TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN
ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH

KHOA TOÁN - TIN HỌC

Báo cáo cuối kì
Xử LÍ SỐ LIỆU THỐNG KÊ

GV bộ môn: Huỳnh Thanh Sơn

Người thực hiện

Châu Gia Kiệt 22110095

Hồ Chí Minh, Tuesday, 14 January 2025

Mục lục

Giới thiệu 2
Chủ đề 2
Chiến lược phân tích và xử lí dữ liệu 2
Mục tiêu nghiên cứu 3
Các phương pháp và chiến lược xử lí dữ liệu 3
Tổng quan về bộ dữ liệu 4
Các biến trong bộ dữ liệu 4
Số lượng và loại dữ liệu 6
Sự phân phố của các biến trong dữ liệu 7
So sánh hiệu suất giữa các biến 9
Thống kê mô tả 10
Phân tích dữ liệu 11
Sự tương quan của bộ dữ liệu 11
So sánh sự khác biệt giữa các nhóm 12
Phân tích đặc điểm nhóm hiệu suất cao (Class A) 15
Phân tích ảnh hưởng của giới tính đến hiệu suất 15
So sánh trực quan các đặc trưng giữa các nhóm và giới tính: . 16
Mô hình phân loại 18
Chuẩn bị dữ liệu 18
Xây dựng và đánh giá các mô hình phân loại 18
Các đặc trưng quan trọng trong từng mô hình và chỉ số 23
Tổng quan hiệu quả về các mô hình xây dựng 27
Tổng kết 28
Thuận lợi 28
Khó khăn 29

Giới thiệu

Chủ đề

Đề tài "Body Performance Data"xuất phát từ sự phát triển mạnh mẽ của các phong trào thể thao hiện nay, thu hút sự tham gia của đa dạng các nhóm tuổi và giới tính. Việc thúc đẩy lối sống lành mạnh, nâng cao sức khỏe cộng đồng, và cải thiện hiệu suất thể chất đang ngày càng trở nên quan trọng. Do đó, việc nghiên cứu và hiểu rõ về các yếu tố ảnh hưởng đến hiệu quả của việc tập thể dục là vô cùng cần thiết. Bộ dữ liệu "BodyPerformance.csv"cung cấp thông tin về 13,393 người tham gia tập thể thao tại Hàn Quốc, với các chỉ số thể chất như chiều cao, cân nặng, huyết áp, tỷ lệ mỡ cơ thể, và kết quả từ các bài kiểm tra thể lực (như nhảy xa, gập bụng, lực kẹp tay, v.v.). Bộ dữ liệu này cung cấp một cơ sở dữ liệu phong phú để phân tích và đưa ra các kết luận khoa học về mối quan hệ giữa các yếu tố cá nhân (như độ tuổi, giới tính, cân nặng) và các chỉ số thể lực. Qua đó, nghiên cứu này sẽ giúp các chuyên gia sức khỏe và huấn luyện viên thể thao có thể đưa ra những phương pháp tối ưu để cải thiện hiệu quả tập luyện cho từng nhóm đối tượng, từ đó nâng cao chất lượng sức khỏe và hiệu suất thể thao của cộng đồng.

Chiến lược phân tích và xử lí dữ liệu

1. Làm sạch dữ liệu, kiểm tra xem bộ dữ liệu có các giá trị bị thiếu (NA), dữ liệu trùng lắp, hay có các giá trị ngoại lai làm nhiễu mô hình hay không.
2. Thực hiện tiền xử lý dữ liệu như xoá các giá trị thiếu, xoá các dữ liệu bị trùng lắp, chuẩn hoá dữ liệu để giải quyết các giá trị ngoại lai.
3. Thực hiện phân tích khám phá dữ liệu (EDA) để hiểu các xu hướng cơ bản như thực hiện thống kê mô tả hay quan sát phân phối của dữ liệu.
4. Sử dụng các kỹ thuật phân tích để đánh giá và so sánh tầm quan trọng từng yếu tố đối với mục tiêu của bài toán cũng như xác định mối quan hệ giữa các biến, như A/B testing hay ANOVA nếu có nhiều nhóm và phân tích tương quan.
5. Áp dụng để xây dựng và đánh giá mô hình dựa trên mục tiêu và yêu cầu của bài toán.

Mục tiêu nghiên cứu

Mục tiêu của nghiên cứu này là phân tích và hiểu rõ hơn về mối quan hệ giữa các yếu tố thể chất cá nhân và hiệu suất thể thao của người tham gia. Trong bối cảnh các phong trào thể thao đang ngày càng phát triển mạnh mẽ và thu hút sự tham gia của nhiều nhóm đối tượng với độ tuổi và giới tính đa dạng, việc nắm bắt được hiệu quả của việc tập luyện thể thao và các yếu tố ảnh hưởng đến hiệu suất thể chất trở nên vô cùng quan trọng. Dữ liệu "BodyPerformance.csv"cung cấp thông tin chi tiết về các chỉ số thể chất của hơn 13,000 người tham gia tập thể thao, sẽ được sử dụng để phân tích các yếu tố như chiều cao, cân nặng, huyết áp, tỷ lệ mỡ cơ thể, và các bài kiểm tra thể lực. Việc phân tích dữ liệu này sẽ giúp đưa ra những kết luận khoa học phục vụ cho các chuyên gia sức khỏe và huấn luyện viên thể thao, giúp cải thiện hiệu quả tập luyện thể thao và tối ưu hóa chương trình sức khỏe cho từng nhóm đối tượng.

Các phương pháp và chiến lược xử lí dữ liệu

Đầu tiên, ta sẽ thực hiện tiền xử lý dữ liệu bao gồm kiểm tra và xử lý các giá trị bị thiếu, dữ liệu bị trùng lắp và các giá trị ngoại lai.

- Kiểm tra giá trị bị thiếu bằng cách sử dụng hàm `aggr()` và tính tổng số lượng giá trị `NA` có trong từng cột trong bộ dữ liệu. Nếu có, ta có thể thực hiện xoá dữ liệu hoặc suy đoán dữ liệu bằng chiến lược single imputation hay multiple imputation.
- Để kiểm tra và xoá dữ liệu bị trùng lắp, ta sử dụng hàm `duplicated()` trong R.
- Vẽ box plot để kiểm tra và trực quan hoá giá trị ngoại lai của các biến. Sau đó, ta áp dụng chiến lược xoá bỏ giá trị ngoại lai dựa trên `IQR` và `winsorization`.

Sau khi tiền xử lý dữ liệu, ta đến bước khám phá dữ liệu (EDA). Ta sử dụng hàm `summary()` để tính thống kê mô tả, các giá trị như trung bình, độ lệch chuẩn,...
Sau đó, ta có thể vẽ phân phối và ước lượng hàm mật độ xác suất của từng biến trong bộ dữ liệu bằng `ggplot2`.
Tiếp theo, để phân tích mối liên hệ giữa các biến, ta có thể sử dụng biểu đồ hệ số tương quan bằng `cor` và `corrplot`. Bên cạnh đó, để so sánh các biến với nhau, ta có thể sử dụng các chiến lược như A/B testing hay ANOVA.

Sau khi thực hiện khám phá dữ liệu để hiểu hơn về mối liên hệ giữa các biến và xu hướng của chúng, ta sẽ xây dựng và đánh giá mô hình phân loại. Trong dự án này, ta sẽ xây dựng và đánh giá ba mô hình là multi logistic, random forest và XGBoost.

Trước khi xây dựng mô hình, ta áp dụng cross-validation để chia tập dữ liệu thành tập huấn luyện (train) và kiểm tra (test) để kiểm tra độ chính xác của của mô hình trên tập dữ liệu chưa được nhìn thấy. Sau khi xử lý các giá trị ngoại lai, dữ liệu bị mất cân bằng nên để xây dựng mô hình hiệu quả, chúng ta sử dụng phương pháp SMOTE trên tập train để tạo ra thêm dữ liệu cho nhóm bị mất cân bằng và bắt tay vào xây dựng ba mô hình phân loại đã nêu. Tiếp theo, để đánh giá mô hình, ta sử dụng confusion matrix để thấy được sự nhầm lẫn giữa các nhóm khi mô hình đang được tổng quát trên tập test, từ đó tính được độ chính xác (accuracy) của mô hình.

Tổng quan về bộ dữ liệu

Bộ dữ liệu BodyPerformance.csv là một nguồn tài liệu quý giá giúp các chuyên gia sức khỏe và thể thao hiểu rõ hơn về các yếu tố ảnh hưởng đến hiệu suất thể lực của con người. Với thông tin chi tiết về các chỉ số thể chất và kết quả bài kiểm tra thể lực, bộ dữ liệu này cung cấp một cơ sở vững chắc để nghiên cứu, phân tích và phát triển các phương pháp tập luyện hiệu quả cho mọi người. Bao gồm 13,393 bản ghi, mỗi bản ghi đại diện cho một cá nhân tham gia vào một chương trình tập luyện thể thao. Các bản ghi này cung cấp thông tin chi tiết về các chỉ số thể chất, các bài kiểm tra thể lực, và phân lớp hiệu suất của mỗi cá nhân. Bộ dữ liệu này đặc biệt hữu ích cho các chuyên gia trong lĩnh vực sức khỏe và thể thao, các nhà nghiên cứu về dinh dưỡng, thể dục thể thao, và các huấn luyện viên trong việc đánh giá và đưa ra các phương pháp tập luyện phù hợp cho từng nhóm đối tượng.

Các biến trong bộ dữ liệu

Bộ dữ liệu BodyPerformance.csv chứa 13,393 bản ghi với 12 biến đặc trưng, mô tả các thông tin về thể chất, thể lực và tình trạng sức khỏe của các cá nhân tham gia chương trình tập luyện thể thao. Dưới đây là mô tả chi tiết về từng biến trong bộ dữ liệu:

1. age (Độ tuổi): Biến này biểu thị độ tuổi của người tham gia, dao động từ 20 đến 64 tuổi. Đây là yếu tố quan trọng vì nó ảnh hưởng trực tiếp đến khả năng phục hồi cơ thể, mức độ chịu đựng trong khi tập luyện và hiệu suất thể thao. Độ tuổi cũng giúp phân nhóm người tham gia để nghiên cứu sự khác biệt về thể lực giữa các nhóm tuổi khác nhau. Loại dữ liệu: integer.
2. gender (Giới tính): Biến này cho biết giới tính của người tham gia, với các giá trị 'F' (nữ) và 'M' (nam). Giới tính ảnh hưởng đến các đặc điểm sinh lý, cấu trúc cơ thể và các chỉ số thể chất như tỷ lệ mỡ cơ thể, sức mạnh cơ bắp, khả năng chịu đựng. Dữ liệu này có thể giúp xác định các chương trình tập luyện hiệu quả cho nam và nữ. Loại dữ liệu: categorical.
3. height_cm (Chiều cao): Biến này cung cấp chiều cao của người tham gia (đơn vị: cm). Chiều cao có thể tác động đến các bài kiểm tra thể lực như nhảy xa hoặc các bài kiểm tra linh hoạt. Chiều cao cũng có mối quan hệ với các chỉ số như cân nặng và tỷ lệ mỡ cơ thể. Loại dữ liệu: numeric.
4. weight_kg (Cân nặng): Biến này đo lường cân nặng của người tham gia (đơn vị: kg). Cân nặng ảnh hưởng đến các chỉ số như chỉ số khối cơ thể (BMI), khả năng thực hiện các bài tập thể lực và mối quan hệ với tỷ lệ mỡ cơ thể. Dữ liệu này có thể giúp tìm ra những cá nhân cần điều chỉnh chế độ ăn uống và luyện tập. Loại dữ liệu: numeric.
5. body fat Đây là một chỉ số quan trọng để đánh giá sức khỏe tổng thể. Tỷ lệ mỡ cơ thể cao có thể dẫn đến các vấn đề về tim mạch, tiểu đường và các bệnh lý khác. Phân tích tỷ lệ mỡ cơ thể giúp các chuyên gia xác định các cá nhân cần cải thiện sức khỏe thông qua việc giảm mỡ và tăng cường cơ bắp. Loại dữ liệu: numeric.
6. diastolic (Huyết áp tâm trương): Biến này đo lường huyết áp tâm trương (đơn vị: mmHg). Huyết áp tâm trương phản ánh áp lực trong động mạch khi tim nghỉ giữa các nhịp đập. Huyết áp cao có thể là dấu hiệu của nguy cơ mắc các bệnh tim mạch, đặc biệt là ở những người ít vận động. Loại dữ liệu: numeric.
7. systolic (Huyết áp tâm thu): Biến này đo huyết áp tâm thu (đơn vị: mmHg), tức là áp lực trong động mạch khi tim đập. Huyết áp tâm thu có ảnh hưởng lớn đến tình trạng sức khỏe tổng thể. Những người có huyết áp cao có nguy cơ mắc các bệnh tim mạch và đột quỵ. Loại dữ liệu: numeric.

8. gripForce (Lực kẹp tay): Biến này đo lường sức mạnh của tay qua bài kiểm tra lực kẹp. Đây là chỉ số thể lực thể hiện sức mạnh cơ bắp của tay và cánh tay. Lực kẹp tay là chỉ số quan trọng để đánh giá mức độ khỏe mạnh của hệ cơ xương, đặc biệt là đối với các bài tập cường độ cao hoặc các hoạt động cần sử dụng lực tay mạnh. Loại dữ liệu: numeric.
9. sit and bend forward_cm (Ngồi và gập người về phía trước): Biến này đo khả năng linh hoạt của cơ thể khi ngồi và gập người về phía trước (đơn vị: cm). Đây là bài kiểm tra khả năng dẻo dai và linh hoạt của cơ thể, đặc biệt là các cơ lưng và cơ đùi. Loại dữ liệu: numeric.
10. sit-ups counts (Số lần gập bụng): Biến này đo số lần gập bụng mà người tham gia có thể thực hiện trong một khoảng thời gian nhất định. Đây là bài kiểm tra sức mạnh cơ bụng và sự chịu đựng của cơ thể, giúp đánh giá khả năng thể lực chung, đặc biệt là sức mạnh và độ bền của vùng cơ bụng. Loại dữ liệu: integer.
11. broad jump_cm (Nhảy xa): Biến này đo khoảng cách người tham gia có thể nhảy xa từ vị trí đứng (đơn vị: cm). Đây là bài kiểm tra sức mạnh và sự linh hoạt của đôi chân, đồng thời phản ánh sự kết hợp giữa khả năng đẩy mạnh và sức mạnh cơ bắp. Loại dữ liệu: numeric.
12. class (Phân lớp hiệu suất): Biến này phân loại hiệu suất thể chất của người tham gia thành 4 lớp: A, B, C, D. Lớp A là nhóm có hiệu suất tốt nhất, trong khi lớp D đại diện cho nhóm có hiệu suất kém nhất. Đây là yếu tố quan trọng để đánh giá mức độ tập luyện của mỗi người và giúp các chuyên gia thể thao xác định các chương trình luyện tập phù hợp. Loại dữ liệu: categorical.

Số lượng và loại dữ liệu

Bộ dữ liệu BodyPerformance.csv bao gồm tổng cộng 13,393 bản ghi, mỗi bản ghi đại diện cho một cá nhân tham gia chương trình tập luyện thể thao. Các bản ghi này chứa 12 biến đặc trưng về thể chất, thể lực và tình trạng sức khỏe, trong đó:

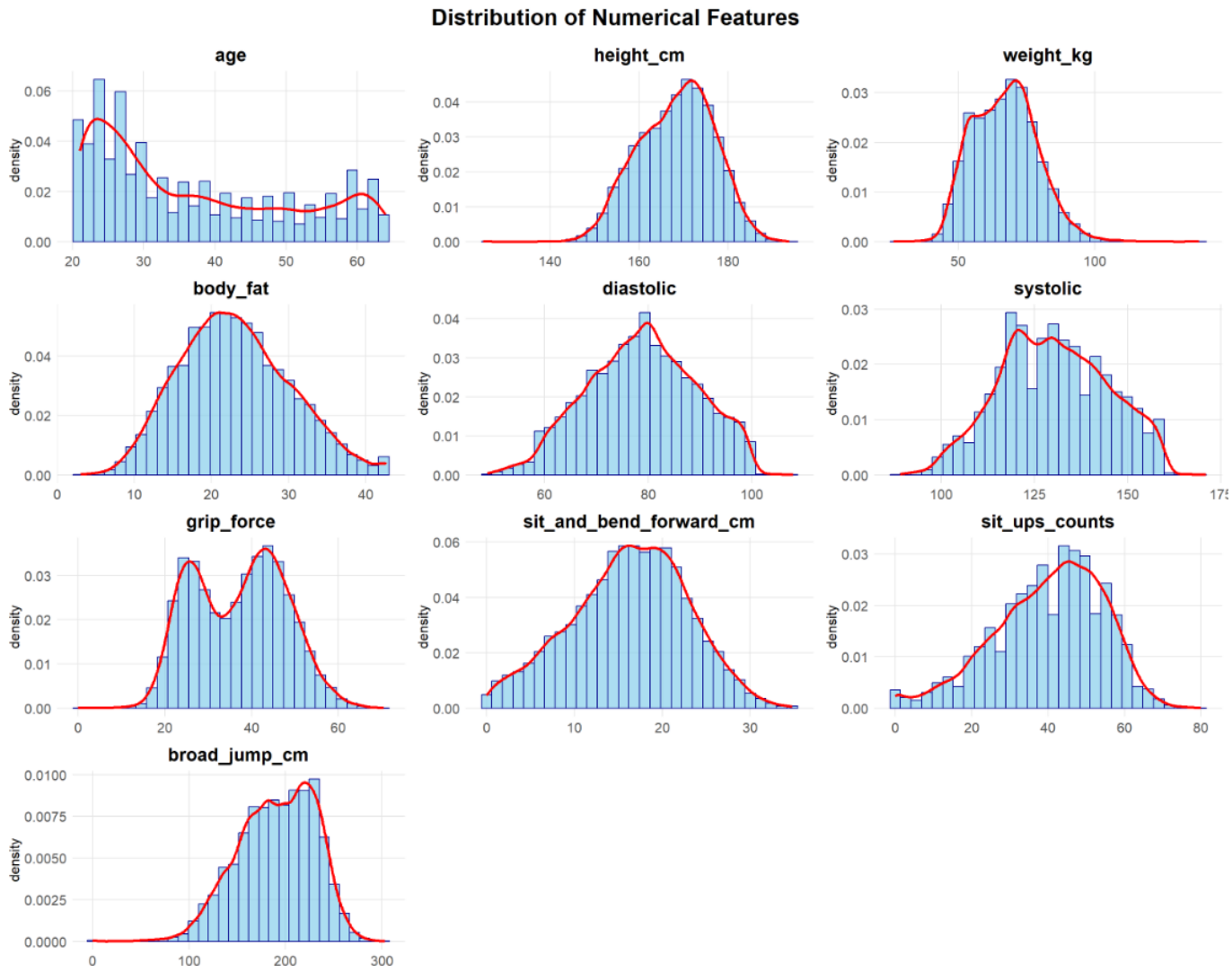
- Các biến về độ tuổi, số lần gập bụng, lực kẹp tay, nhảy xa, huyết áp là kiểu numeric hoặc integer.
- Các biến về giới tính, phân lớp hiệu suất là kiểu categorical.

Bộ dữ liệu không chứa giá trị thiếu trong các bản ghi, giúp đảm bảo tính toàn vẹn

và độ chính xác trong việc phân tích và mô hình hóa dữ liệu.

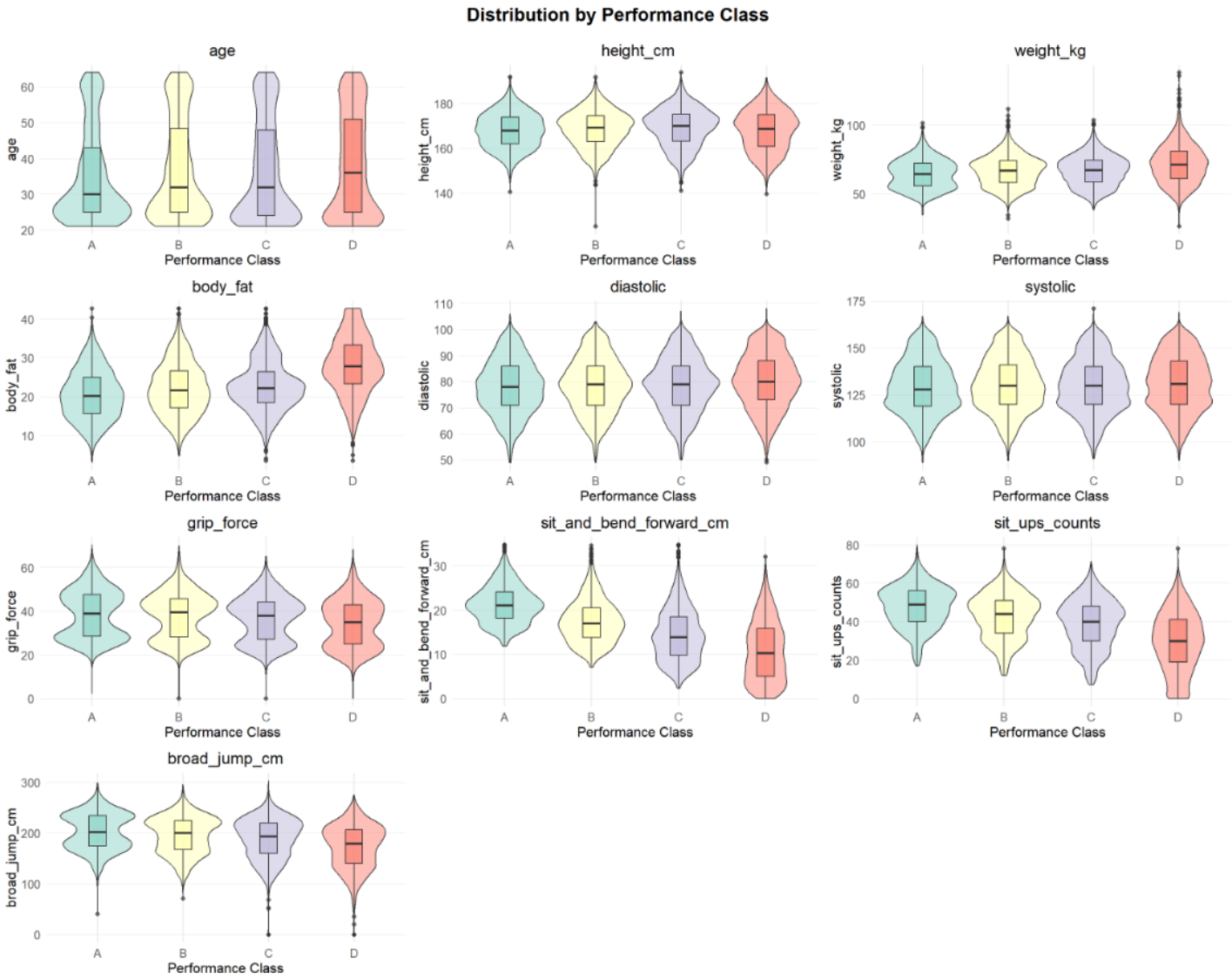
Sự phân phố của các biến trong dữ liệu

Các biến trong bộ dữ liệu có sự phân bố khá đa dạng. Ví dụ, độ tuổi của người tham gia phân bổ từ 20 đến 64, với sự phân bố tương đối đều giữa các nhóm tuổi. Cân nặng, chiều cao và tỷ lệ mỡ cơ thể có sự biến động lớn, phản ánh sự khác biệt về thể chất giữa các cá nhân. Đặc biệt, các bài kiểm tra thể lực như nhảy xa và gập bụng cho thấy sự phân bố rộng về khả năng thể lực của các cá nhân tham gia, với nhóm có hiệu suất cao (lớp A) và nhóm có hiệu suất thấp (lớp D) rõ rệt. Sau khi làm sạch dữ liệu và loại bỏ một số giá trị ngoại lai, ta có thể biểu diễn phân phối của dữ liệu và so sánh các nhóm hiệu suất với nhau như sau:



- **age**: có sự phân bố không đồng đều, chủ yếu tập trung ở độ tuổi từ 20-40, giảm dần ở các độ tuổi cao hơn. Ở độ tuổi 60 bắt đầu có dấu hiệu xuất hiện lại nhiều hơn.
- **body_fat**: Phân bố khá cân đối, dạng gần giống phân phối chuẩn. Đỉnh tập trung vào khoảng 20-30 (người tham gia.
- **broad_jump_cm**: Phân bố khá đối xứng, tập trung nhiều ở khoảng 150-250 (cm). Đồ thị hơi lệch trái chứng tỏ có thể có vài giá trị ngoại lai thấp hơn mức xu hướng chung.
- **diastolic**: Phân bố khá cân đối, tập trung ở khoảng 70-90 (mmHg), đây là mức huyết áp bình thường.
- **grip_force**: Phân bố dạng hai đỉnh (bimodal), có thể là do sự khác biệt giữa hai giới tính hoặc nhóm tuổi. Hai đỉnh nằm trong khoảng 20-40 và 50-60.
- **height_cm**: Phân bố dạng chuông, khá đồng đều, không có giá trị bất thường nào rõ rệt, tập trung ở khoảng 160-180 (cm).
- **sit_and_bend_forward_cm**: Phân bố đối xứng, tập trung ở khoảng 10-30 (cm). Điều này cho thấy sự linh hoạt trung bình trong việc tập thể dục của những người tham gia.
- **sit_ups_counts**: Phân bố gần chuẩn, tập trung ở khoảng 30-50 (lần).
- **systolic**: Phân bố khá cân đối, tập trung trong khoảng 120-140 (mmHg), đây là mức huyết áp phổ biến.
- **weight_kg**: Tập trung ở khoảng 50-80 (kg). Phân bố hơi lệch phải, có thể có một số giá trị ngoại lai cao hơn mức phổ biến bình thường.

So sánh hiệu suất giữa các biến



- age: Nhóm D có tuổi trung bình cao nhất, khoảng 35-45 tuổi, trong khi đó nhóm A lại là nhóm có số tuổi trung bình thấp nhất, khoảng dưới 30-40. Điều này là dễ hiểu vì thực tế, người trẻ thường khoẻ hơn người lớn tuổi.
- body_fat: Có xu hướng tăng dần từ nhóm A đến D, trong đó nhóm D có chỉ số mỡ cao nhất, khoảng 30-40 (%), còn nhóm A có chỉ số mỡ thấp nhất, khoảng 20-25 (
- broad_jump_cm: Giảm nhẹ dần từ nhóm A đến D. Phần lớn các giá trị của các nhóm nằm trong khoảng 150-200 (cm). Có một số giá trị ngoại lai thấp hơn mức phân bố chung ở tất cả các nhóm.

- diastolic: Khá tương đồng giữa các nhóm, nhóm D cao hơn những nhóm còn lại một chút. Đây có thể là vì những người có sức khoẻ thấp hơn sẽ có huyết áp cao hơn. Đó cũng là dấu hiệu cho thấy đặc trưng này có vẻ sẽ phân biệt tốt nhóm D so với các nhóm khác. Hầu hết các giá trị nằm trong khoảng 70-90 (mmHg) và nhóm D có một số giá trị ngoại lai thấp.
- grip_force: Xu hướng giảm nhẹ từ nhóm A đến D, trong đó nhóm A có lực nắm tay cao nhất. Độ phân tán khá tương đồng giữa các nhóm.
- height_cm: Khá đồng đều giữa các nhóm, hầu hết nằm trong khoảng 160-180cm và có một số giá trị ngoại lai ở tất cả các nhóm. Điều này cho thấy đây sẽ là một đặc trưng khó để phân biệt các nhóm với nhau.
- sit_and_bend_forward_cm: Giảm dần từ nhóm A đến D trong đó nhóm A có độ co giãn người linh hoạt tốt nhất còn nhóm D là thấp nhất. Dễ dàng thấy độ khác nhau của đặc trưng này giữa các biến là khá lớn, đặc biệt là ở nhóm A và D, chứng tỏ đây là một đặc trưng quan trọng ảnh hưởng đến sự chính xác khi phân loại các nhóm hiệu suất.
- sit_ups_counts: Tương tự như độ co giãn người linh hoạt (sit_and_bend_forward_cm), số lần gập bụng có xu hướng giảm từ nhóm A đến D. Sự khác biệt giữa các nhóm với nhau tuy không đáng kể bằng sit_and_bend_forward_cm nhưng vẫn rõ ràng để phân biệt tốt.
- systolic: Khá tương đồng giữa các nhóm, hầu hết trong khoảng 120-140 (mmHg), độ phân tán tương đối đồng đều.
- weight_kg: Tăng nhẹ từ nhóm A đến D, nhóm D có cân nặng trung bình cao nhất. Điều này có thể giúp phân biệt tốt nhóm D với các nhóm còn lại. Có nhiều giá trị ngoại lai cao hơn mức chung ở tất cả các nhóm.

Thống kê mô tả

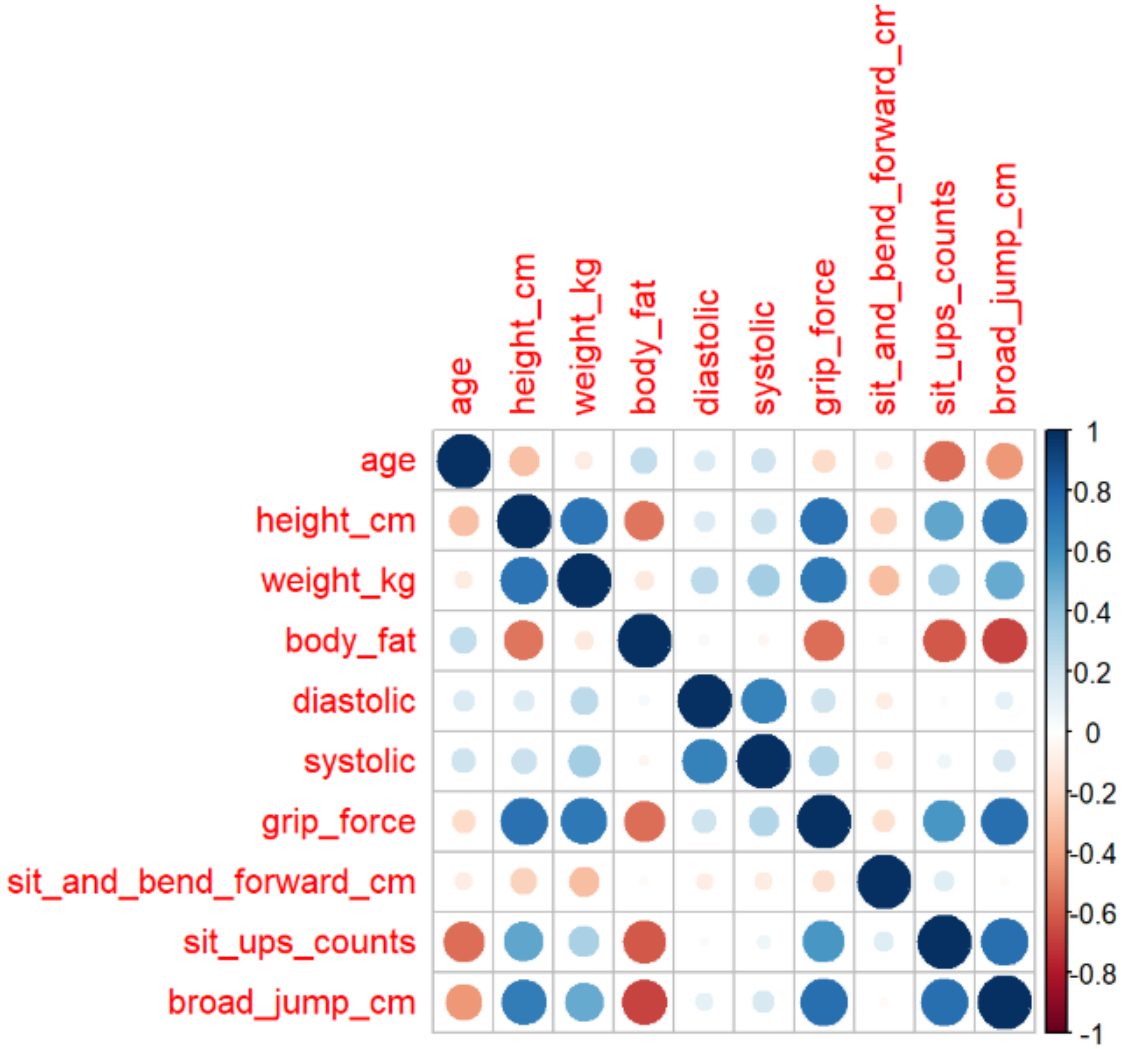
Ta tính các số liệu thống kê như giá trị nhỏ nhất, giá trị lớn nhất, trung vị, trung bình và độ lệch chuẩn, sau đó rút ra bảng thống kê đơn giản:

Biến	GTNN	GTLN	TV	TB	DLC
age	21.00	64.0	32.0	36.8	13.63
body_fat	3.05	42.7	22.6	23.1	7.17
broad_jump_cm	0.00	303.0	194.0	190.8	39.87
diastolic	49.00	108.0	79.0	78.8	10.42
grip_force	0.00	70.5	38.0	37.0	10.69
height_cm	125.00	193.8	169.0	168.4	8.42
sit_and_bend_forward_cm	0.00	34.8	16.6	16.2	6.62
sit_ups_counts	0.00	80.0	42.0	40.2	14.19
systolic	89.00	171.0	130.0	130.2	14.38
weight_kg	26.30	138.1	67.2	67.1	11.78

Bảng 1: Bảng thống kê mô tả

Phân tích dữ liệu

Sự tương quan của bộ dữ liệu



Tương quan dương mạnh:

1. height_cm và weight_kg: những người có chiều cao càng cao thì có xu hướng nặng hơn.
2. body_fat và weight_kg: lượng chất béo của cơ thể cao thì cân nặng cũng cao.
3. systolic và diastolic: Sự tương quan dương mạnh giữa hai đặc trưng này thể hiện rằng huyết áp tâm trương càng cao thì huyết áp tâm thu cũng càng cao.

Tương quan dương vừa:

1. grip và height_cm: những người có chiều cao cao hơn thì có vẻ lực nắm cũng cao hơn.
2. grip và weight_kg: tương tự, những người nặng hơn thì lực nắm cũng có vẻ cao hơn.

Tương quan yếu hoặc không có tương quan: age và các biến khác: sự tương quan giữa độ tuổi và đa số các biến khác là không rõ ràng, nghĩa là tuổi không phải là một đặc trưng ảnh hưởng quá lớn đến những đặc điểm thể chất khác trong bộ dữ liệu này.
Tương quan âm: age và sit_and_bend_forward_cm: có sự tương quan âm yếu, chứng tỏ rằng một số người càng lớn tuổi thì càng khó ngồi gập người về phía trước.

So sánh sự khác biệt giữa các nhóm

Ta sẽ sử dụng phương pháp ANOVA để so sánh sự khác nhau của các biến định lượng giữa các nhóm với nhau.

Biến	Df (Nhóm)	Sum Sq (Nhóm)	Mean Sq (Nhóm)	Tổng phần dư	p-value	Kết luận
age	3	15250	5083.3	2337658	$< 2.2 \times 10^{-16}$	Khác biệt đáng kể
height_cm	3	3728	1242.55	894578	$< 2.2 \times 10^{-16}$	Khác biệt đáng kể
weight_kg	3	78306	26102.0	1679470	$< 2.2 \times 10^{-16}$	Khác biệt đáng kể
body_fat	3	92861	30954.0	557780	$< 2.2 \times 10^{-16}$	Khác biệt đáng kể
diastolic	3	8327	2775.67	1368383	$< 2.2 \times 10^{-16}$	Khác biệt đáng kể
systolic	3	6651	2216.99	2612349	$< 2.2 \times 10^{-16}$	Khác biệt đáng kể
grip_force	3	30441	10147.0	713688	$< 2.2 \times 10^{-16}$	Khác biệt đáng kể
sit_and_bend_forward_cm	3	179828	59943.0	375874	$< 2.2 \times 10^{-16}$	Khác biệt đáng kể
sit_ups_counts	3	529594	176531.0	2020013	$< 2.2 \times 10^{-16}$	Khác biệt đáng kể
broad_jump_cm	3	1414897	471632.0	18722617	$< 2.2 \times 10^{-16}$	Khác biệt đáng kể

Bảng 2: Bảng kết quả phân tích phương sai (ANOVA)

Nhận xét:

1. age: Giá trị p $<$ 2.2e-16 (rất nhỏ), chứng tỏ sự khác biệt đáng kể giữa các nhóm trong biến age. Tổng bình phương phần dư cao (2337658) cho thấy sự phân tán trong dữ liệu.
2. height_cm: Giá trị p $<$ 2.2e-16, chỉ ra sự khác biệt đáng kể giữa các nhóm về chiều cao. Tổng bình phương phần dư (894578) cho thấy mức độ biến thiên trong chiều cao không lớn.
3. weight_kg: Giá trị p $<$ 2.2e-16, cho thấy sự khác biệt có ý nghĩa giữa các nhóm trong cân nặng. Tổng bình phương phần dư (1679470) cho thấy sự phân tán trong dữ liệu cân nặng.
4. body_fat: Giá trị p $<$ 2.2e-16, chỉ ra rằng lượng mỡ cơ thể có sự khác biệt rõ rệt giữa các nhóm. Tổng bình phương phần dư (557780) nhỏ hơn các biến khác, cho thấy sự phân tán thấp hơn.
5. diastolic: Giá trị p $<$ 2.2e-16, chỉ ra sự khác biệt có ý nghĩa giữa các nhóm trong huyết áp tâm trương. Tổng bình phương phần dư (1368336) phản ánh mức độ biến thiên.
6. systolic: Giá trị p $<$ 2.2e-16, cho thấy sự khác biệt đáng kể về huyết áp tâm thu giữa các nhóm.
7. grip_force: Giá trị p $<$ 2.2e-16, chứng minh rằng lực bóp tay khác nhau đáng kể giữa các nhóm.
8. sit_and_bend_forward_cm: Giá trị p $<$ 2.2e-16, cho thấy sự khác biệt đáng kể giữa các nhóm về độ linh hoạt khi ngồi gập người về phía trước.
9. sit_ups_counts: Giá trị p $<$ 2.2e-16, phản ánh sự khác biệt đáng kể về số lần gập bụng giữa các nhóm.
10. broad_jump_cm: Giá trị p $<$ 2.2e-16, chỉ ra rằng khả năng nhảy xa có sự khác biệt rõ rệt giữa các nhóm.

Kết luận:

1. Tất cả p-value đều rất nhỏ.
2. Phân tích theo nhóm chỉ số

- Chỉ số sức mạnh (grip_force, broad_jump_cm):
  - Có sự phân biệt rõ rệt giữa các class.
  - Class A và B thể hiện ưu thế vượt trội.
- Chỉ số sức bền (sit_ups_counts):
  - Sự khác biệt tăng dần theo class.
  - Class A thể hiện khả năng vượt trội đáng kể.
- Chỉ số thể trạng (body_fat, weight_kg):
  - Có xu hướng giảm dần từ class D đến A.
  - Class A có tỷ lệ mỡ cơ thể thấp nhất.

3. Ý nghĩa:
   - Các chỉ số thể chất đều có vai trò quan trọng trong việc phân loại hiệu suất.
   - Sức mạnh và sức bền là yếu tố phân biệt rõ rệt nhất giữa các class.
   - Thể trạng (đặc biệt là tỷ lệ mỡ cơ thể) có ảnh hưởng đáng kể đến hiệu suất.
   - Các chỉ số sức khỏe (huyết áp) tuy ít ảnh hưởng hơn nhưng vẫn có ý nghĩa thống kê.

Phân tích đặc điểm nhóm hiệu suất cao (Class A)

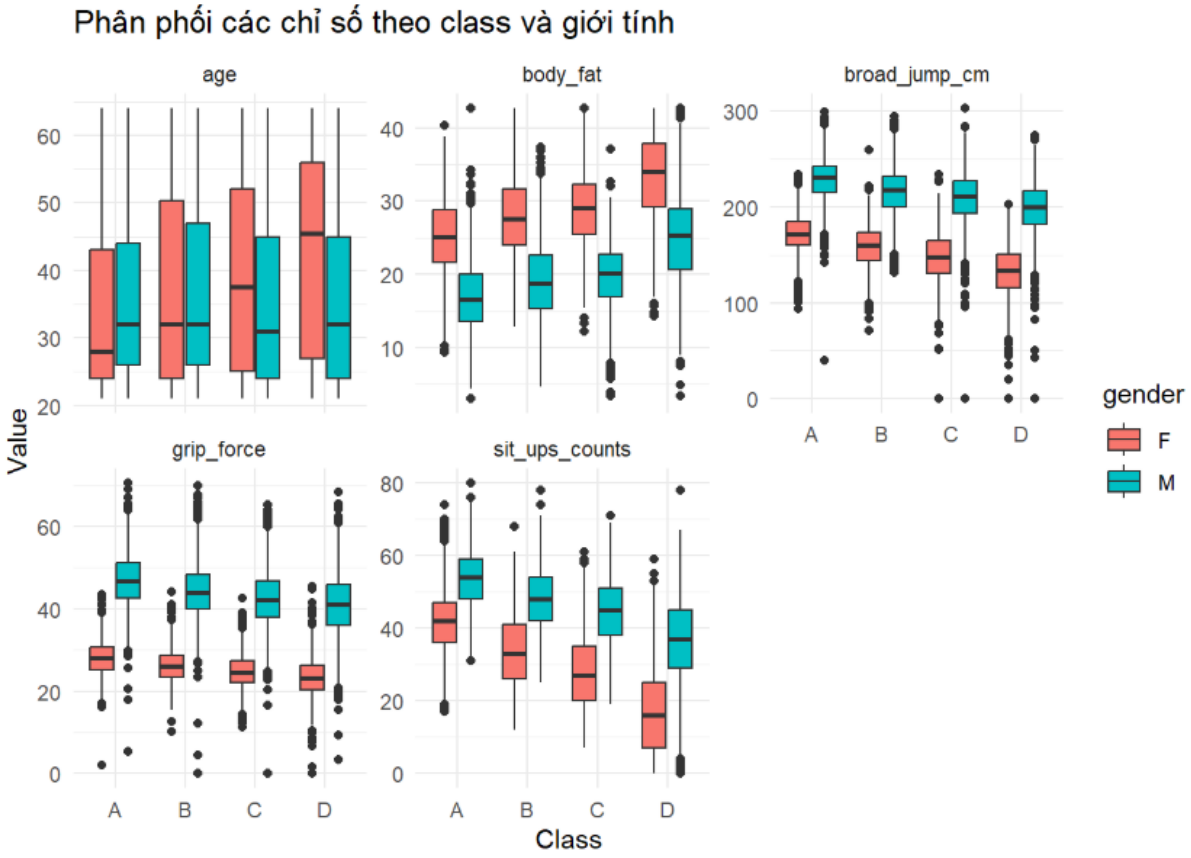
Ở đây, ta so sánh trung bình các đặc trưng của nhóm A so với các nhóm khác. Có thể thấy nhóm A cao hơn các nhóm khác ở gần như mọi đặc trưng, chỉ có đặc trưng thấp hơn là tuổi và lượng chất béo trong cơ thể. Điều này là dễ hiểu vì lớp có hiệu suất cao, sức khoẻ tốt thì lượng chất béo ít hơn và tuổi cũng sẽ trẻ hơn.

Phân tích ảnh hưởng của giới tính đến hiệu suất

Permutation ANOVA Hai Chiều: Kiểm tra sự tương tác giữa giới tính và class đối với các biến quan trọng (important_vars). Điều này giúp xác định xem sự khác biệt giữa các lớp hiệu suất có phụ thuộc vào giới tính hay không. Giúp hiểu sâu hơn về mối quan hệ phức tạp giữa giới tính và hiệu suất, cung cấp cơ sở khoa học để thiết kế chương trình tập luyện phù hợp.

Biến	Thành phần	Df	R Sum Sq	R Mean Sq	Pr(Prob)
Age	gender1	1	13822	13822.4	< 2.2e-16
	class1	3	25025	8341.6	< 2.2e-16
	gender1:class1	3	36706	12235.4	< 2.2e-16
	Residuals	12663	2290791	180.9	
Body Fat	gender1	1	220348	220348	< 2.2e-16
	class1	3	96500	32167	< 2.2e-16
	gender1:class1	3	384	128	< 2.2e-16
	Residuals	12663	335356	26	
Grip Force	gender1	1	953074	953074	< 2.2e-16
	class1	3	45969	15323	< 2.2e-16
	gender1:class1	3	607	202	< 2.2e-16
	Residuals	12663	451887	36	
Sit Ups Counts	gender1	1	699735	699735	< 2.2e-16
	class1	3	608187	202729	< 2.2e-16
	gender1:class1	3	26115	8705	< 2.2e-16
	Residuals	12663	1312700	104	
Broad Jump (cm)	gender1	1	10838491	10838491	< 2.2e-16
	class1	3	1833902	611301	< 2.2e-16
	gender1:class1	3	46437	15479	< 2.2e-16
	Residuals	12663	7869750	621	

So sánh trực quan các đặc trưng giữa các nhóm và giới tính:



Từ kết quả phân tích Permutation ANOVA và box plot so sánh, chúng ta có thể rút ra một số kết luận quan trọng:

1. Vai trò của giới tính trong hiệu suất thể chất:
   - Giới tính là yếu tố có ảnh hưởng mạnh mẽ nhất đến tất cả các chỉ số thể chất.
   - Tất cả các p-value đều nhỏ hơn 2.2e-16, cho thấy sự khác biệt giữa nam và nữ là rất có ý nghĩa thống kê.
   - Nam giới thường có ưu thế về mặt thể chất, đặc biệt trong các chỉ số như lực nắm tay và nhảy xa.
2. Sự phân hóa theo class hiệu suất:
   - Có sự khác biệt rõ rệt giữa các class về mọi chỉ số thể chất.
   - Class càng cao, khoảng cách về thể lực giữa nam và nữ càng lớn.
   - Class A nổi bật với các chỉ số thể lực vượt trội và tỷ lệ mỡ cơ thể thấp.
3. Mối tương tác giữa giới tính và class:
   - Sự khác biệt nam-nữ không đồng đều giữa các class.
   - Ở class càng cao, sự chênh lệch về thể lực giữa nam và nữ càng rõ rệt.
   - Điều này gợi ý rằng nam và nữ có thể cần các chương trình tập luyện khác nhau để phát triển tối ưu.
4. Ý nghĩa thực tiễn:
   - Cần xây dựng chương trình tập luyện riêng biệt cho nam và nữ.

Mô hình phân loại

Chuẩn bị dữ liệu

Đầu tiên, ta sử dụng phương pháp tập xác thực để chia dữ liệu thành hai phần là **tập huấn luyện** (training set) và **tập kiểm tra** (test set). Sau đó, vì ta đã loại bỏ một số giá trị ngoại lai bất thường nên dữ liệu huấn luyện ở các nhóm có sự mất cân bằng, ta sử dụng phương pháp SMOTE để giải quyết.

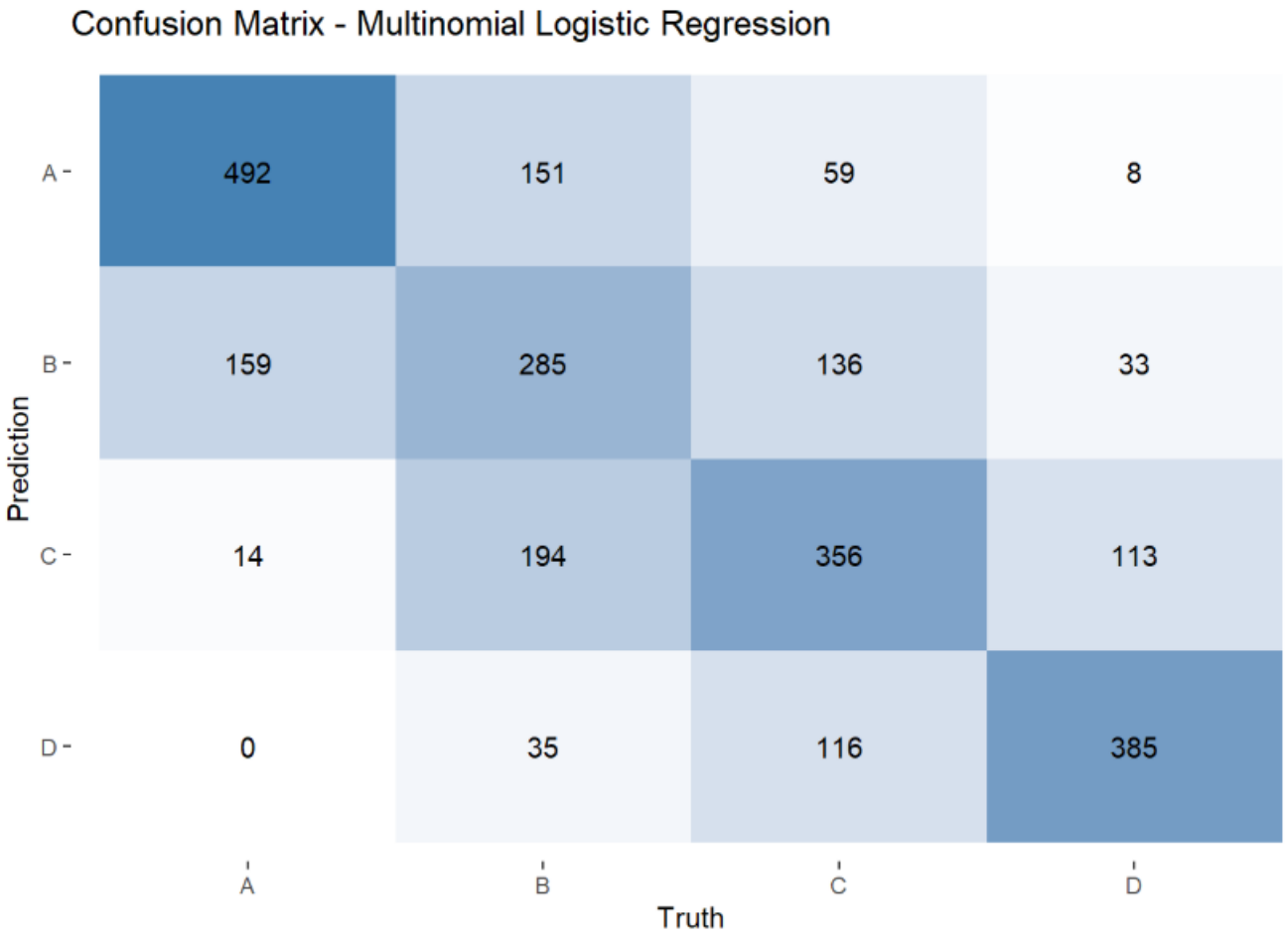
Xây dựng và đánh giá các mô hình phân loại

Trong dự án này, ta sẽ thử nghiệm ba mô hình phân loại là Multinomial Logistic Regression, Random Forest và XGBoost.
Đối với mô hình Multinomial Logistic Regression, ta thử nghiệm với dữ luyện train ban đầu đã chuẩn bị và sau khi biến đổi tương tác, log, square.

Multinomial Logistic Regression: Các chỉ số đánh giá và Confusion matrix:

Metric	A	B	C	D
Precision	0.7398496	0.4285714	0.5337331	0.7142857
Recall	0.6929577	0.4649266	0.5258493	0.7182836
Additional Metrics				
Accuracy	0.5985804			
Kappa	0.4633906			
Macro F1	0.6023017			

Bảng 3: Evaluation Metrics



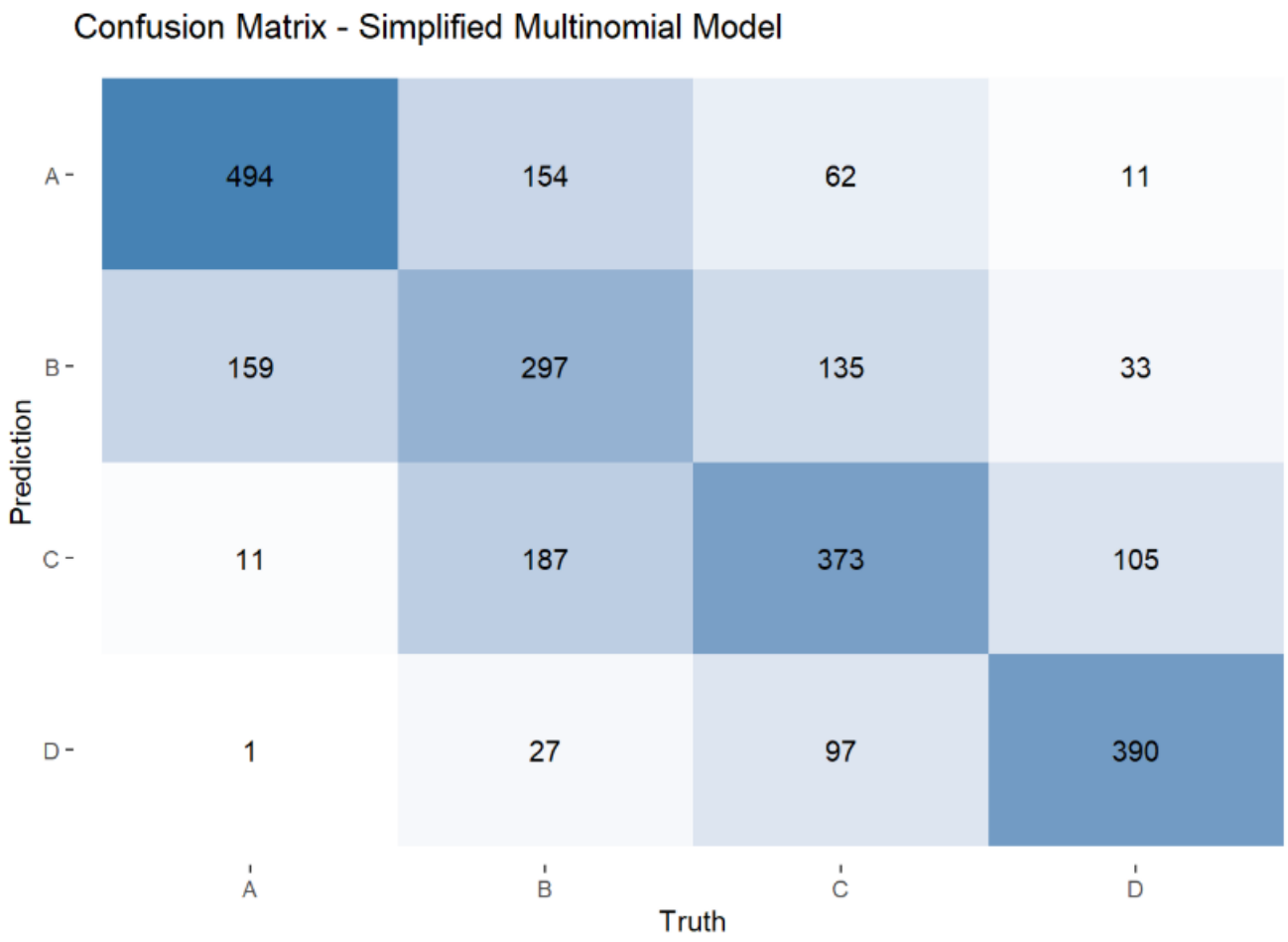
Nhận xét: Nhìn chung, với độ chính xác chỉ khoảng 60%, mô hình không thực sự tốt. Nó có thể phân biệt đúng nhóm A và D với độ chính xác hơn 70% nhưng ở nhóm

B và C chỉ có trên dưới 50%, nhầm lẫn khá nhiều. Confusion matrix ở trên biểu thị sự nhầm lẫn khá nhiều ở hai nhóm giữa, B và C.

Multinomial Logistic Regression - simplified: Các chỉ số đánh giá và Confusion matrix:

Metric	**A**	**B**	**C**	**D**
Precision	0.7428571	0.4466165	0.5592204	0.7235622
Recall	0.6851595	0.4759615	0.5517751	0.7572816
Additional Metrics				
Accuracy	0.612776			
Kappa	0.4820824			
Macro F1	0.6178041			

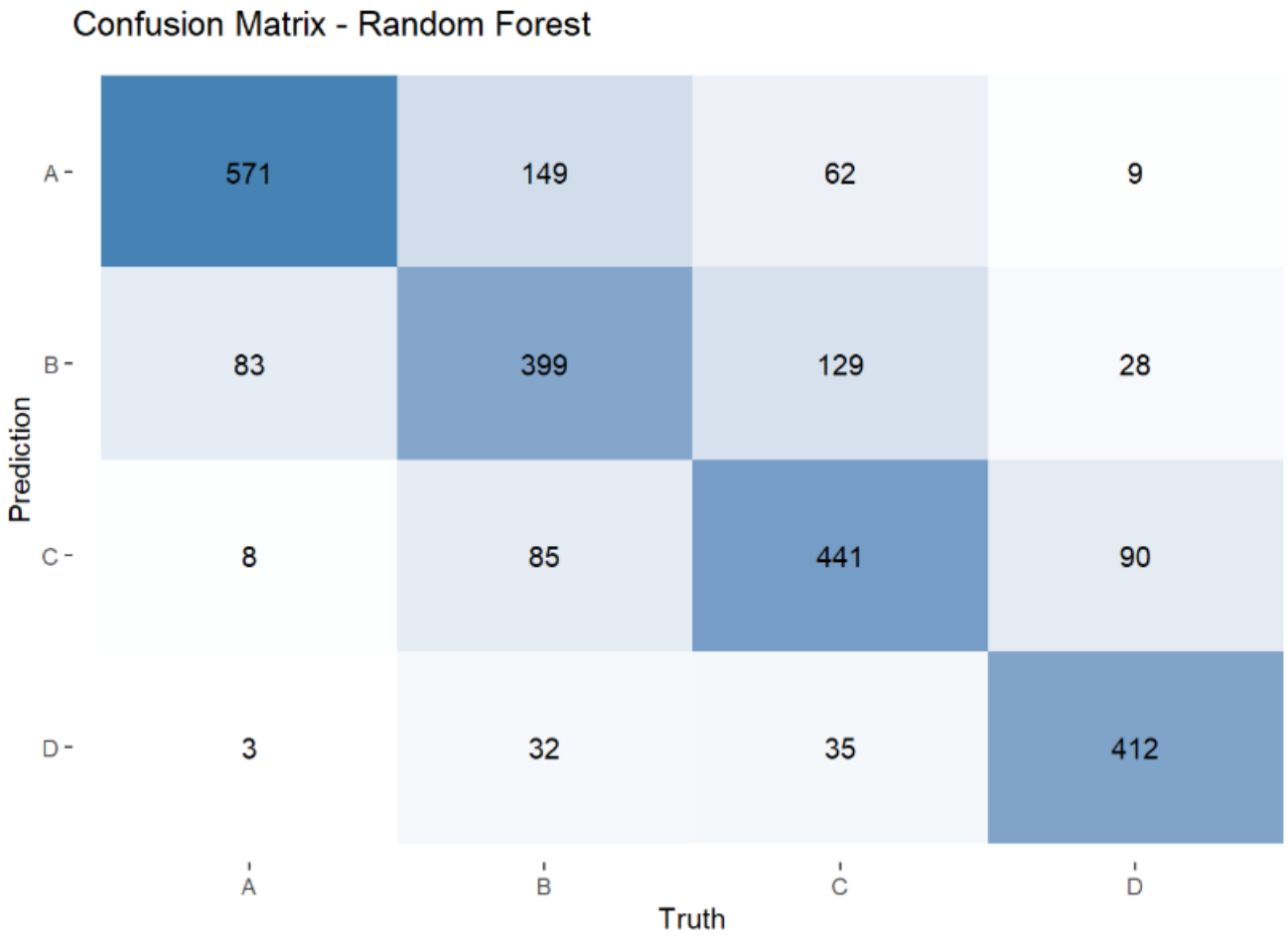
Bảng 4: Evaluation Metrics



Random Forest: Các chỉ số đánh giá và Confusion matrix:

Metric	A	B	C	D
Precision	0.8586466	0.6000000	0.6611694	0.7643785
Recall	0.7218710	0.6244131	0.7067308	0.8547718
Additional Metrics				
Accuracy	0.7188486			
Kappa	0.6236386			
Macro F1	0.7239856			

Bảng 5: Evaluation Metrics



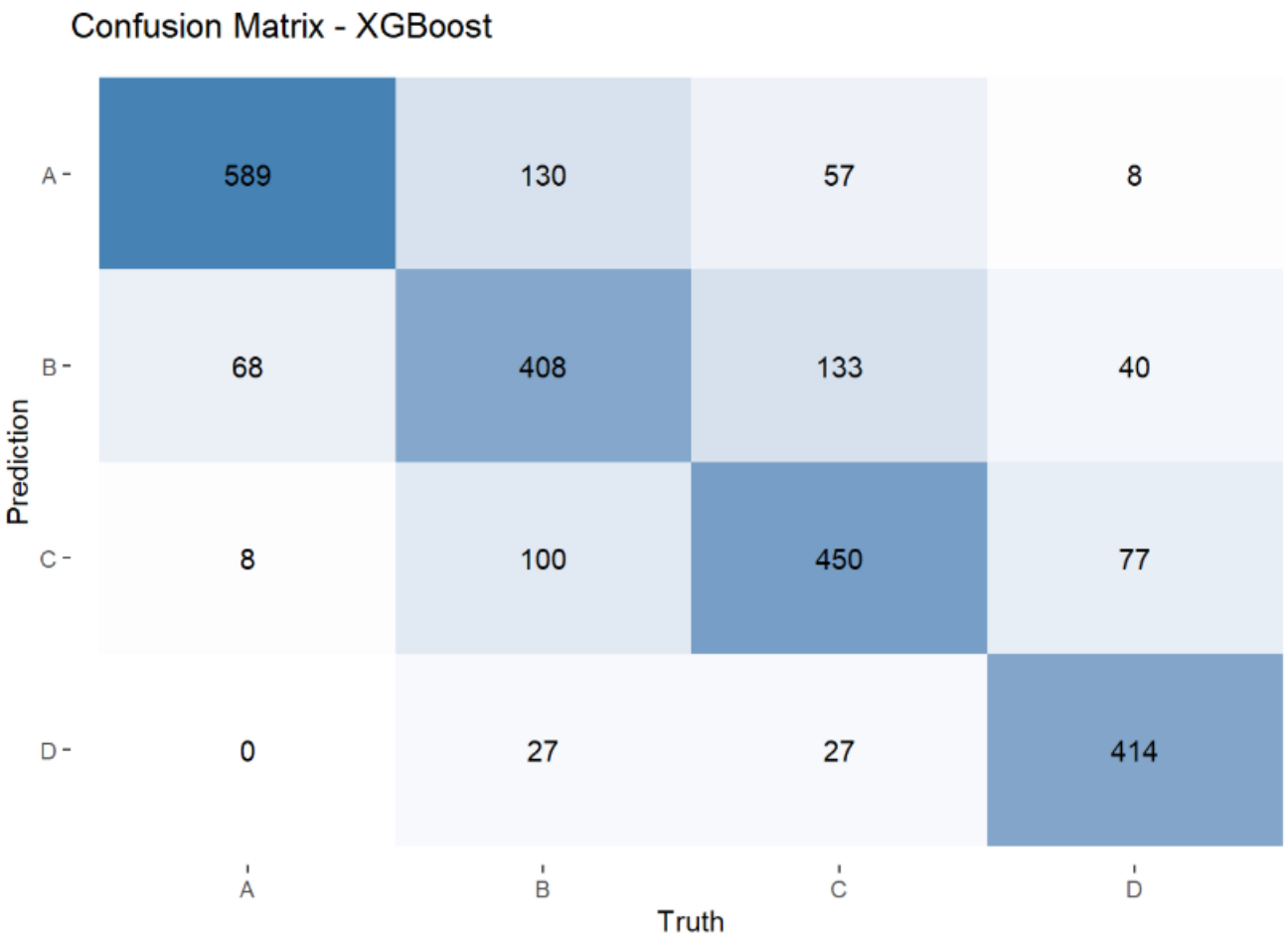
Nhận xét: Đến mô hình Random Forest, có thể thấy độ chính xác đã được cải thiện khá đáng kể,lên đến trên 70%. Trong đó, lớp A có độ chính xác khi phân biệt cao

nhất, trên 80%. Hai lớp A và C mặc dù có khả năng phân biệt đúng còn thấp nhưng đã được cải thiện lên hơn 60%, có thể nói là khá tốt. Từ confusion matrix có thể thấy sự nhầm lẫn giữa các lớp đã ít hơn.

XGBoost (Extreme Gradient Boosting): Các chỉ số đánh giá và Confusion matrix:

Metric	**A**	**B**	**C**	**D**
Precision	0.8857143	0.6135338	0.6746627	0.7680891
Recall	0.7512755	0.6286595	0.7086614	0.8846154
Additional Metrics				
Accuracy		0.7338328		
Kappa		0.6435647		
Macro F1		0.7393809		

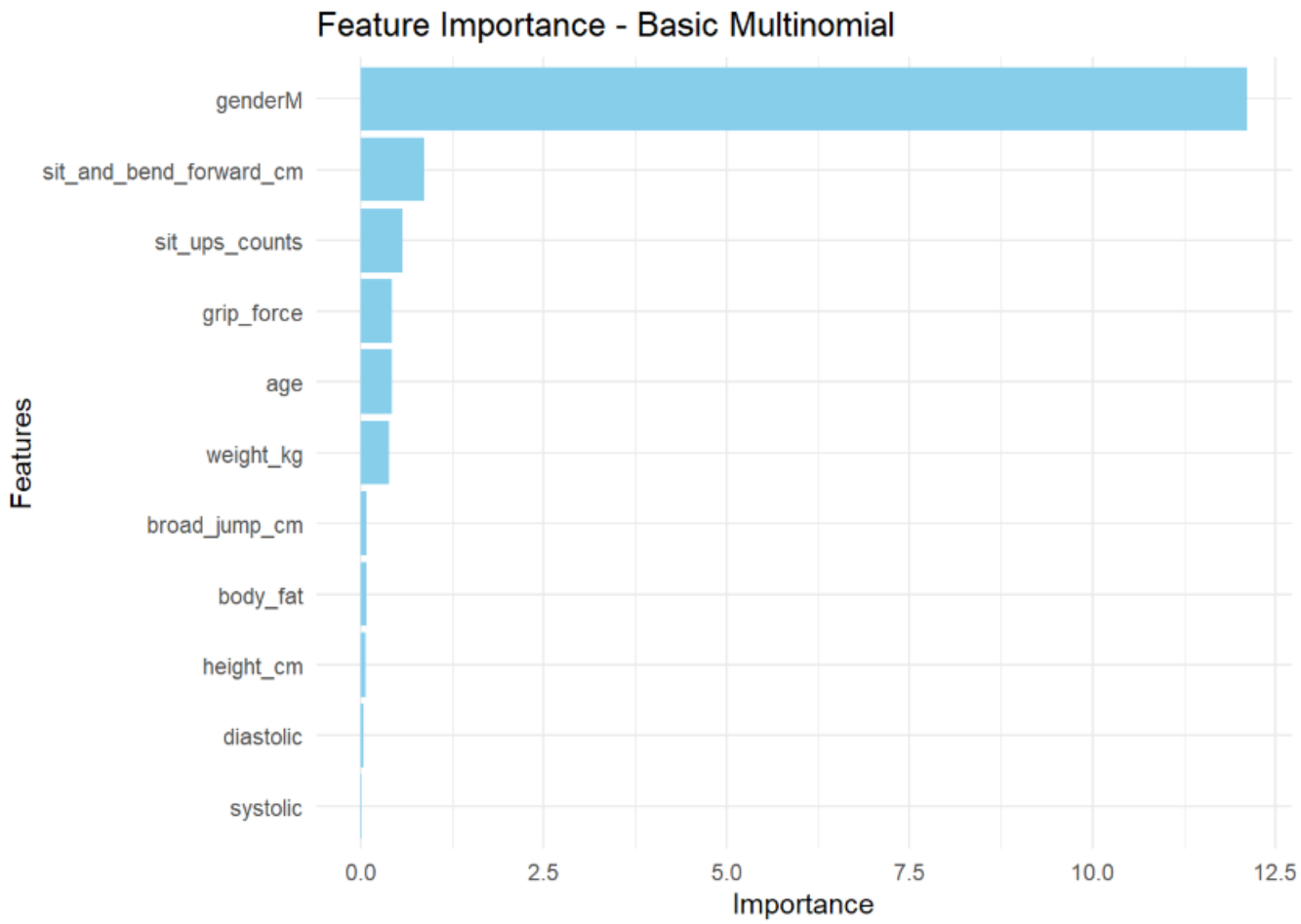
Bảng 6: Evaluation Metrics



Nhận xét: Mô hình này có sự cải thiện so với Random Forest. Cụ thể, lớp A có khả năng phân biệt đúng cao nhất, gần 90%. Lớp B và C cũng đã được cải thiện khá đáng kể. Có thể thấy từ confusion matrix là mặc dù lớp B và C khi phân biệt vẫn còn nhầm lẫn với nhau, lớp B còn bị nhầm với lớp A nữa nhưng đã được cải thiện rất tốt so với mô hình đầu tiên.

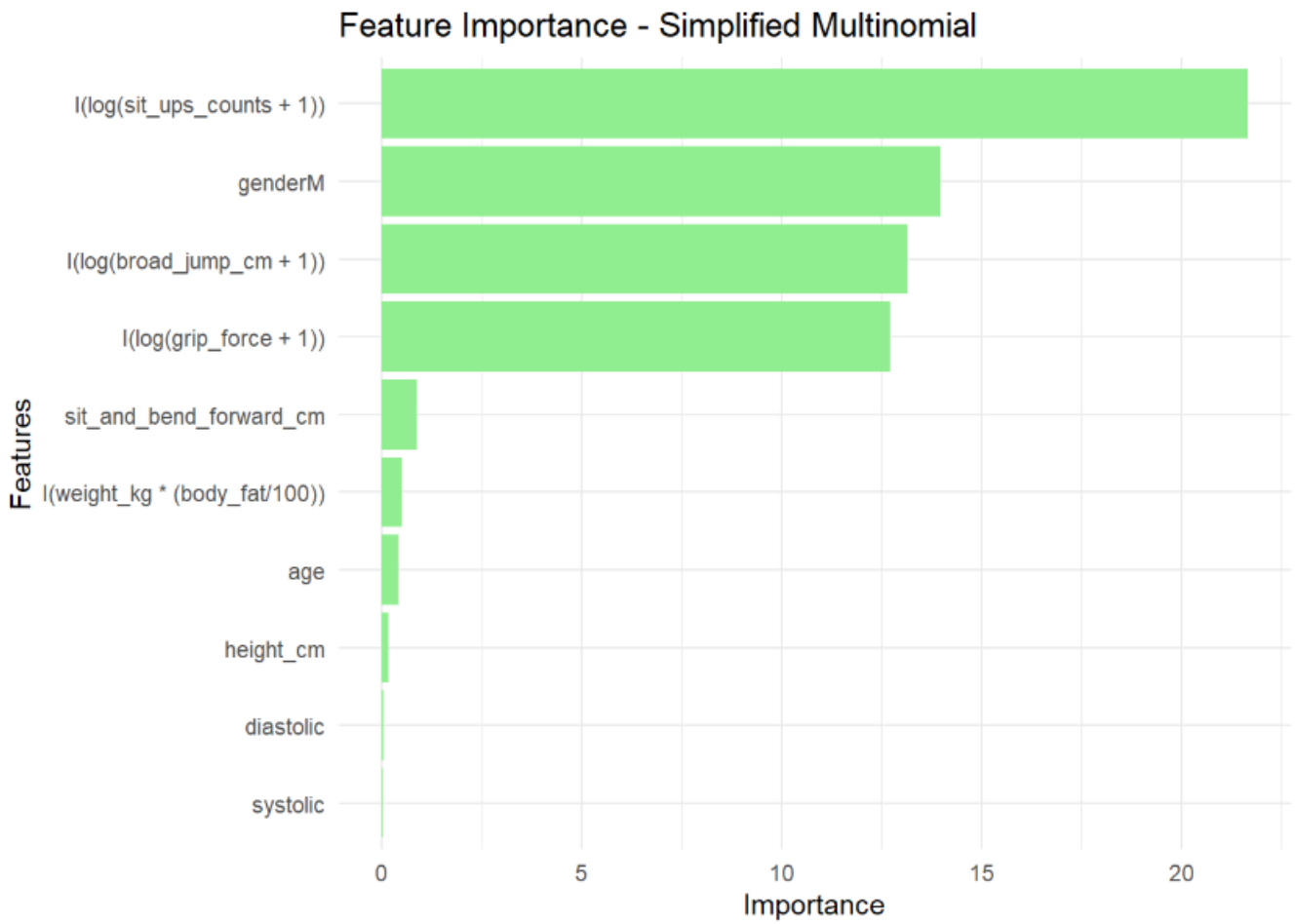
Các đặc trưng quan trọng trong từng mô hình và chỉ số

Multinomial Logistic Regression:



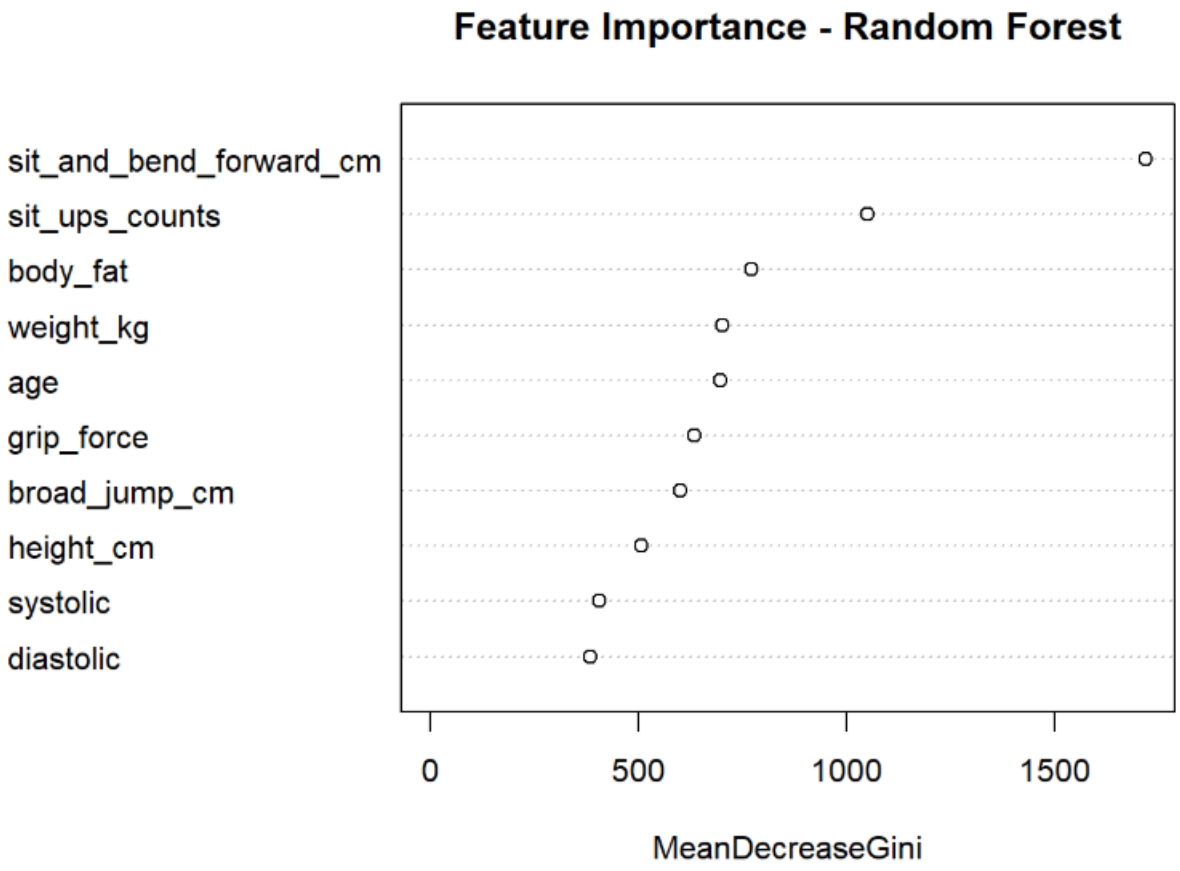
Theo mô hình, đặc trưng giới tính là quan trọng nhất, hơn hẳn so với các đặc trưng khác. Mặc dù trên thực tế, giới tính không phải là yếu tố duy nhất quyết định đến hiệu suất luyện tập và sức khoẻ mà phải có sự góp mặt không hề nhỏ của các đặc trưng khác thì mới có thể phân loại hiệu quả được. Đây có thể lý do khiến mô hình này có hiệu quả không được tốt.

Multinomial Logistic Regression - simplified:



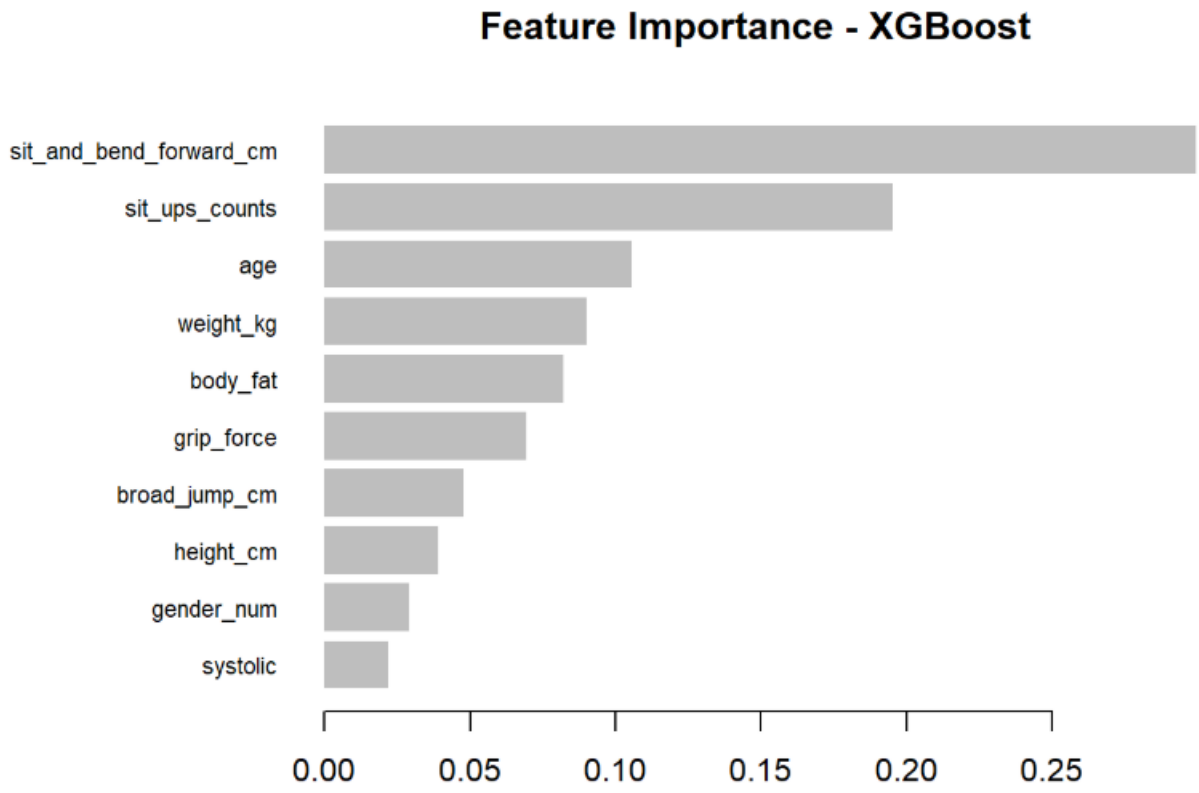
Sau khi thực hiện biến đổi tương tác, log, square, một số đặc trưng khác như số lần gập bụng, nhảy xa, lực nắm đã được nâng độ quan trọng trong việc huấn luyện mô hình, đặc biệt là số lần gập bụng. Tuy nhiên, giới tính vẫn còn khá cao, đứng thứ hai, do đó mô hình mặc dù có cải thiện nhưng chưa đáng kể.

Random Forest:



Ở mô hình này, có thể thấy các đặc trưng đóng vai trò quan trọng trong việc huấn luyện đã được thay đổi đáng kể. Một số đặc trưng quan trọng trong phần phân tích, so sánh trước đó và trên thực tế đã tham gia nhiều hơn vào việc huấn luyện như ngồi gập bụng, lượng chất béo trong cơ thể cân nặng,...

XGBoost:



Đối với XGBoost, các đặc trưng đóng vai trò quan trọng trong việc huấn luyện khá giống với Random Forest, có một số đặc trưng còn được cải thiện độ quan trọng như số lần gập bụng, tuổi cân nặng,... vốn quan trọng để phân loại sức khoẻ trên thực tế.

Tổng quan hiệu quả về các mô hình xây dựng

Sau khi thử nghiệm với ba loại mô hình, ta rút ra được một số kết luận như sau:

1. **Multinomial Logistic Regression:** Có độ chính xác thấp nhất, các lớp dễ bị nhầm lẫn với nhau khi phân loại, cho thấy mô hình này không hiệu quả.
2. **Random Forest:** Mô hình đã cải thiện đáng kể khả năng phân loại hiệu suất. Mặc dù vẫn còn nhầm lẫn khá nhiều trong việc phân loại nhưng mức độ hiệu quả của mô hình đã được cải thiện.
3. **XGBoost** Mô hình có độ chính xác cao nhất trong các mô hình được thử nghiệm và đánh giá. Mặc dù vẫn còn nhiều nhầm lẫn khi phân biệt, đặc biệt là lớp B và C nhưng độ hiệu quả đã tốt hơn. Vậy trong các mô hình đã được thử nghiệm,

mô hình XGBoost (Extreme Gradient Boosting) là hiệu quả nhất.

Tổng kết

Bộ dữ liệu bodyPerformance.csv mang lại nhiều thuận lợi nhưng cũng đi kèm một số thách thức trong quá trình xử lý và xây dựng mô hình phân loại. Việc giải quyết các khó khăn này sẽ giúp nâng cao hiệu suất và giá trị ứng dụng của mô hình trong thực tế.

Dự án nghiên cứu và phân tích dữ liệu Body Performance Data đã được hoàn thiện với nhiều kết quả đáng khích lệ, mang lại những đóng góp quan trọng cả về mặt học thuật lẫn thực tiễn.

Thuận lợi

1. Dễ dàng truy cập và xử lý: Tập dữ liệu bodyPerformance.csv đã được tiền xử lý, giúp giảm thiểu công sức và tập trung nhanh vào các bước phát triển mô hình.
2. Cơ hội sử dụng nhiều thuật toán: Tập dữ liệu phù hợp cho nhiều thuật toán khác nhau như Logistic Regression, Random Forest, Support Vector Machines, v.v.
3. Hỗ trợ trực quan hóa: Dữ liệu với nhiều thuộc tính dễ dàng được trực quan hóa qua các biểu đồ như scatter plot, heatmap, box plot, giúp hiểu rõ hơn về dữ liệu.
4. Đa dạng đặc trưng: Các đặc trưng như tuổi, giới tính, chiều cao, cân nặng, tỷ lệ mỡ cơ thể, huyết áp, lực nắm, gập bụng, nhảy xa cung cấp lượng thông tin phong phú.
5. Kích thước mẫu lớn: Với 13.393 mẫu, bộ dữ liệu đủ lớn để huấn luyện và kiểm tra mô hình, tăng độ chính xác và khả năng tổng quát hóa.
6. Thông tin thể chất đầy đủ: Bộ dữ liệu chứa các đặc điểm thể chất phong phú, hỗ trợ mô hình phân loại hiệu quả.
7. Phân loại trực quan: Một số đặc trưng như tỷ lệ mỡ cơ thể hay số lần gập bụng liên quan trực tiếp đến hiệu suất, làm bài toán dễ hiểu và giải thích.
8. Phù hợp với nhiều mô hình: Bộ dữ liệu có thể sử dụng cho Logistic Regression, Random Forest, Gradient Boosting, hoặc Deep Learning.

9. Ứng dụng thực tế cao: Kết quả phân loại có thể được dùng để đánh giá thể chất, đề xuất bài tập luyện, hoặc hỗ trợ lĩnh vực y tế.
10. Kiểm tra hiệu suất mô hình: Dữ liệu lớn giúp chia nhỏ tập huấn luyện, kiểm tra, và xác thực để đánh giá chính xác hơn hiệu suất mô hình.

Khó khăn

1. Xử lý các dữ liệu outlier: Một số đặc trưng như huyết áp tâm trương (diastolic), huyết áp tâm thu (systolic), và độ linh hoạt khi ngồi và cúi về phía trước (sit and bend forward_cm) có thể xuất hiện các giá trị outlier bất hợp lý, gây ảnh hưởng đến hiệu suất và độ tin cậy của mô hình.
2. Mất cân bằng dữ liệu giữa các lớp: Sau khi xử lý các dữ liệu outlier, có thể xảy ra tình trạng mất cân bằng dữ liệu giữa các lớp phân loại, dẫn đến việc mô hình học không đồng đều và giảm hiệu suất trên các lớp nhỏ.
3. Mô hình logistic không hiệu quả: Mô hình Multinomial Logistic Regression không đạt hiệu quả cao trong bài toán phức tạp này, yêu cầu sử dụng các mô hình phi tuyến tính như Random Forest hoặc XGBoost để cải thiện hiệu suất.
4. Đánh giá mô hình: Do có nhiều lớp phân loại (A, B, C, D), việc đánh giá mô hình cần được thực hiện với các chỉ số phù hợp như F1-score, Precision, Recall cho từng lớp, thay vì chỉ dựa vào độ chính xác tổng thể (Accuracy).

Kết quả nổi bật

- **Hoàn thiện quy trình xử lý dữ liệu:**
  - Tiền xử lý dữ liệu được thực hiện kỹ lưỡng, bao gồm loại bỏ giá trị thiếu, dữ liệu trùng lặp và xử lý các giá trị ngoại lai bằng các phương pháp như IQR và winsorization.
  - Sử dụng phương pháp SMOTE để tái cân bằng dữ liệu giữa các lớp hiệu suất, cải thiện hiệu quả huấn luyện mô hình.
- **Phân tích khám phá dữ liệu (EDA):**
  - Thực hiện thống kê mô tả và phân tích tương quan để hiểu rõ các yếu tố quan trọng, ví dụ như mối quan hệ giữa chiều cao và cân nặng, hay tương quan âm yếu giữa tuổi tác và độ linh hoạt.

  - Phân tích các chỉ số như tuổi, tỷ lệ mỡ cơ thể, khả năng nhảy xa và số lần gập bụng, qua đó xác định các đặc trưng có ý nghĩa trong phân loại hiệu suất.

- **Xây dựng và đánh giá mô hình:**
  - Triển khai và so sánh ba mô hình: Multinomial Logistic Regression, Random Forest, và XGBoost.
  - Mô hình XGBoost đạt hiệu quả cao nhất với độ chính xác 73.3bật ở khả năng phân loại tốt giữa các nhóm hiệu suất.
  - Xác định các đặc trưng quan trọng như số lần gập bụng, tỷ lệ mỡ cơ thể và khả năng nhảy xa, góp phần cải thiện hiệu quả dự đoán của mô hình.