ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH

TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN

Môn học: Xử lý số liệu thống kê

Body Performance

Sinh viên thực hiện:

1. Lê Nho Hãn – 22110054
2. Tạ Quang Duy – 22110047
3. Lê Hoàng An – 22110002
4. Phạm Hoàng Dũng – 22110043
5. Lê Trần Gia Bảo – 21110246
6. Nguyễn Lê Khánh Duy – 22110046

Giảng viên môn học:

Tô Đức Khánh

Mục lục

1 Giới thiệu

1.1 Bài toán và mục tiêu

Trong những năm gần đây, các phong trào tập luyện thể thao đã trở nên phổ biến, thu hút sự tham gia của nhiều nhóm tuổi và giới tính. Tập luyện không chỉ giúp cải thiện sức khỏe thể chất mà còn nâng cao chất lượng cuộc sống. Tuy nhiên, hiệu quả của việc tập luyện có thể khác nhau ở từng người, phụ thuộc vào các yếu tố như độ tuổi, giới tính, chỉ số cơ thể và khả năng thể chất.

Báo cáo này phân tích dữ liệu từ 13,393 người tham gia tập luyện tại Hàn Quốc, được lưu trữ trong tập dữ liệu bodyPerformance.csv, với mục tiêu:

Đánh giá mối quan hệ giữa các yếu tố thể chất (tuổi, giới tính, chiều cao, cân nặng, v.v.) và hiệu suất tập luyện.

Xác định các yếu tố chính ảnh hưởng đến phân lớp hiệu suất (A, B, C, D). Đưa ra các khuyến nghị dựa trên kết quả phân tích, nhằm tối ưu hóa hiệu quả tập luyện cho từng nhóm đối tượng.

Phương pháp phân tích bao gồm mô tả dữ liệu, phân tích mối quan hệ giữa các biến, và xây dựng các mô hình dự đoán. Kết quả sẽ cung cấp những thông tin giá trị cho các chuyên gia sức khỏe, giúp họ thiết kế các chương trình tập luyện phù hợp, mang lại hiệu quả tối ưu cho từng cá nhân.

1.2 Phương pháp và chiến lược phân tích

Mô tả dữ liệu:
Sử dụng các bảng tổng hợp, biểu đồ hộp (boxplot), biểu đồ phân tán (scatter plot), histogram.
Phân tích tương quan:
Tính hệ số tương quan giữa các biến liên tục. Sử dụng biểu đồ heatmap để trực quan hóa.
Phân tích so sánh nhóm:
Sử dụng kiểm định t-test hoặc ANOVA để so sánh các nhóm theo phân lớp hiệu suất.
Xây dựng mô hình:

Hồi quy logistic để dự đoán xác suất đạt hiệu suất cao. Random forest để đánh giá mức độ quan trọng của các yếu tố.

2 Xử lí dữ liệu

2.1 Làm sạch và chuyển đổi dữ liệu

Xem các hàng đầu để biết cơ bản về dữ liệu

	age	gender	height_cm	weight_kg	body fat_%	diastolic	systolic	gripForce	sit and bend forward_cm	sit-ups counts	broad jump_cm	class
0	27.0	M	172.3	75.24	21.3	80.0	130.0	54.9	18.4	60.0	217.0	C
1	25.0	M	165.0	55.80	15.7	77.0	126.0	36.4	16.3	53.0	229.0	A
2	31.0	M	179.6	78.00	20.1	92.0	152.0	44.8	12.0	49.0	181.0	C
3	32.0	M	174.5	71.10	18.4	76.0	147.0	41.4	15.2	53.0	219.0	B
4	28.0	M	173.8	67.70	17.1	70.0	127.0	43.5	27.1	45.0	217.0	B

Kiểm tra missing value

```
##                   age                gender             height_cm
##                     0                     0                     0
##             weight_kg          body.fat_.              diastolic
##                     0                     0                     0
##              systolic            gripForce sit.and.bend.forward_cm
##                     0                     0                     0
##         sit.ups.counts         broad.jump_cm                 class
##                     0                     0                     0
```

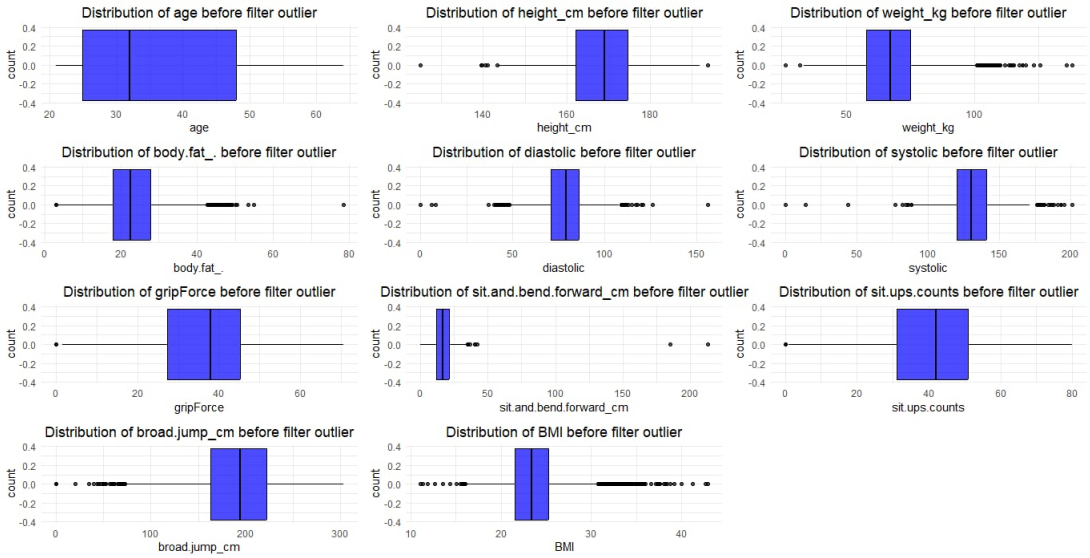
Dữ liệu không có missing value

Xem bảng tổng hợp dữ liệu

```
> summary(data)
      age           gender            height_cm       weight_kg        body.fat_.
 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       gripForce     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.   :303.0
```

Ta thấy cột sit and bend forward_cm có giá trị lớn bất thường và giá trị âm. Ta cần loại bỏ các giá trị này

Vẽ boxplot xem các giá trị ngoại lai

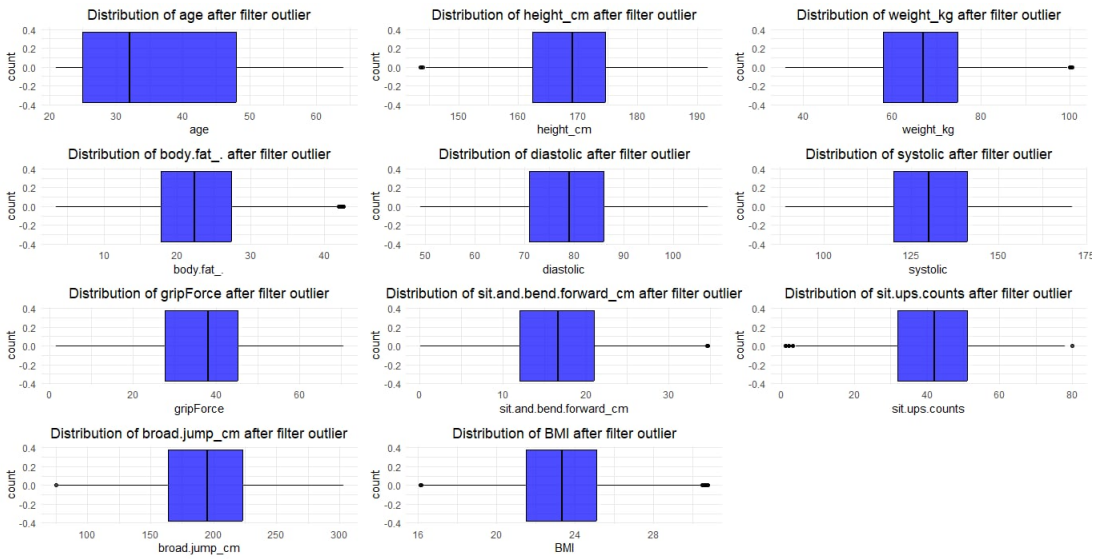


Các điểm dữ liệu ngoại lai bất thường:

- Cột body_fat% : % mỡ cơ thể $> 70\%$ là không hợp lí
- Cột Systolic: systolic < 50 là không hợp lí
- Cột Diastolic: diastolic < 20 và > 150 là không hợp lí

Cần loại bỏ các giá trị ngoại lai này

Sau khi loại bỏ:



2.2 Thêm các đặc trưng cần thiết

Để bài phân tích trở nên đầy đủ. Chúng tôi tính toán và thêm vào dữ liệu các đặc trưng sau:

• BMI: chỉ số BMI được tính bằng công thức: BMI = $\frac{\text{Cân nặng (kg)}}{\text{Chiều cao}^2\text{(m)}}$ Đây là chỉ số được sử dụng phổ biến trong việc đánh giá sức khỏe

• bmi_category: Dựa vào BMI ta chia thành 4 nhóm:

BMI < 18.5: Nhóm "Underweight" (Thiếu cân).

18.5 <= BMI < 25: Nhóm "Normal" (Bình thường).

25 <= BMI < 30: Nhóm "Overweight" (Thừa cân).

BMI >= 30: Nhóm "Obese" (Béo phì).

• systolic_category: Dựa vào Systolic chia làm 3 nhóm:

Systolic >= 120 và <= 130: Nhóm "Normal".

Systolic < 120: Nhóm "Low".

Systolic > 130: Nhóm "High".

• diastolic_category : Dựa vào Diastolic chia làm 3 nhóm

Diastolic >= 70 và <= 90: Nhóm "Normal".

Diastolic < 70: Nhóm "Low".

Diastolic > 90: Nhóm "High".

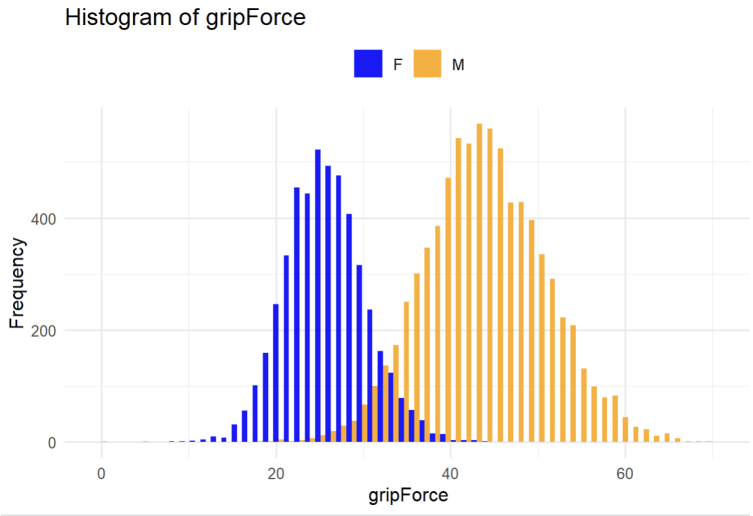
Dữ liệu sau khi thêm các đặc trưng:

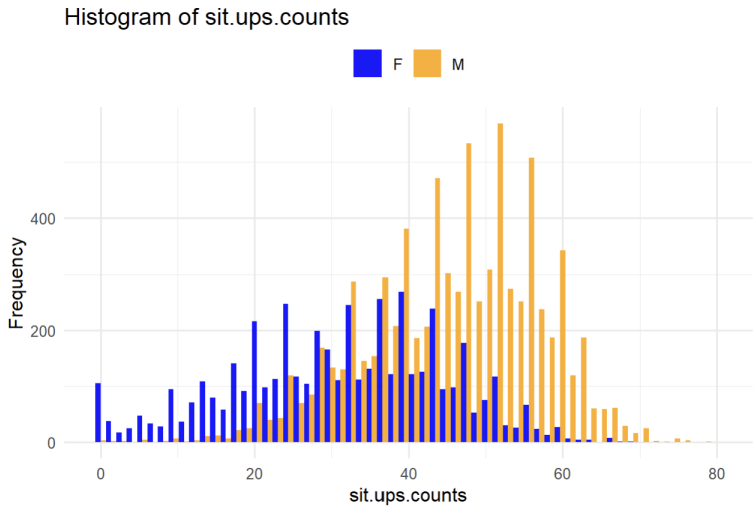
```
  age gender height_cm weight_kg body.fat_. diastolic systolic gripForce
1  27      M     172.3     75.24       21.3        80      130      54.9
2  25      M     165.0     55.80       15.7        77      126      36.4
3  31      M     179.6     78.00       20.1        92      152      44.8
4  32      M     174.5     71.10       18.4        76      147      41.4
5  28      M     173.8     67.70       17.1        70      127      43.5
  sit.and.bend.forward_cm sit.ups.counts broad.jump_cm class group_age  BMIndex
1                    18.4             60           217     C     Youth 25.34418
2                    16.3             53           229     A     Youth 20.49587
3                    12.0             49           181     C     Youth 24.18143
4                    15.2             53           219     B     Youth 23.34956
5                    27.1             45           217     B     Youth 22.41244
   situation _systolic _diastolic
1 Overweight    Normal     Normal
2     Normal    Normal     Normal
3     Normal      High       High
4     Normal      High     Normal
5     Normal    Normal     Normal
```

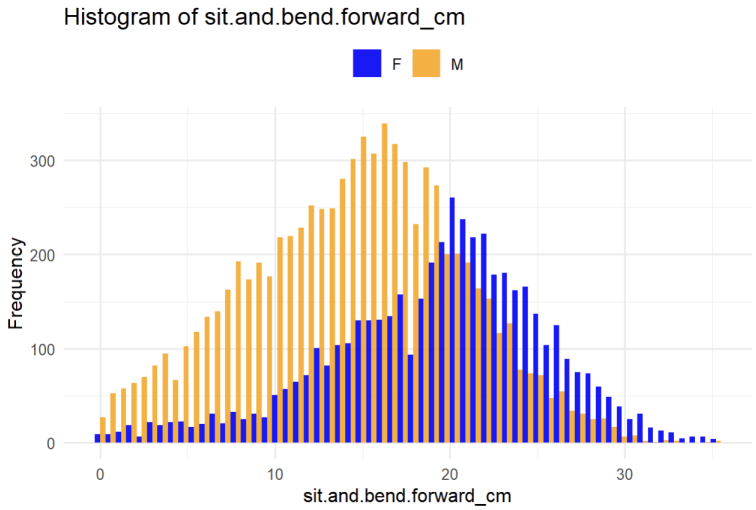
3 Khám phá dữ liệu

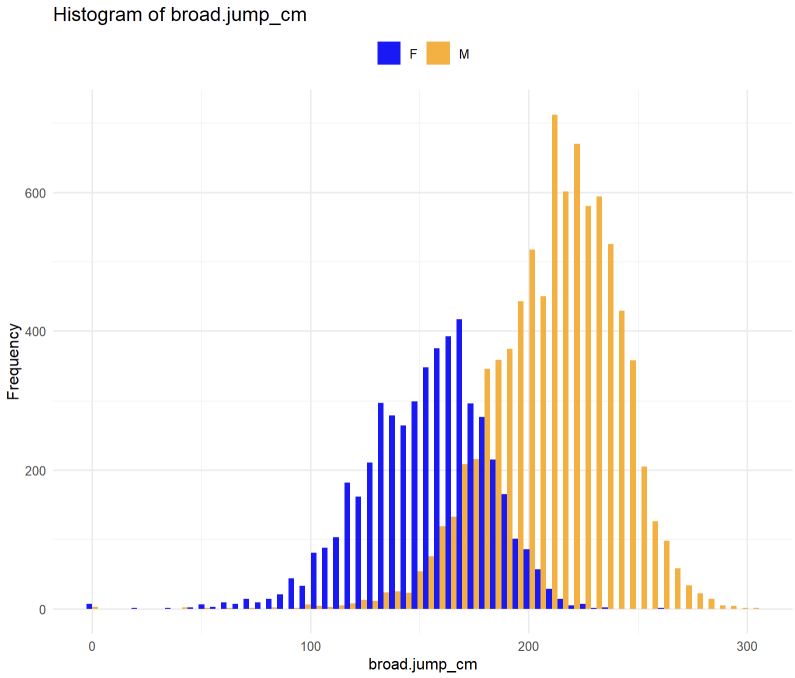
3.1 Giới tính và tuổi với hiệu suất tập thể dục

Đầu tiên ta cùng xem giới tính và độ tuổi có ảnh hưởng đến hiệu suất tập luyện các môn thể dục như thế nào







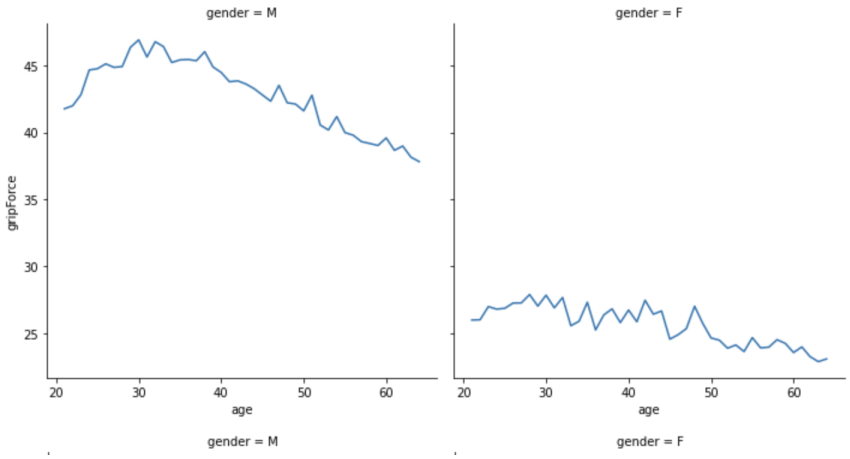


Thông qua các histogram ta thấy có sự khác biệt giữa nam và nữ:

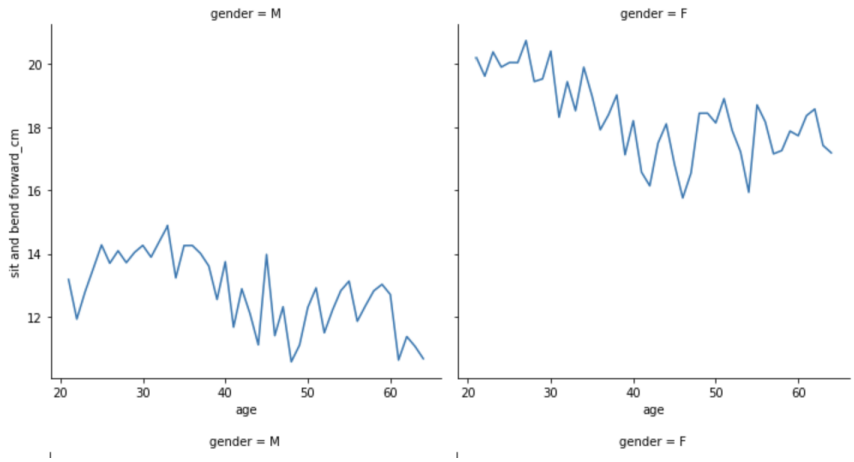
• Lực kẹp, số lần gập bụng và nhảy xa của nam lớn hơn nhiều so với nữ

• số lần ngồi và gập người về trước của nữ lớp hơn nam tương đối- có vẻ đây là bài tập yêu thích của giới tính nữ

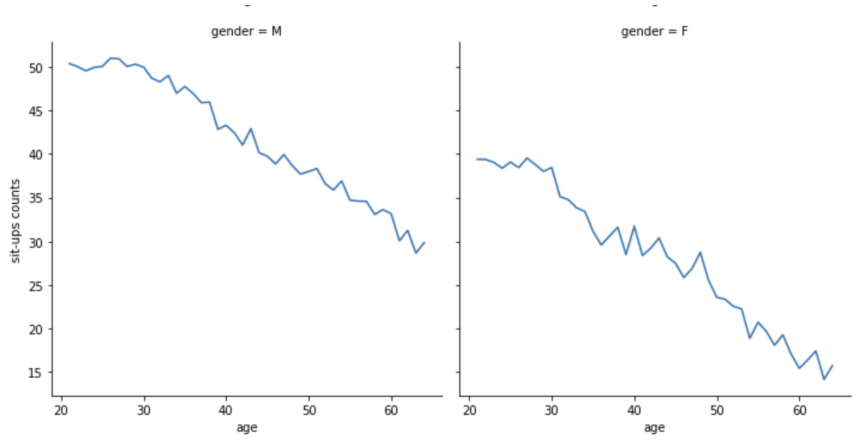
Tuổi và lực kẹp



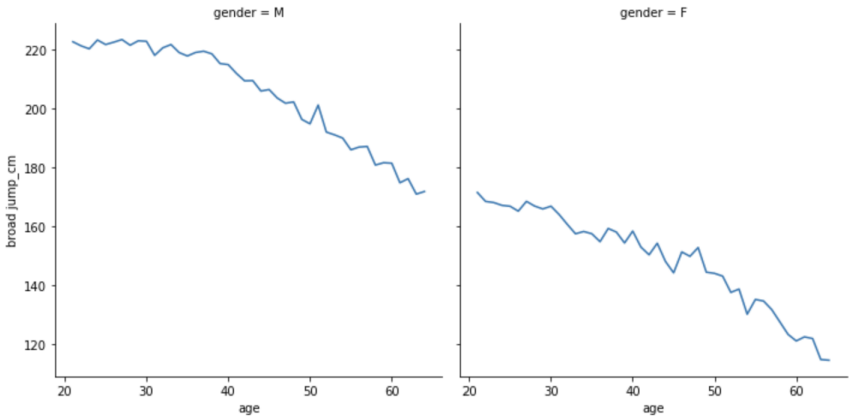
Tuổi và số lần ngồi và gập về trước



Tuổi và số lần gập bụng



Tuổi và khoảng cách nhảy xa



Ta có thể thấy rõ hơn sự khác biệt giữa nam và nữ trong hiệu suất các bài thể dục
Hiệu suất tập thể dục cũng giảm dần theo độ tuổi

Bây giờ ta thực hiện A/B testing để xem sự khác biệt này có phải do ngẫu nhiên hay có ý nghĩa thống kê

3.1.1 Thực hiện A/B testing đối với giới tính

Qua cái histogram ta có thể thấy có sự khác biệt về các bài thể dục giữa nam và nữ. Bây giờ ta sử dụng A/B testing với kiểm định hoán vị xem sự khác biệt này là do ngẫu nhiên hay có tính thống kê.
Giả thuyết không (H0):
Không có sự khác biệt có ý nghĩa thống kê giữa nam và nữ về các bài thể dục.
H0: $\mu_{nam} = \mu_{nu}$

Giả thuyết đối (H1):

Có sự khác biệt có ý nghĩa thống kê giữa nam và nữ về các bài thể dục.

H1: $\mu_{nam} \neq \mu_{nu}$

Sau khi thực hiện ta được kết quả:

```
Kiểm định hoán vị cho biến: gripForce
Chênh lệch trung bình thực tế: 17.82315
P-value: 0

Kiểm định hoán vị cho biến: sit.and.bend.forward_cm
Chênh lệch trung bình thực tế: -5.030667
P-value: 0

Kiểm định hoán vị cho biến: sit.ups.counts
Chênh lệch trung bình thực tế: 14.50011
P-value: 0

Kiểm định hoán vị cho biến: broad.jump_cm
Chênh lệch trung bình thực tế: 59.1805
P-value: 0
```

Các giá trị p-value đều = 0 (tiến về 0) chứng tỏ sự khác biệc ở các bài tập giữa nam và nữ đối với hiệu suất thể thao không phải ngẫu nhiên và hoàn toàn có ý nghĩa thống kê.

3.1.2 Thực hiện A/B testing đối với nhóm tuổi

Qua cái histogram ta có thể thấy có sự khác biệt về các bài thể dục giữa các nhóm tuổi. Bây giờ ta chia những người < 40 tuổi thuộc nhóm A và > 40 tuổi thuộc nhóm B để sử dụng A/B testing với kiểm định hoán vị xem sự khác biệt này là do ngẫu nhiên hay có tính thống kê.

Giả thuyết không (H0):

Không có sự khác biệt có ý nghĩa thống kê giữa nhóm A (người < 40 tuổi) và nhóm B (người > 40 tuổi) về các bài thể dục.

H0: $\mu_A = \mu_B$

Giả thuyết đối (H1):

Có sự khác biệt có ý nghĩa thống kê giữa nhóm A và nhóm B về các bài thể dục

H1: $\mu_A \neq \mu_B$

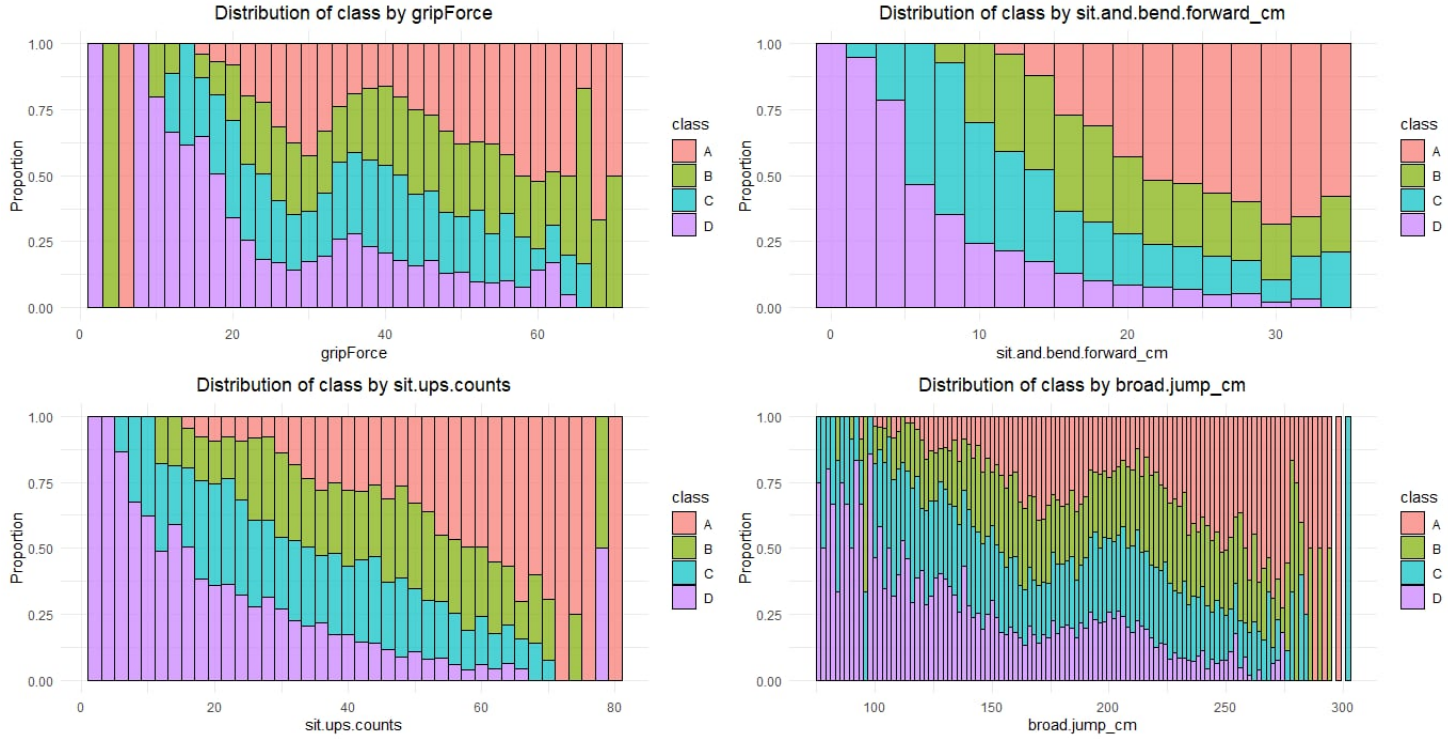
Sau khi thực hiện ta được kết quả:

```
Kiểm định hoán vị cho biến: gripForce
Chênh lệch trung bình thực tế: 4.690098
P-value: 0

Kiểm định hoán vị cho biến: sit.and.bend.forward_cm
Chênh lệch trung bình thực tế: 1.258966
P-value: 0

Kiểm định hoán vị cho biến: sit.ups.counts
Chênh lệch trung bình thực tế: 15.29625
P-value: 0

Kiểm định hoán vị cho biến: broad.jump_cm
Chênh lệch trung bình thực tế: 34.94387
P-value: 0
```

Các giá trị p-value đều = 0 (tiến về 0) chứng tỏ sự khác biệc ở các bài tập giữa 2 nhóm tuổi đối với hiệu suất thể thao không ngẫu nhiên và hoàn toàn có ý nghĩa thống kê.

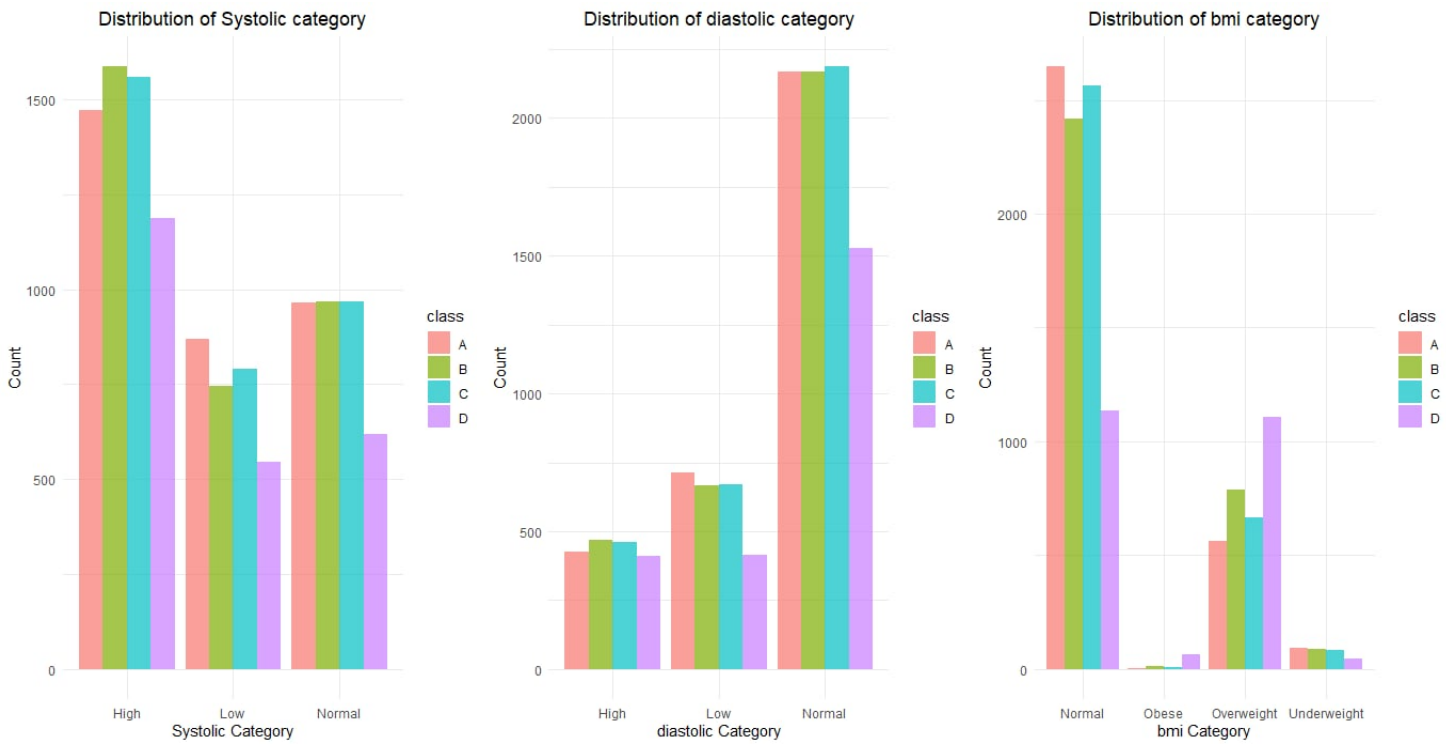
3.2 Các chỉ số khác với hiệu suất thể thao

Các lớp hiệu suất và chỉ số các môn thể dục



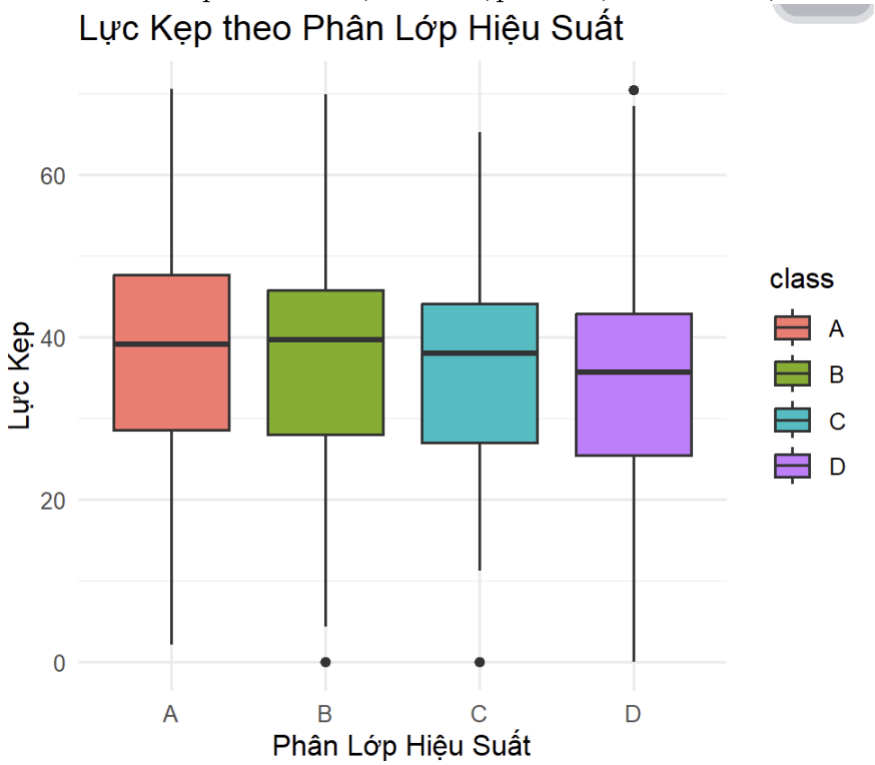
Các lớp hiệu suất với trạng thái cơ thể, chỉ số huyết áp tâm thu, tâm trương

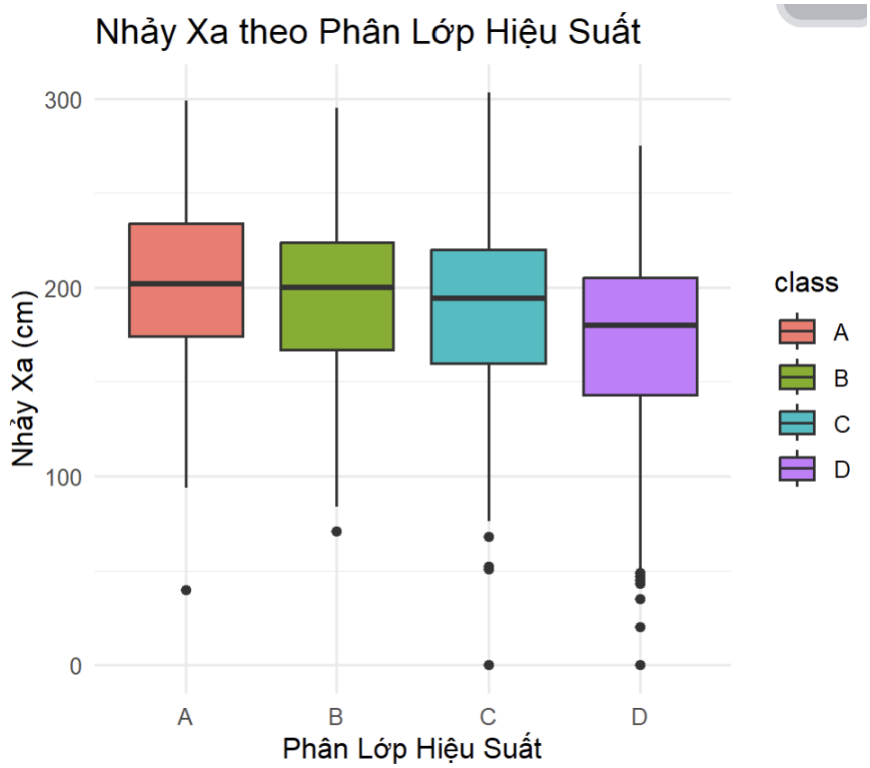
Các lớp hiệu suất và chỉ số khối cơ thể

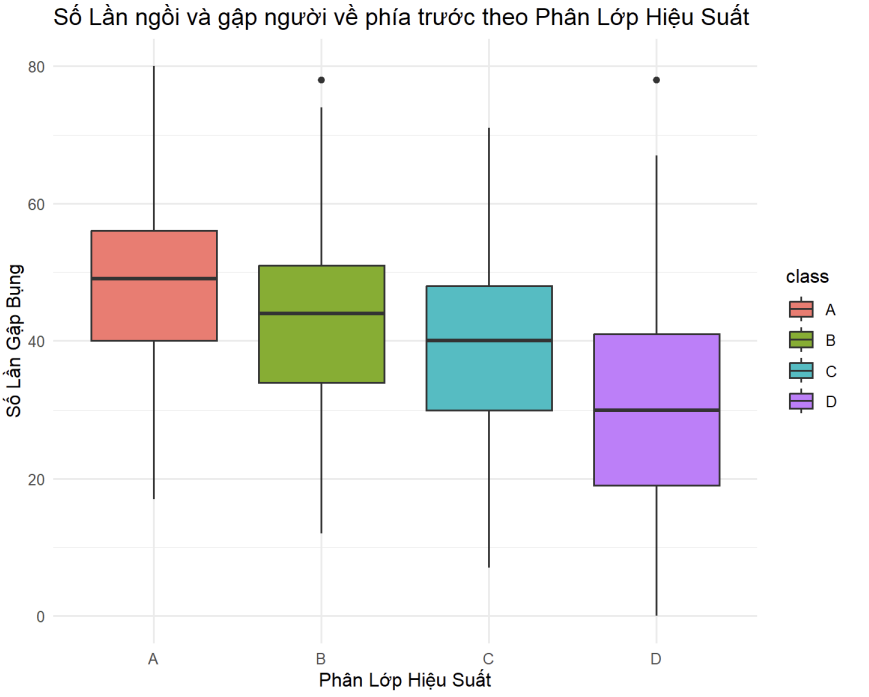


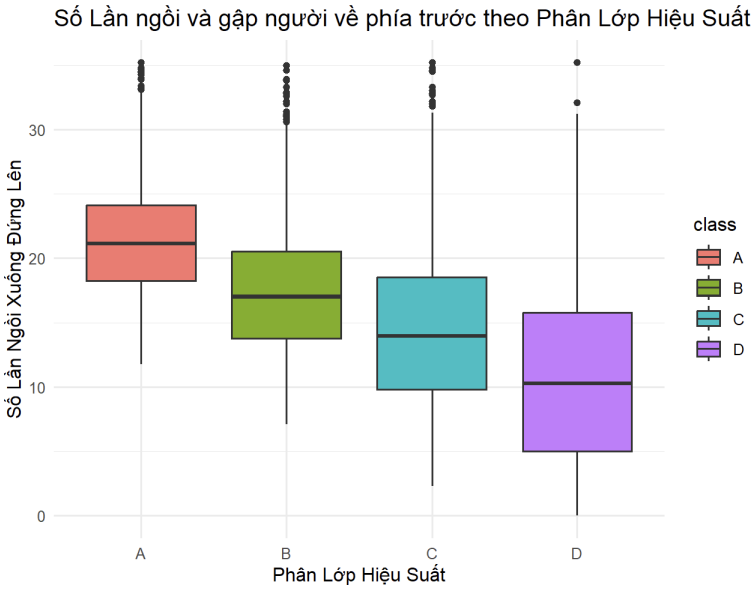
3.3 Các bài tập thể dục với hiệu suất các môn thể dục

Ta vẽ các Box plot xem hiệu suất tập thể dục đối với hiệu suất




Nhảy Xa theo Phân Lớp Hiệu Suất
Nhảy Xa (cm)
300
200
100
0
A
B
C
D
Phân Lớp Hiệu Suất
class
A
B
C
D


Số Lần ngồi và gập người về phía trước theo Phân Lớp Hiệu Suất
Số Lần Gập Bụng
80
60
40
20
0
A
B
C
D
Phân Lớp Hiệu Suất
class
A
B
C
D



3.3.1 Sử dụng Permutation ANOVA kiểm định sự khác biệt của các lớp hiệu suất

Sử dụng ANOVA xem có sự khác biệc giữa lớp hiệu suất thể thao(A,B,C,D) trong việc tập cấc môn thể dục

Null hypothesis (H0): Không có sự khác biệt có ý nghĩa giữa các lớp hiệu suất (A, B, C, D) về các biến kiểm tra (lực kẹp, nhảy xa, số lần gập bụng, tỷ lệ mỡ cơ thể, v.v.).

Alternative hypothesis (H1): Có sự khác biệt có ý nghĩa giữa ít nhất một cặp lớp hiệu suất về các biến bạn đang kiểm tra.

```
Hoán vị ANOVA cho biến: gripForce

        Approximative K-Sample Fisher-Pitman Permutation Test

data:  gripForce by class (A, B, C, D)
chi-squared = 200.47, p-value < 1e-04


Hoán vị ANOVA cho biến: sit.and.bend.forward_cm

        Approximative K-Sample Fisher-Pitman Permutation Test

data:  sit.and.bend.forward_cm by class (A, B, C, D)
chi-squared = 4030.1, p-value < 1e-04


Hoán vị ANOVA cho biến: sit.ups.counts

        Approximative K-Sample Fisher-Pitman Permutation Test

data:  sit.ups.counts by class (A, B, C, D)
chi-squared = 2266.3, p-value < 1e-04


Hoán vị ANOVA cho biến: broad.jump_cm

        Approximative K-Sample Fisher-Pitman Permutation Test

data:  broad.jump_cm by class (A, B, C, D)
chi-squared = 665.15, p-value < 1e-04
```

P-value rất nhỏ ($<$ 1e-4) cho thấy rằng có rất ít khả năng rằng sự khác biệt giữa các nhóm chỉ là ngẫu nhiên. Điều này giúp bác bỏ giả thuyết không có sự khác biệt giữa các nhóm (class). Tất cả các môn thể dục gripForce, sit and bend forward_cm, sit-ups counts, broad jump_cm đều cho thấy sự khác biệt có ý nghĩa thống kê giữa các lớp hiệu suất (A, B, C, D). Điều này cho thấy rằng các lớp hiệu suất khác nhau có mức độ thể chất khác nhau đối với các môn thể dục này

4 Áp dụng mô hình học máy dự đoán phân loại các lớp hiệu suất

Để dự đoán và phân loại hiệu suất thể thao (A, B, C, D) dựa trên các yếu tố như tuổi, giới tính, chiều cao, cân nặng, lực kẹp, số lần gập bụng, và khả năng nhảy xa, chúng tôi sẽ áp dụng các mô hình học máy. Việc này không chỉ giúp xác định độ chính xác của dự đoán mà còn cho phép đánh giá mức độ ảnh hưởng của từng yếu tố đối với hiệu suất. Hai mô hình chính được sử dụng là Logistic Regression để phân tích mối quan hệ tuyến tính và Random Forest để khai thác các mối quan hệ phi tuyến tính cũng như đánh giá tầm quan trọng của các đặc trưng.

4.1 Chuẩn bị dữ liệu

Ta tiến hành các xử lí sau:

• Chuyển đổi các cột phân loại(category) thành số.

• Chuẩn hóa các cột numeric dùng hàm scale() trong thư viện caret.

Sau khi hoàn thành các bước trên ta được tập dữ liệu:

```
> head(indexData)
         age gender  height_cm   weight_kg body.fat_.  diastolic    systolic
1 -0.7169611      1  0.4634846  0.68937067 -0.2491219  0.1163632 -0.01441668
2 -0.8635344      1 -0.4032474 -0.96085249 -1.0226616 -0.1638292 -0.28633295
3 -0.4238146      1  1.3302165  0.92366161 -0.4148804  1.2371326  1.48112279
4 -0.3505280      1  0.7246915  0.33793426 -0.6497050 -0.2572266  1.14122746
5 -0.6436745      1  0.6415802  0.04931498 -0.8292767 -0.8176114 -0.21835388
6 -0.0573815      0 -0.3557552 -0.99480770 -0.1524294 -1.3779961 -0.76218642
   gripForce sit.and.bend.forward_cm sit.ups.counts broad.jump_cm class
1  1.6755574             0.323067090      1.3968520     0.6583551     3
2 -0.0553408             0.006328152      0.9033166     0.9591532     1
3  0.7305805            -0.642232532      0.6212964    -0.2440392     3
4  0.4124695            -0.159582721      0.9033166     0.7084881     2
5  0.6089499             1.635271263      0.3392762     0.6583551     2
6 -1.2342228             0.715220061     -0.9298148    -0.9459015     2
     BMIndex situation index_systolic index_diastolic
1  0.6254321         3              2               2
2 -1.0536248         2              2               2
3  0.2227504         2              3               3
4 -0.0653395         2              3               2
5 -0.3898819         2              2               2
6 -1.1385515         2              1               1
```

• Chia dữ liệu thành tập train và test với tỉ lệ 7: 3.

```
> dim(train_data)
[1] 8922   16
> dim(test_data)
[1] 3823   16
```

4.2 Random Forest

4.2.1 Mục tiêu

Mục tiêu của việc áp dụng mô hình Random Forest là sử dụng một tập hợp các cây quyết định (decision trees) để xây dựng một mô hình học máy mạnh mẽ có khả năng phân loại hoặc dự đoán một biến mục tiêu .

Random Forest có khả năng tìm ra các đặc trưng quan trọng (feature importance).Giúp xác định các biến (đặc trưng) nào ảnh hưởng nhiều nhất đến quyết định của mô hình.

4.2.2 Xây dựng mô hình

Chúng ta sẽ dùng K fold cross validation để đánh giá và tìm tham số tối ưu

Sử dụng hàm train() trong thư viện Caret và được kết quả như sau.

```
Random Forest

8923 samples
  15 predictor
   4 classes: '1', '2', '3', '4'

No pre-processing
Resampling: Cross-Validated (5 fold)
Summary of sample sizes: 7138, 7140, 7138, 7139, 7137
Resampling results across tuning parameters:

  mtry  Accuracy   Kappa
   2    0.6901284  0.5850830
   8    0.7250939  0.6317065
  15    0.7182581  0.6225150

Accuracy was used to select the optimal model using the largest value.
The final value used for the model was mtry = 8.
```

Ta thấy với mtry = 8 mô hình cho độ chính xác tốt nhất.

Bây giờ, ta áp dụng Random forest với mtry = 8 cho tập test xem mô hình hoạt động ra sao

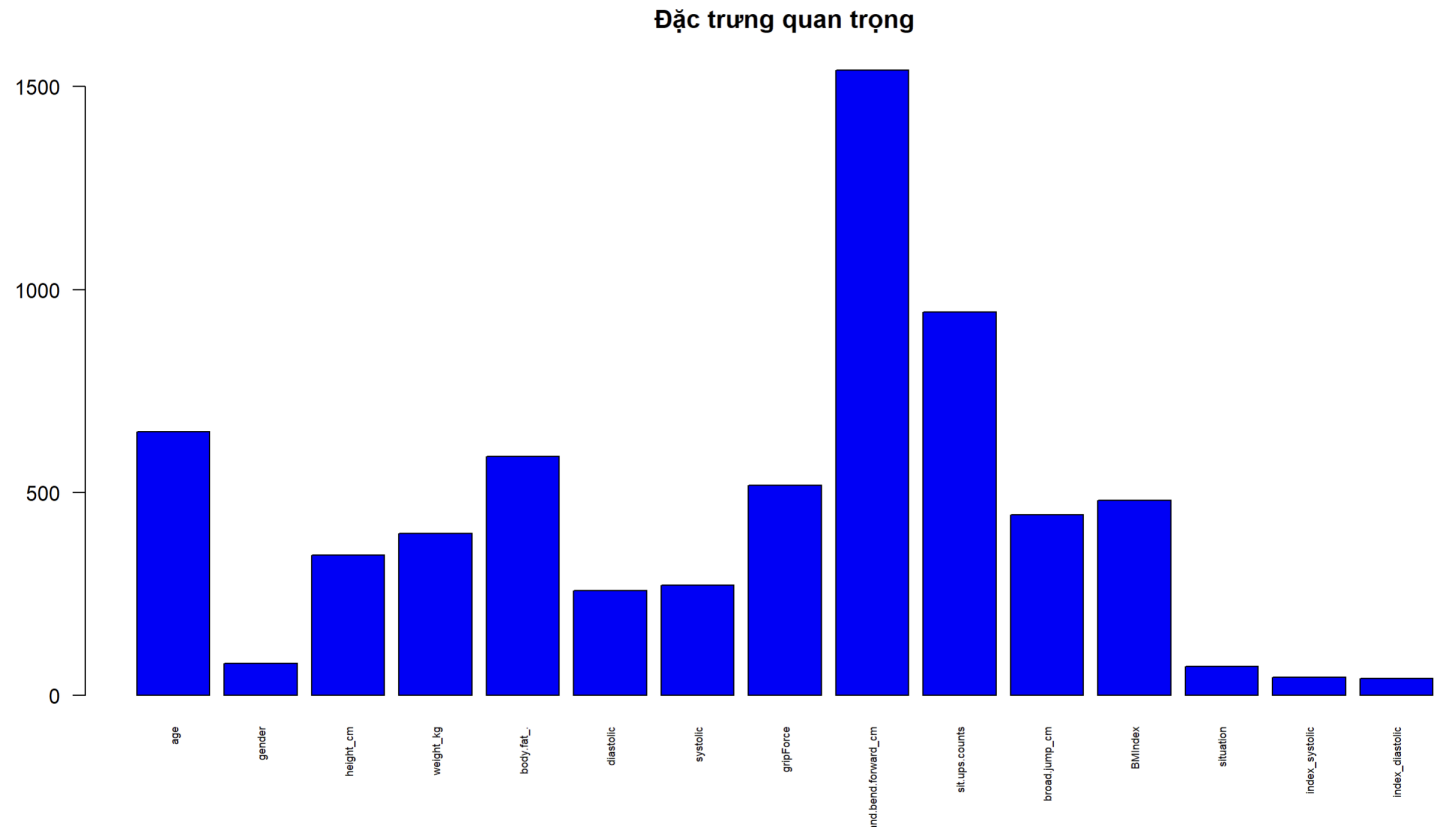
```
Confusion Matrix and Statistics

          Reference
Prediction   1   2   3   4
         1 852 236  87  11
         2 132 609 206  48
         3  18 121 687 125
         4   1  37  24 628

Overall Statistics

               Accuracy : 0.7263
                 95% CI : (0.7119, 0.7404)
    No Information Rate : 0.2627
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.6334

 Mcnemar's Test P-Value : < 2.2e-16
```

Confusion matrix cho thấy số lượng dự đoán đúng và sai cho mỗi lớp (1, 2, 3, 4) từ mô hình Random Forest.

Accuracy của mô hình là 72,63% . Vì đây là bài toán áp dụng dự đoán phân loại cho 4 lớp nên với độ chính xác này là chấp nhận được.

4.2.3 Các đặc trưng quan trọng

Sử dụng hàm impotance() trong Random Forest để xem ảnh hưởng của các đặc trưng



• Các chỉ số tập thể dục ảnh hưởng lớn đến mô hình phân loại, ngoài ra tuổi và % mỡ cơ thể cũng quan trọng

• Có thể thấy ảnh hưởng của các biết phân loại(gender, situation, index_systolic, index_diastolic) không đáng kể trong Random forest.

Ở phần tiếp theo, chúng tôi gộp hiệu suất thành 2 nhóm(A, B : hiệu suất tốt, C, D: hiệu suất không tốt) và sử dụng Logistic Regression để dự đoán phân loại

4.3 Logistic Regression

4.3.1 Mục tiêu

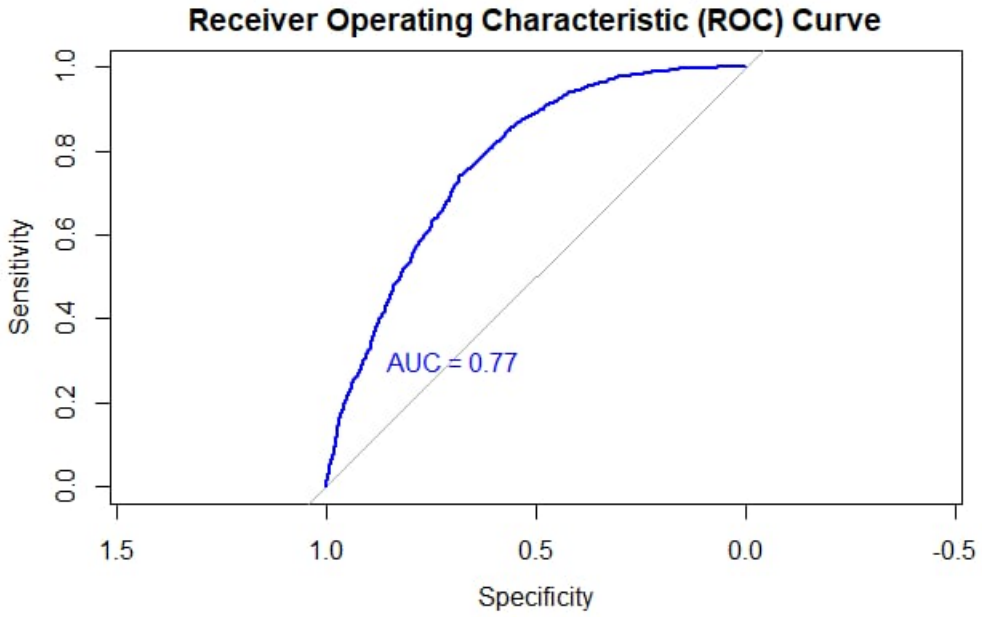
Mô hình này giúp phân loại dữ liệu thành các nhóm cụ thể (chẳng hạn như nhóm hiệu suất A, B, C, D). Logistic regression là một phương pháp mạnh mẽ để dự đoán các nhãn nhị phân hoặc đa lớp (sau khi chuyển đổi thành nhãn nhị phân). Mô hình sẽ chỉ ra mối quan hệ giữa các biến đầu

vào (như tuổi, giới tính, chỉ số cơ thể...) và khả năng gia nhập vào nhóm hiệu suất cao (A) hay không.

4.3.2 Ý nghĩa

Tìm mối quan hệ giữa biến độc lập và biến phụ thuộc: Mô hình giúp xác định những yếu tố nào có tác động mạnh nhất đến hiệu suất thể thao. Dự đoán kết quả: Logistic regression có thể dự đoán người tham gia có khả năng đạt hiệu suất thể thao cao hay không dựa trên các đặc trưng của họ. Tăng cường quyết định: Mô hình này có thể giúp các chuyên gia thể thao và sức khỏe đưa ra quyết định về các yếu tố cần cải thiện.

```
            age       grip_force   sit_ups_counts    broad_jump_cm
       2.137385         3.753003         3.285427         4.068901
body_fat_percent         systolic        diastolic              bmi
       3.028318         2.038890         1.871095         2.447703
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 1116  455
         1  611 1533

               Accuracy : 0.7131
                 95% CI : (0.6982, 0.7276)
    No Information Rate : 0.5351
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.4198

 Mcnemar's Test P-Value : 2.061e-06

            Sensitivity : 0.6462
            Specificity : 0.7711
         Pos Pred Value : 0.7104
         Neg Pred Value : 0.7150
             Prevalence : 0.4649
         Detection Rate : 0.3004
   Detection Prevalence : 0.4229
      Balanced Accuracy : 0.7087

       'Positive' Class : 0
```



4.3.3 Kết luận

Các yếu tố dự đoán có ý nghĩa (p-value < 0.05) bao gồm tuổi, số lần chống đẩy, tỷ lệ mỡ cơ thể, huyết áp tâm trương, và BMI, trong khi các yếu tố như lực nắm tay, nhảy xa (cm) và huyết áp tâm thu không đóng góp đáng kể. Điều này cho thấy các yếu tố này có thể không phải là chỉ số mạnh để dự đoán kết quả trong bộ dữ liệu này.

Mô hình đạt được độ chính xác 71.31%, cao hơn so với tỷ lệ không thông tin (NIR) là 53.51%, cho thấy mô hình có hiệu quả vượt trội so với việc dự đoán ngẫu nhiên. Các chỉ số Kappa trung bình và các chỉ số như độ chính xác, nhạy cảm, và đặc hiệu cho thấy mô hình có thể được sử dụng để dự đoán, mặc dù có thể cải thiện bằng cách nghiên cứu thêm các tương tác giữa các biến hoặc thêm các yếu tố khác.

5 Tổng kết

Bài báo cáo bao gồm các bước thực hiện làm sạch dữ liệu, kiểm định giả thuyết, kỹ thuật tạo đặc trưng và ứng dụng các mô hình học máy để dự đoán hiệu suất thể dục. Phân tích thống kê và phương pháp đánh giá mô hình được trình bày theo định dạng có cấu trúc, mang tính học thuật.