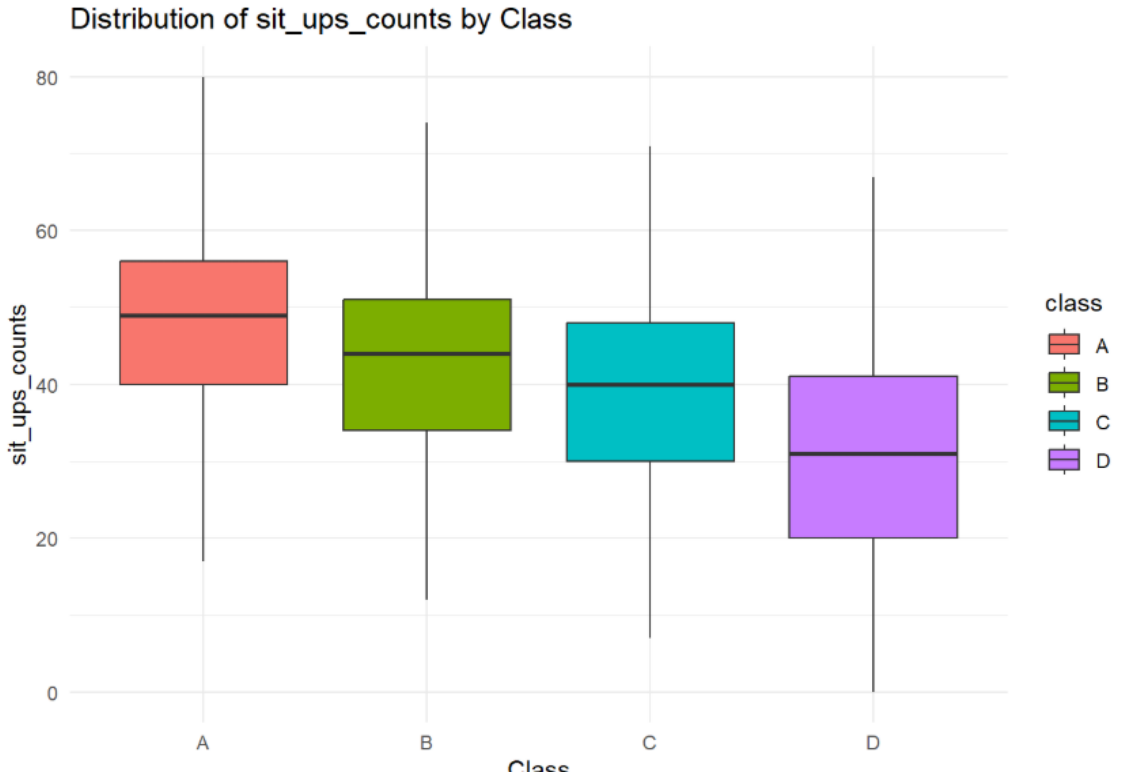


- Biến grip_force:

- Biến sit_and_bend_forward:



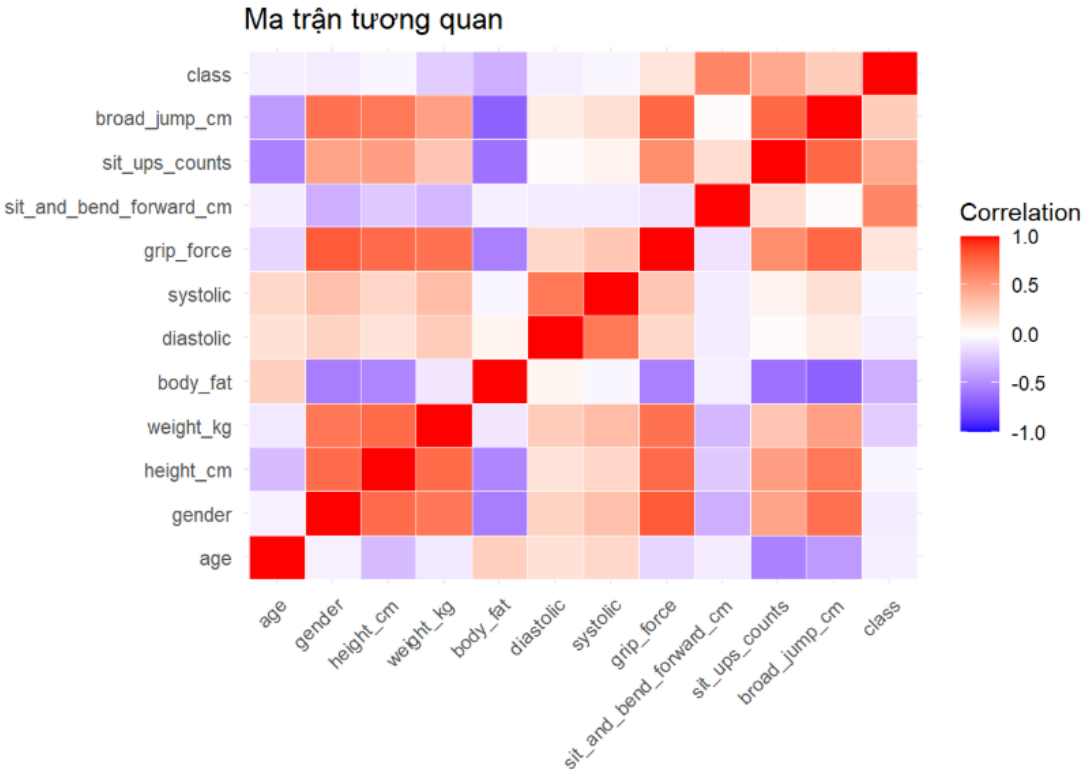
- Biến sit_ups_counts



- Biến broad_jump_cm

1.2. Xử lý số liệu

- Do dữ liệu không có missing data nên không cần xử lý missing data.
- Dữ liệu ban đầu có nhiều outlier ở các nhóm nên cần xử lý outlier cho từng nhóm.
- Việc xử lý outlier cho từng nhóm nhằm giữ được tính chất của dữ liệu hiệu suất của các nhóm (VD body_fat của nhóm D khác với các nhóm khác nên khi lọc outlier theo 4 nhóm thì số liệu của nhóm D bị ảnh hưởng nhiều nhất và gây ra dữ liệu không chính xác).
- Ma trận tương quan để thể hiện sự tương quan giữa các biến với nhau:



2. MỤC TIÊU PHÂN TÍCH

- Mục tiêu: dự đoán hiệu suất của việc tập thể dục và các biến ảnh hưởng đến hiệu suất.
- So sánh sự khác nhau về hiệu suất giữa các nhóm bằng ANOVA test các biến định lượng.

3. PHƯƠNG PHÁP VÀ CHIẾN LƯỢC PHÂN TÍCH

- Dùng ANOVA test để so sánh giữa các nhóm trong biến class.
- Classification: Dùng Multinomial logistic, LDA, QDA và Random Forest để so sánh model tìm một model tốt nhất.

4. PHÂN TÍCH ANOVA TEST

Với biểu đồ trong phần 1.1. ta thấy các biến height_cm, systolic, diastolic không có sự khác biệt đáng kể giữa các nhóm. Tuy nhiên, kết quả cho ra tất cả các biến đều có p-value <0.05 do đó luôn có ít nhất một nhóm khác biệt với các nhóm còn lại trong biến class.

5. XÂY DỰNG MÔ HÌNH VÀ PHÂN TÍCH

Biến phân loại mục tiêu là biến class và các biến x feature là tất cả 11 biến mà dữ liệu đã cho sẵn. Huấn luyện mô hình trên tập train và kiểm tra trên tập test.

5.1. Multinominal Logistic

Biến phân loại mục tiêu là biến class và các biến x feature là tất cả 11 biến mà dữ liệu đã cho sẵn.

- Huấn luyện mô hình trên tập train và kiểm tra trên tập test.

```
[1] "Accuracy: 0.619857795835449"
          Actual
Predicted   A   B   C   D
        A 739 205  72  12
        B 259 464 192  60
        C  22 289 506 150
        D   0  39 197 732
```

- Sau đó ta kiểm tra các giá trị p-value của các hệ số của các biến trong mô hình.

```
  (Intercept) age weight_kg    height_cm     body_fat    diastolic   systolic
B           0   0         0 1.366443e-01 8.888073e-01 1.831951e-01 0.27833114
C           0   0         0 5.829557e-03 6.591629e-01 5.254671e-03 0.04935171
D           0   0         0 2.781775e-12 4.778272e-09 7.330586e-05 0.02393082
  grip_force sit_and_bend_forward_cm sit_ups_counts broad_jump_cm genderM
B          0                       0              0  6.661338e-16       0
C          0                       0              0  0.000000e+00       0
D          0                       0              0  0.000000e+00       0
```

Dùng nhóm A làm nhóm tham chiếu cho các nhóm B, C và D so sánh, đồng thời nhìn vào bảng số liệu ta có thể thấy rằng:

- Biến height_cm trong nhóm B có p_val > 0.05.
- Biến body_fat trong nhóm B, C đều có p_val > 0.05
- Biến diastolic trong nhóm B, C có p_val > 0.05.
- Biến systolic trong nhóm B đều có p_val > 0.05.

Qua việc xử lý như trên, có thể thấy biến body_fat và biến diastolic không có ảnh hưởng gì nhiều đến hiệu suất của mô hình phân loại biến class.
Còn các biến còn lại cũng không có ảnh hưởng hoặc sẽ ảnh hưởng ít đến quá trình thống kê và đưa ra dự đoán nên có thể loại bỏ các biến này trong mô hình.
Nhìn chung, các biến này đều liên quan đến nhóm B.

- Sau đó ta loại bỏ các biến đã nhận xét ở trên và chạy lại mô hình

```
# weights:  36 (24 variable)
initial  value 12735.886296
iter  10 value 10000.438877
iter  20 value 9897.648406
iter  30 value 7904.115215
final  value 7903.783114
converged
[1] "Accuracy: 0.618080243778568"
[1] "Confusion Matrix:"
          Actual
Predicted   A   B   C   D
        A 739 202  75  12
        B 261 469 196  55
        C  20 291 507 168
        D   0  35 189 719
```

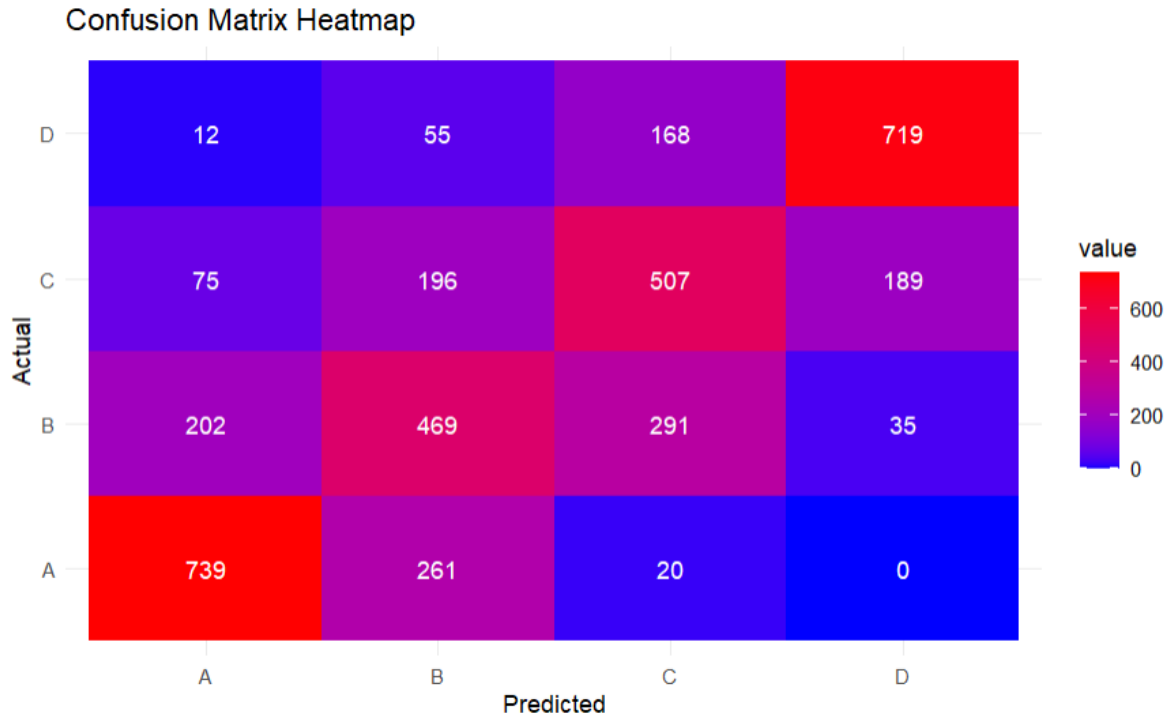
- Ta chạy hàm đánh giá mô hình tổng quát, ta được

```
[1] "Evaluation Metrics:"
$Precision
        A         B         C         D
0.7245098 0.4704112 0.5243020 0.7536688

$Recall
        A         B         C         D
0.7188716 0.4780836 0.5141988 0.7624602

$F1_Scores
        A         B         C         D
0.7216797 0.4742164 0.5192012 0.7580390
```

```
$Macro_Precision
[1] 0.6182229

$Macro_Recall
[1] 0.6184035

$Macro_F1
[1] 0.6182841

$Accuracy
[1] 0.6180802

$Kappa
[1] 0.4906537
```

Sau khi loại bỏ các biến không ảnh hưởng thì hiệu suất là 0.618, so với hiệu suất lúc ban đầu là 0.619, hiệu suất của mô hình không giảm nhiều so với ban đầu. Điều này cho thấy các biến đã bị loại bỏ thật sự không ảnh hưởng đáng kể tới hiệu suất của mô hình.
Mô hình hoạt động tốt với các lớp A và D, với F1-Score lần lượt là 0.721 và 0.758, nhưng hiệu suất còn hạn chế ở các lớp B (F1-Score 0.474) và C (F1-Score 0.519). Độ chính xác tổng thể đạt 61.80%, các chỉ số tổng hợp như Macro Precision (0.618), Macro Recall (0.618), và Macro F1 (0.618) cho thấy mô hình duy trì sự cân đối nhưng chưa thực sự vượt trội. Chỉ số Kappa (0.490) phản ánh mức độ đồng thuận trung bình giữa mô hình và dữ liệu thực tế, tốt hơn dự đoán ngẫu nhiên nhưng vẫn cần cải thiện.

- Biểu diễn ma trận nhầm lẫn qua heatmap



Heatmap của ma trận nhầm lẫn cho thấy mô hình hoạt động tốt nhất trên Class A và Class D, với số lượng phân loại đúng lần lượt là 739 và 719. Tuy nhiên, vẫn xảy ra nhầm lẫn đáng kể, đặc biệt giữa Class A với Class B (202 mẫu nhầm) và giữa Class C với Class B (291 mẫu nhầm). Class B và Class C có sự nhầm lẫn lẫn nhau khá nhiều, cho thấy các đặc trưng phân biệt giữa hai lớp này chưa rõ ràng. Điều này có thể là nguyên nhân dẫn đến sự chênh lệch trong hiệu suất phân loại giữa các lớp.

5.2. Mô hình LDA

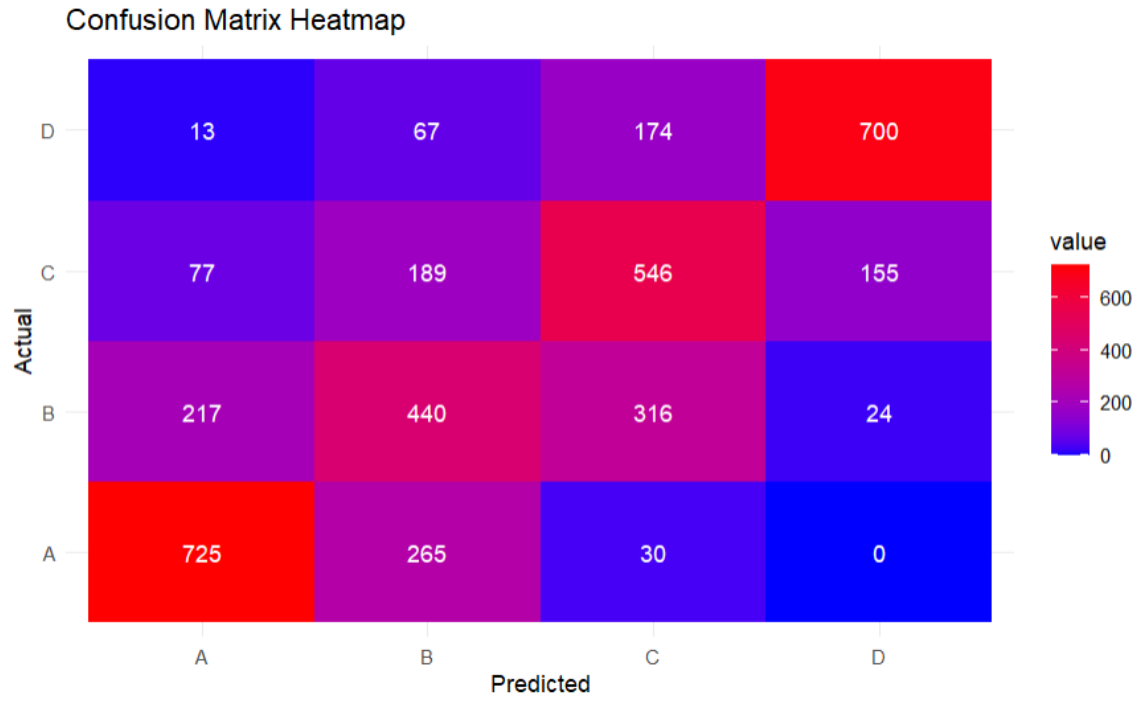
- Lấy tập train và tập test như ở mô hình trên và kết quả sau khi huấn luyện trên tập train.

```
Call:
lda(class ~ age + weight_kg + height_cm + body_fat + systolic +
    diastolic + grip_force + sit_and_bend_forward_cm + sit_ups_counts +
    broad_jump_cm + gender, data = train_data, maxit = 1500)

Prior probabilities of groups:
        A         B         C         D
0.2485033 0.2490476 0.2520954 0.2503538

Group means:
       age weight_kg height_cm body_fat systolic diastolic grip_force sit_and_bend_forward_cm sit_ups_counts broad_jump_cm   genderM
A 35.14455  64.44451  167.9035 20.45569  129.463  77.96163   38.69708              21.303316       48.14717      203.2501 0.5606658
B 36.89729  66.52468  168.6502 21.92157  130.281  78.44580   38.01610              17.382740       42.73837      195.6967 0.6464161
C 36.52763  66.74383  169.2620 22.47904  129.674  78.47064   36.66625              14.323402       38.86010      189.4685 0.6727116
D 37.98783  71.69750  168.6080 27.64033  130.830  80.03883   34.69743               7.924117       30.01478      174.1665 0.6565217

Coefficients of linear discriminants:
                            LD1         LD2         LD3
age                   -0.046551687  0.015719167  0.034160234
weight_kg              0.046848417  0.092133536 -0.003805877
height_cm             -0.007376237 -0.110431258 -0.026165844
body_fat               0.025086802  0.032983744  0.087662824
systolic              -0.003026230 -0.001282414  0.014818755
diastolic              0.006385492  0.004931975 -0.035825151
grip_force            -0.044335238  0.030505319  0.016434001
sit_and_bend_forward_cm -0.103802771 -0.031192913  0.059784637
sit_ups_counts        -0.071621443  0.025692202  0.001128416
broad_jump_cm         -0.006388751  0.016436049 -0.002191084
genderM                1.297979683 -2.540723971  2.422138229

Proportion of trace:
   LD1    LD2    LD3
0.9799 0.0186 0.0015
```

- Ta chạy hàm đánh giá mô hình tổng quát.

```
[1] "Evaluation Metrics:"
$Precision
        A         B         C         D
0.7107843 0.4413240 0.5646329 0.7337526

$Recall
        A         B         C         D
0.7025194 0.4578564 0.5121951 0.7963595

$F1_Scores
        A         B         C         D
0.7066277 0.4494382 0.5371372 0.7637752

$Macro_Precision
[1] 0.6126234

$Macro_Recall
[1] 0.6172326

$Macro_F1
[1] 0.6142446

$Accuracy
[1] 0.6122397

$Kappa
[1] 0.4828447
```

- Heatmap cho ma trận nhầm lẫn.

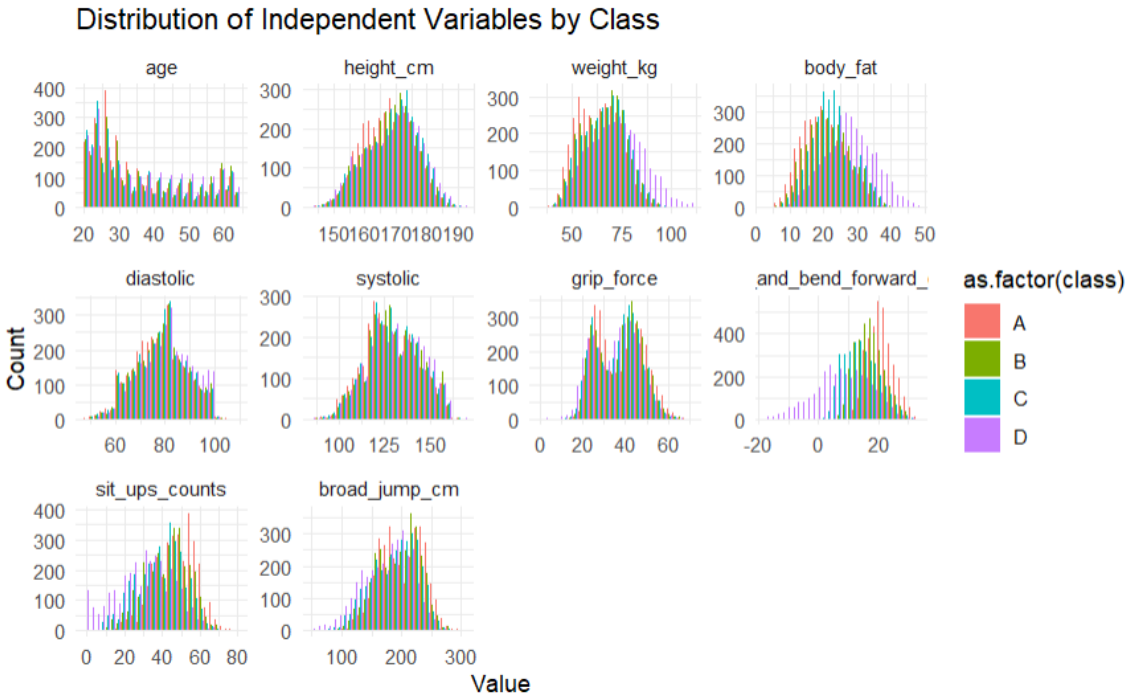


- Kiểm tra tính đồng nhất của ma trận hiệp phương sai (Box's M Test).

```
        Box's M-test for Homogeneity of Covariance Matrices

data:  as.matrix(x)
Chi-Sq (approx.) = 5390.6, df = 198, p-value < 2.2e-16
```

Do p_vals < 0.05 nên không thể cho rằng giả định các nhóm đồng nhất phương sai với nhau. Việc không thể chắc chắn rằng các nhóm đồng nhất phương sai với nhau thì có thể gây ra ảnh hưởng về hiệu quả dự đoán của mô hình LDA.

- Vẽ phân phối của các biến trong dữ liệu.



Qua việc kiểm tra phân phối chuẩn của từng nhóm dữ liệu loại A, B, C và D cho thấy rằng biến age, grip_force và broad_jump_cm không tuân theo phân phối chuẩn nên cũng ảnh hưởng đến việc phân loại của mô hình.

Do việc sử dụng LDA có thể không đảm bảo độ chính xác, nên việc chuyển sang QDA là hợp lý hơn, vì QDA không yêu cầu giả định về sự đồng nhất phương sai và phân phối của dữ liệu.

5.3. Mô hình QDA

- Xây dựng mô hình và chạy trên tập test.

```
          Actual
Predicted   A   B   C   D
        A 773 185  65  10
        B 234 564 206  61
        C   9 223 620 160
        D   4  25  76 723
Accuracy: 0.6805485
```

```
Call:
qda(class ~ age + weight_kg + height_cm + body_fat + diastolic +
    systolic + grip_force + sit_and_bend_forward_cm + sit_ups_counts +
    broad_jump_cm + gender, data = train_data)

Prior probabilities of groups:
        A         B         C         D
0.2485033 0.2490476 0.2520954 0.2503538

Group means:
       age weight_kg height_cm body_fat diastolic systolic grip_force
A 35.14455  64.44451  167.9035 20.45569  77.96163  129.463   38.69708
B 36.89729  66.52468  168.6502 21.92157  78.44580  130.281   38.01610
C 36.52763  66.74383  169.2620 22.47904  78.47064  129.674   36.66625
D 37.98783  71.69750  168.6080 27.64033  80.03883  130.830   34.69743
  sit_and_bend_forward_cm sit_ups_counts broad_jump_cm   genderM
A               21.303316       48.14717      203.2501 0.5606658
B               17.382740       42.73837      195.6967 0.6464161
C               14.323402       38.86010      189.4685 0.6727116
D                7.924117       30.01478      174.1665 0.6565217
```

Trong 4 nhóm, các biến age, height_cm, systolic, diastolic, và gender không có sự chênh lệch đáng kể (không quá 5%). Vì vậy, các biến này được loại bỏ để thực hiện lại mô hình.

- Chạy lại mô hình sau khi bỏ đi 5 biến nêu trên.

```
          Actual
Predicted   A   B   C   D
        A 747 254 102  17
        B 223 410 208  59
        C  44 287 539 176
        D   6  46 118 702
Accuracy: 0.6089385
```

Kết quả chạy lại mô hình cho thấy việc loại bỏ các biến trên không hoàn toàn chính xác:

- Biến age: Mặc dù không có sự chênh lệch trong 4 nhóm A, B, C và D và có tương quan thấp với biến class, nhưng đồ thị phân phối cho thấy biến này không tuân theo phân phối chuẩn. Hơn nữa, trong đồ thị tương quan, age có mối quan hệ mạnh mẽ với hai biến sit_ups_counts và broad_jump_cm. Do đó, biến age đóng vai trò bổ trợ và không nên loại bỏ.
- Biến height_cm: Phân phối chuẩn nhưng có ảnh hưởng mạnh đến các biến khác trong ma trận tương quan và ít ảnh hưởng đến biến class. Điều này cho thấy khả năng xảy ra đa cộng tuyến, nên có thể loại bỏ biến này.
- Biến systolic và diastolic: Cả hai không có tương quan đáng kể với các biến khác và không có sự chênh lệch trung bình trong 4 nhóm A, B, C và D. Vì vậy, việc loại bỏ hai biến này là hợp lý.
- Biến gender: Là biến định tính phân biệt nam và nữ, cần giữ lại do biểu đồ heatmap cho thấy gender có ảnh hưởng đến các biến khác, tương tự như vai trò của biến age.

- Thử nghiệm các biến độc lập khác với QDA.

```
          Actual
Predicted   A   B   C   D
        A 768 184  63  12
        B 241 565 217  58
        C   7 225 606 159
        D   4  23  81 725
Accuracy: 0.6764855
```

Kết quả cho thấy khi sử dụng tất cả các biến, độ chính xác đạt 0,680. Tuy nhiên, sau khi loại bỏ ba biến height_cm, diastolic, và systolic, độ chính xác giảm nhẹ xuống 0,676. Điều này chứng tỏ các biến này không ảnh hưởng nhiều đến hiệu suất của mô hình.

Ngoài ra, khi xét biến body_fat trong mô hình Multinomial Logistic, kết quả cho thấy biến này không có vai trò quan trọng trong việc phân loại các giá trị của biến class. Do đó, việc loại bỏ biến body_fat cũng có thể được cân nhắc để đơn giản hóa mô hình.

- Chạy lại mô hình sau khi loại bỏ biến body_fat.

```
          Actual
Predicted   A   B   C   D
        A 772 183  64   9
        B 239 559 211  47
        C   6 232 602 181
        D   3  23  90 717
Accuracy: 0.6729304
```

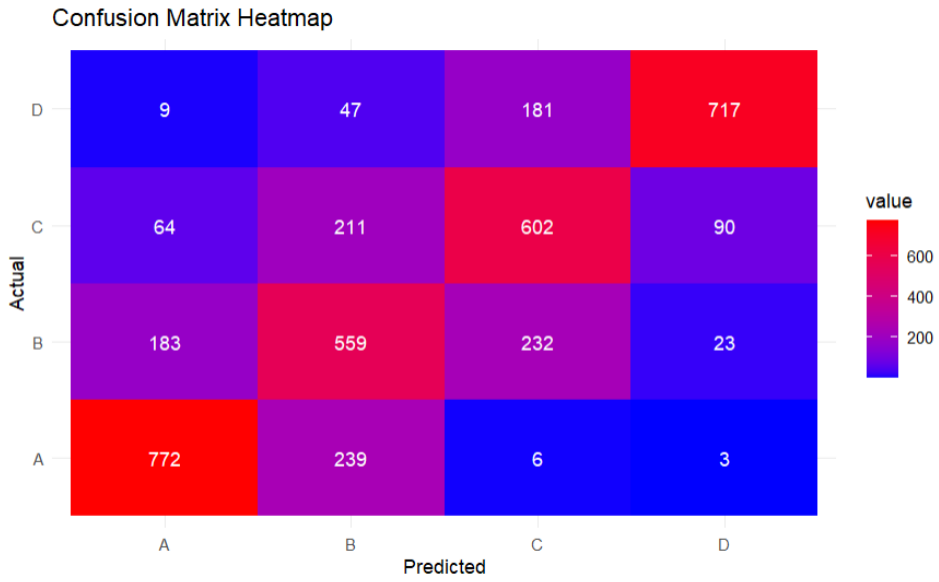
Hiệu suất của mô hình chỉ giảm 0,4%, điều này cho thấy biến body_fat có thể gặp vấn đề đa cộng tuyến giống như biến height_cm và thực sự không ảnh hưởng đáng kể đến mô hình. Vì vậy, việc loại bỏ biến body_fat là hợp lý.

- Hàm đánh giá đa lớp.

```
[1] "Evaluation Metrics:"
$Precision
        A         B         C         D
0.7568627 0.5606820 0.6225440 0.7515723

$Recall
        A         B         C         D
0.7509728 0.5293561 0.5896180 0.8607443

$F1_Scores
        A         B         C         D
0.7539062 0.5445689 0.6056338 0.8024622
```

```
$Macro_Precision
[1] 0.6729153

$Macro_Recall
[1] 0.6826728

$Macro_F1
[1] 0.6766428

$Accuracy
[1] 0.6729304

$Kappa
[1] 0.5636664
```

- Heatmap cho ma trận nhầm lẫn.



5.4. Random Forest

Qua việc đọc slide và tìm hiểu tài liệu, nhóm em thấy rằng phương pháp phân nhóm khi sử dụng model "RandomForest" phù hợp với bài toán (có biến định tính lẫn biến định lương, nhiều biến đầu vào,...).

> *Khái niệm: Random Forest là một tập hợp các cây quyết định (decision trees) được xây dựng theo nguyên lý bagging (Bootstrap Aggregating). Mỗi cây trong rừng được huấn luyện trên một mẫu con khác nhau của tập dữ liệu và các đặc trưng cũng được chọn ngẫu nhiên tại mỗi điểm phân chia (split) trong cây.*

- Áp dụng với tất cả các biến đầu vào.

```
          Actual
Predicted   A   B   C   D
        A 859 202  81  11
        B 150 634 176  50
        C   8 120 643 108
        D   3  41  67 785
[1] "Accuracy: 0.741747079735907"
```

```
[1] "Evaluation Metrics:"
$Precision
        A         B         C         D
0.8421569 0.6359077 0.6649431 0.8228512

$Recall
        A         B         C         D
0.7450130 0.6277228 0.7315131 0.8761161

$F1_Scores
        A         B         C         D
0.7906121 0.6317887 0.6966414 0.8486486
```

```
$Macro_Precision
[1] 0.7414647

$Macro_Recall
[1] 0.7450912

$Macro_F1
[1] 0.7419227

$Accuracy
[1] 0.7417471

$Kappa
[1] 0.6553414
```

Mô hình có khả năng dự đoán chính xác 74.17% ở mức khá và có thể chấp nhận được. Nhìn vào Precision Recall, F1 và Accuracy các chỉ số khá là tương đương nhau chứng tỏ hiệu suất mô hình khá ổn định. Giá trị Kappa là 0.655 tuy không cao nhưng cũng có thể chấp nhận được đối với mô hình khi phân loại trong bài toán trên. Phân tích từng lớp cho thấy khả năng phân loại hiệu suất A và D khá tốt nhưng ở lớp B và C thì giá trị khá thấp (có thể do nhiều nguyên nhân như dữ liệu dành cho lớp B và C chưa đủ tốt).

- Phân tích mức độ quan trọng của từng biến.

	A	B	C	D	MeanDecreaseAccuracy	MeanDecreaseGini
age	62.55917	54.253264	48.903472	8.598975	73.15064	507.00245
gender	35.49852	28.909149	24.258033	11.838716	39.07735	96.44737
height_cm	30.63456	21.570471	19.577899	18.226819	43.02584	435.54180
weight_kg	58.33983	31.281657	45.188508	39.997104	77.52990	609.18772
body_fat	42.73943	26.085809	63.431190	53.168123	74.88378	678.29054
diastolic	10.43764	2.185702	6.348282	6.037018	12.59468	337.83095
systolic	12.61870	5.995448	8.877820	5.501251	17.17488	354.04253
grip_force	68.11862	42.481408	23.395533	11.067115	74.54972	553.58873
sit_and_bend_forward_cm	248.76446	104.620556	105.636503	125.967213	200.76782	1718.41726
sit_ups_counts	164.96260	84.765636	67.347935	67.733741	131.09241	958.16350
broad_jump_cm	68.01216	27.198388	16.653955	20.608215	65.01324	519.23668
age_group	32.71239	31.594038	27.427462	3.560962	37.55080	120.53625

Dựa trên kết quả kiểm tra mức độ quan trọng của các biến, các biến "diastolic", "systolic", "height_cm" và "gender" đều không có tác động đáng kể hoặc ảnh hưởng rất ít đến hiệu suất của mô hình Random Forest và các mô hình khác trước đó. Do đó, quyết định được đưa ra là loại bỏ 4 biến này nhằm giảm sự phức tạp của mô hình, qua đó cải thiện hiệu quả tính toán và tăng khả năng tổng quát hóa trên dữ liệu mới.

- Chạy lại khi không có 4 biến trên.

```
          Actual
Predicted   A   B   C   D
        A 860 203  81  11
        B 149 632 176  50
        C   8 120 643 108
        D   3  42  67 785
[1] "Accuracy: 0.741493143727781"


[1] "Evaluation Metrics:"
$Precision
        A         B         C         D
0.8431373 0.6339017 0.6649431 0.8228512

$Recall
        A         B         C         D
0.7445887 0.6276068 0.7315131 0.8751394

$F1_Scores
        A         B         C         D
0.7908046 0.6307385 0.6966414 0.8481902
```

```
$Macro_Precision
[1] 0.7412083

$Macro_Recall
[1] 0.744712

$Macro_F1
[1] 0.7415937

$Accuracy
[1] 0.7414931

$Kappa
[1] 0.6550024
```

Dựa vào các hệ số đánh giá mô hình dự đoán thì kết quả sau khi bỏ 4 biến đầu vào không ảnh hưởng đến mô hình. Vì vậy việc bỏ đi các biến trên là hoàn toàn phù hợp.

6. KẾT LUẬN

Qua việc đánh giá hiệu suất của mô hình và lựa chọn ra các biến có sự ảnh hưởng mạnh mẽ tới biến phân loại class thì ta có thể kết luận rằng:

- 2 biến diastolic và systolic thật sự không ảnh hưởng quá nhiều đến hiệu suất của mô hình, 3 biến height_cm, body_fat và gender có ảnh hưởng một phần nhỏ nhưng không đáng kể nên có thể loại bỏ. Do đó, 5 biến này là các biến không ảnh hưởng tới việc phân loại các đối tượng trong class.
- Các biến còn lại như biến sit_ups_counts, broad_jump_cm và age là những biến có ảnh hưởng mạnh mẽ nhất tới việc phân loại.

Random Forest nổi bật là lựa chọn tối ưu trong 4 mô hình phân loại đã sử dụng. Ưu điểm nổi bật của mô hình này thể hiện ở khả năng xử lý đồng thời nhiều loại biến, bao gồm cả biến định tính và định lượng. Đặc biệt, Random Forest có cơ chế tích hợp giúp hạn chế hiện tượng overfitting, đồng thời không đòi hỏi các điều kiện nghiêm ngặt về tính độc lập giữa các biến đầu vào hay yêu cầu tuân theo phân phối chuẩn. Những đặc điểm này khiến Random Forest trở thành công cụ phân loại linh hoạt và đáng tin cậy cho đồ án này.

Do 3 biến sit_ups_counts, broad_jump_cm và age là những biến có ảnh hưởng mạnh mẽ đến mô hình phân loại hiệu suất, nhưng do biến age là biến không thay đổi được. Vì vậy, chỉ có thể dựa vào 2 biến sit_ups_counts và broad_jump_cm để nâng cao hiệu suất của người tập thể dục. Đối với sit_ups_counts, cần thực hiện các bài tập phát triển sức mạnh, bao gồm các biến thể của động tác gập bụng và plank, đồng thời tăng dần cường độ. Song song đó, để nâng cao chỉ số broad_jump_cm, cần tập trung vào các bài tập phát triển sức mạnh như plyometric, squat jump và box jump. Quá trình tập luyện cần được đánh giá định kỳ, có sự điều chỉnh phù hợp dựa trên tiến độ cá nhân, đồng thời đảm bảo chế độ nghỉ ngơi và dinh dưỡng để tối ưu hiệu quả và phòng tránh chấn thương.