**BỘ GIÁO DỤC VÀ ĐÀO TẠO**

**TRƯỜNG ĐẠI HỌC NGUYỄN TẤT THÀNH**

**KHOA CÔNG NGHỆ THÔNG TIN**





**TIỂU LUẬN ĐỒ ÁN CHUYÊN NGÀNH KHOA HỌC DỮ LIỆU**

**ĐỀ TÀI:** **Phân Tích và Xử Lý Dữ Liệu Hiệu Suất Thể Thao: Thông Số Cân Nặng và Kết Quả Tập Luyện**

**Giảng viên hướng dẫn: Th.S DƯƠNG MINH TUẤN**

**Sinh viên thực hiện: PHAN THANH TÂM**

**MSSV: 2100008769**

**Chuyên ngành: Khoa học dữ liệu**

**Môn học: Đồ án chuyên ngành khoa học dữ liệu**

**Khóa: 2021**

**Tp.HCM, 24 tháng 06 năm 2024**

**BỘ GIÁO DỤC VÀ ĐÀO TẠO**

**TRƯỜNG ĐẠI HỌC NGUYỄN TẤT THÀNH**

**KHOA CÔNG NGHỆ THÔNG TIN**





**TIỂU LUẬN ĐỒ ÁN CHUYÊN NGÀNH KHOA HỌC DỮ LIỆU**

**ĐỀ TÀI: Phân Tích và Xử Lý Dữ Liệu Hiệu Suất Thể Thao: Thông Số Cân Nặng và Kết Quả Tập Luyện**

**Giảng viên hướng dẫn: Th. S. DƯƠNG MINH TUẤN**

**Sinh viên thực hiện: PHAN THANH TÂM**

**MSSV: 2100008769**

**Chuyên ngành: Khoa học dữ liệu**

**Môn học: Đồ án chuyên ngành khoa học dữ liệu**

**Khóa: 2021**

**Tp.HCM, 24 tháng 06 năm 2024**

# LỜI CẢM ƠN

Em muốn bày tỏ lòng biết ơn chân thành và trân trọng tới tất cả **quý Thầy Cô tại Khoa Công Nghệ Thông Tin, trường Đại học Nguyễn Tất Thành** đã dành thời gian và kiến thức quý báu để chia sẻ với chúng em trong suốt quá trình học tập và nghiên cứu tại trường.

Đặc biệt, em muốn gửi lời cảm ơn đặc biệt và sâu sắc nhất tới Thầy **Th.S Dương Minh Tuấn**, người đã ân cần hướng dẫn và truyền đạt cho em những kiến thức, kỹ năng cần thiết để hoàn thành dự án Phân Tích và Xử Lý Dữ Liệu Hiệu Suất Thể Thao: Thông Số Cân Nặng và Kết Quả Tập Luyện này.

Nhờ vào sự dạy bảo và sự hỗ trợ từ Thầy **Th.S Dương Minh Tuấn**, em đã không chỉ nắm vững về lĩnh vực này mà còn phát triển kỹ năng nghiên cứu và phân tích một cách tự tin hơn.

Lời cảm ơn này cũng là một lời tri ân sâu sắc đến tất cả các Thầy Cô đã luôn tạo điều kiện tốt nhất cho sự phát triển của chúng em, từ việc tài trợ tài liệu, hỗ trợ kỹ thuật cho đến sự khích lệ và định hướng đúng đắn trong quá trình nghiên cứu.

Xin được chân thành cảm ơn và chúc quý Thầy Cô ngày càng thành công và phồn thịnh trong sứ mệnh cao quý của mình.

Em xin chân thành cảm ơn.

TP.HCM, ngày 24 tháng 06 năm 2024

# LỜI MỞ ĐẦU

Trong kỷ nguyên số hóa hiện nay, dữ liệu không chỉ là một nguồn tài nguyên quý giá mà còn là chìa khóa để mở ra những hiểu biết sâu sắc và đổi mới trong nhiều lĩnh vực, bao gồm cả thể thao. Báo cáo này tập trung vào việc phân tích và xử lý dữ liệu hiệu suất thể thao, đặc biệt là các thông số liên quan đến cân nặng và kết quả tập luyện trong năm 2024, thời điểm chứng kiến sự phát triển mạnh mẽ và cạnh tranh khốc liệt trong ngành thể thao.

Chúng em đã thực hiện một nghiên cứu chi tiết để phân tích các yếu tố ảnh hưởng đến hiệu suất tập luyện, từ các thông số cơ bản như cân nặng đến các chỉ số thành tích quan trọng. Mục tiêu của báo cáo là cung cấp cái nhìn toàn diện về cách các yếu tố này tương tác và ảnh hưởng đến kết quả tập luyện.

Thông qua báo cáo này, chúng em hy vọng mang đến những thông tin giá trị không chỉ cho các vận động viên, huấn luyện viên, mà còn cho các tổ chức và cá nhân quan tâm đến việc tối ưu hóa hiệu suất thể thao. Chúng em tin rằng báo cáo sẽ là cơ sở cho những nghiên cứu tiếp theo và góp phần vào sự phát triển bền vững và hiệu quả trong lĩnh vực thể thao.

Và một lần nữa chúng em xin chân thành cảm ơn sự hỗ trợ của tất cả **quý Thầy Cô tại Khoa Công Nghệ Thông Tin, trường Đại học Nguyễn Tất Thành** và **Thầy TS. Dương Minh Tuấn** đã giúp đỡ và hỗ trợ chúng em trong quá trình thực hiện đồ án này.

# NHẬN XÉT CỦA GIẢNG VIÊN HƯỚNG DẪN

Điểm đồ án:

*TPHCM, Ngày …… tháng …… năm*

**Giáo viên hướng dẫn**

(Ký tên, đóng dấu)

|  |  |
| --- | --- |
| TRƯỜNG ĐẠI HỌC NGUYỄN TẤT THÀNH  **TRUNG TÂM KHẢO THÍ** | **KỲ THI KẾT THÚC HỌC PHẦN**  **HỌC KỲ 2 – NĂM HỌC 2023-2024** |

# PHIẾU CHẤM THI TIỂU LUẬN/ĐỒ ÁN

Môn thi: **Đồ án chuyên ngành khoa học dữ liệu**

Lớp học phần: **21DTH2C**

Nhóm sinh viên thực hiện :

1.Phan Thanh Tâm:2100008769 Tham gia đóng góp: 100%

2.Trần Minh Ngọc:2100010462 Tham gia đóng góp: 100%

Ngày thi:11/09/2024

Phòng thi:L.502

Đề tài tiểu luận/báo cáo của sinh viên : Phân Tích và Xử Lý Dữ Liệu Hiệu Suất Thể Thao: Thông Số Cân Nặng và Kết Quả Tập Luyện

Phần đánh giá của giảng viên (căn cứ trên thang rubrics của môn học):

|  |  |  |  |
| --- | --- | --- | --- |
| **Tiêu chí (theo CĐR HP)** | **Đánh giá của GV** | **Điểm tối đa** | **Điểm đạt được** |
| Cấu trúc của báo cáo |  | 2 |  |
| Nội dung |  |  |  |
| * Các nội dung thành phần |  | 2 |  |
| * Lập luận |  | 4 |  |
| * Kết luận |  | 1 |  |
| Trình bày |  | 1 |  |
| **TỔNG ĐIỂM** |  | 10 |  |

|  |  |
| --- | --- |
|  | **Giảng viên chấm thi**  *(ký, ghi rõ họ tên)* |

Th.S Dương Minh Tuấn

|  |  |
| --- | --- |
| Trường Đại học Nguyễn Tất Thành  **Khoa Công Nghệ Thông Tin**  🙜 🙜 🙝 🙝 | CỘNG HÒA XÃ HỘI CHỦ NGHĨA VIỆT NAM  **Độc lập – Tự do – Hạnh phúc**  🙜 🙜 🙝 🙝 |

# NHIỆM VỤ ĐỒ ÁN CHUYÊN NGÀNH

Họ và tên: **phan thanh tâm** MSSV: **2100008769**

Chuyên ngành: **Khoa học dữ liệu** Lớp:**21DTH2C** Email:[phantam10072003@gmail.com](mailto:phantam10072003@gmail.com)SĐT: 0353892615

Tên đề tài: **Phân Tích và Xử Lý Dữ Liệu Hiệu Suất Thể Thao: Thông Số Cân Nặng và Kết Quả Tập Luyện**

Giáo viên hướng dẫn: **Dương Minh Tuấn**

Thời gian thực hiện: **24/06/2024** **đến 11/09/2024**

Nhiệm vụ/nội dung (mô tả chi tiết nội dung, yêu cầu, phương pháp… ) :

* Tìm hiểu định nghĩa và ý nghĩa làm sạch dữ liệu (Data Cleaning) và phân tích dữ liệu (Data Analysis).
* Trình bày nội dung và các bước thực hiện làm sạch dữ liệu và phân tích dữ liệu.
* Làm sạch dữ liệu (Data Cleaning):
* Thu thập và khám phá dữ liệu: Xác định và thu thập dữ liệu từ nhiều nguồn khác nhau và khám phá sơ bộ để đánh giá cấu trúc và độ tin cậy.
* Xử lý dữ liệu: Bao gồm xử lý dữ liệu thiếu, loại bỏ nhiễu và chuẩn hóa để chuẩn bị cho phân tích.
* Phân tích dữ liệu (Data Analysis):
* Mô tả và phân tích: Sử dụng thống kê mô tả và biểu đồ để mô tả dữ liệu và phân tích mối quan hệ giữa các biến.
* Áp dụng phương pháp phân tích: Sử dụng các kỹ thuật và mô hình để tìm ra các insights từ dữ liệu đã được làm sạch.
* Thực nghiệm và triển khai các công cụ và phương pháp sử dụng.
* Ngôn ngữ lập trình:
* Sử dụng Python làm ngôn ngữ chính để thực hiện các phần của nghiên cứu.
* Phương pháp mô hình hóa dữ liệu:
* Phân tích nhân tố (Factor Analysis)
* Mô hình hóa hỗn hợp (Mixture Modeling)

**Nội dung và yêu cầu đã được thông qua Bộ môn.**

*TP.HCM, ngày 30 tháng 06 năm 2022*

|  |  |
| --- | --- |
| **TRƯỞNG BỘ MÔN**  *(Ký và ghi rõ họ tên)* | **GIÁO VIÊN HƯỚNG DẪN**  *(Ký và ghi rõ họ tên)* |

MỤC LỤC

[- LỜI CẢM ƠN 3](#_Toc176638392)

[- LỜI MỞ ĐẦU 4](#_Toc176638393)

[- NHẬN XÉT CỦA GIẢNG VIÊN HƯỚNG DẪN 5](#_Toc176638394)

[- PHIẾU CHẤM THI TIỂU LUẬN/ĐỒ ÁN 6](#_Toc176638395)

[- NHIỆM VỤ ĐỒ ÁN CHUYÊN NGÀNH 7](#_Toc176638396)

[MỤC LỤC 9](#_Toc176638397)

[DANH MỤC HÌNH 10](#_Toc176638398)

[DANH MỤC CHỮ VIẾT TẮT 11](#_Toc176638399)

[CHƯƠNG 1. GIỚI THIỆU 12](#_Toc176638400)

[1.1. Mô Tả Tóm Tắt Về Đề Tài 12](#_Toc176638401)

[1.2. Mục Tiêu Và Nhiệm Vụ Nghiên Cứu 12](#_Toc176638402)

[1.2.1. Đối tượng và phạm vi nghiên cứu 13](#_Toc176638403)

[1.2.2. Phương pháp nghiên cứu 13](#_Toc176638404)

[1.3. Ý nghĩa khoa học và thực tiễn 14](#_Toc176638405)

[CHƯƠNG 2. CƠ SỞ LÝ THUYẾT 14](#_Toc176638406)

[2.1. Hiểu về hiệu suất thể thao 14](#_Toc176638407)

[2.2. Làm Sạch Dữ Liệu (Data Cleaning) 15](#_Toc176638408)

[2.3. Phân Tích Dữ Liệu (Data Analysis) 17](#_Toc176638409)

[2.4. Mô Hình Hóa Hỗn Hợp (Mixture Modeling) 18](#_Toc176638410)

[2.5. Phân Tích Nhân Tố (Factor Analysis) 19](#_Toc176638411)

[2.6. Chuẩn Hóa Dữ Liệu (Data Normalization) 19](#_Toc176638412)

[CHƯƠNG 3. THỰC NGHIỆM 20](#_Toc176638413)

[3.1. Phân tích dữ liệu ban đầu 20](#_Toc176638414)

[3.2. Kiểm tra và tiền sử lý dữ liệu 26](#_Toc176638415)

[3.3. Thống kê mô tả dữ liệu. 29](#_Toc176638416)

[CHƯƠNG 4.KẾT LUẬN 31](#_Toc176638417)

[Tài Liệu Tham Khảo 1](#_Toc176638418)

# DANH MỤC HÌNH

[Hình 1: Data Cleaning mang đến vô số lợi ích cho doanh nghiệp. 15](#_Toc176638504)

[Hình 2: Các phương pháp phân tích dữ liệu. 17](#_Toc176638505)

[Hình 3: Tổng quan về phân tích nhân tố khám phá EFA. 19](#_Toc176638506)

[Hình 4: Đọc dữ liệu từ tệp CSV và in 10 dòng đầu tiên của dataframe. 20](#_Toc176638507)

[Hình 5: Kiểm Tra Giá Trị Thiếu. 22](#_Toc176638508)

[Hình 6: Xử Lý Dữ Liệu Nhiễu và Bất Thường. 23](#_Toc176638509)

[Hình 7: Chuẩn Hóa Dữ Liệu. 24](#_Toc176638510)

[Hình 8: Loại Bỏ Các Cột Không Cần Thiết. 25](#_Toc176638511)

[Hình 9: Kiểm Tra Mối Quan Hệ Giữa Các Đặc Trưng. 25](#_Toc176638512)

[Hình 10: Chia Dữ Liệu Thành Tập Huấn Luyện và Tập Kiểm Tra. 26](#_Toc176638513)

[Hình 11: Phân Tích Nhân Tố (Factor Analysis). 27](#_Toc176638514)

[Hình 12: Mô Hình Hóa Hỗn Hợp (Gaussian Mixture Modeling). 28](#_Toc176638515)

[Hình 13: In thống kê mô tả cho các cột. 29](#_Toc176638516)

# DANH MỤC CHỮ VIẾT TẮT

(EFA) Exploratory Factor Analysis.

(AVS) Advanced Video Coding.

(PII) Personally Identifiable Information.

(ETL) Extract, Transform, Load.

(GMM) Gaussian Mixture Model.

# CHƯƠNG 1. GIỚI THIỆU

## Mô Tả Tóm Tắt Về Đề Tài

Việc chọn đề tài “Phân Tích và Xử Lý Dữ Liệu Hiệu Suất Thể Thao: Thông Số Cân Nặng và Kết Quả Tập Luyện” có nhiều khía cạnh quan trọng và hữu ích. Đề tài này không chỉ giúp phân tích các yếu tố ảnh hưởng đến hiệu suất thể thao mà còn mở rộng cái nhìn về cách các thông số cơ thể, cụ thể là cân nặng, có thể tác động đến kết quả của quá trình tập luyện.

Cụ thể, nghiên cứu này sẽ tập trung vào việc xác định mối quan hệ giữa cân nặng của vận động viên và các chỉ số hiệu suất thể thao của họ. Ví dụ, chúng ta có thể khám phá xem cân nặng có ảnh hưởng đến sức bền, tốc độ, khả năng phục hồi hay các yếu tố khác liên quan đến hiệu suất thể thao như thế nào. Bằng cách phân tích dữ liệu từ các bài tập luyện và thi đấu, chúng ta có thể rút ra những kết luận về cách tối ưu hóa các chương trình luyện tập, giúp vận động viên cải thiện hiệu suất của mình một cách hiệu quả hơn.

Ngoài việc phân tích mối quan hệ giữa cân nặng và kết quả tập luyện, đề tài này cũng chú trọng đến việc làm sạch dữ liệu. Làm sạch dữ liệu là một bước quan trọng để đảm bảo rằng các phân tích và kết luận được đưa ra dựa trên dữ liệu chính xác và đáng tin cậy. Điều này bao gồm việc loại bỏ các dữ liệu không chính xác, khắc phục các lỗi trong dữ liệu và xử lý các vấn đề liên quan đến thiếu sót dữ liệu. Khi dữ liệu được xử lý một cách chính xác, chúng ta có thể đưa ra những khuyến nghị có cơ sở vững chắc về cách tối ưu hóa chế độ tập luyện, chế độ ăn uống và các yếu tố khác để cải thiện hiệu suất thể thao.

Từ những phân tích và kết luận rút ra từ nghiên cứu này, các huấn luyện viên và vận động viên có thể áp dụng các phương pháp luyện tập phù hợp hơn với từng cá nhân, từ đó đạt được kết quả tốt hơn trong các cuộc thi đấu và cải thiện sức khỏe tổng thể. Hơn nữa, nghiên cứu này cũng có thể đóng góp vào việc phát triển các phương pháp và công cụ phân tích hiệu suất thể thao mới, giúp nâng cao chất lượng đào tạo và thi đấu trong tương lai.

## 1.2. Mục Tiêu Và Nhiệm Vụ Nghiên Cứu

* Mục tiêu:

Mục tiêu nghiên cứu của đề tài “Phân Tích và Xử Lý Dữ Liệu Hiệu Suất Thể Thao: Thông Số Cân Nặng và Kết Quả Tập Luyện” là đa dạng và sâu rộng, nhằm đáp ứng các nhu cầu và thách thức trong việc tối ưu hóa hiệu suất thể thao thông qua việc sử dụng các kỹ thuật phân tích dữ liệu tiên tiến. Như là:

Làm sạch dữ liệu hiệu suất thể thao: Một trong những mục tiêu quan trọng nhất của nghiên cứu là thực hiện quy trình làm sạch dữ liệu nhằm loại bỏ những dữ liệu không chính xác, thiếu sót hoặc bị lỗi. Điều này bao gồm việc kiểm tra và sửa chữa các giá trị bị sai, điều chỉnh dữ liệu không đầy đủ, và xử lý các điểm dữ liệu ngoại lệ có thể ảnh hưởng đến tính chính xác của các phân tích. Làm sạch dữ liệu là bước tiền đề cần thiết để đảm bảo rằng mọi phân tích và kết luận đưa ra dựa trên dữ liệu có độ tin cậy cao.

Phân tích các yếu tố ảnh hưởng đến hiệu suất thể thao: Nghiên cứu sẽ tiến hành phân tích sâu rộng các yếu tố có thể ảnh hưởng đến hiệu suất thể thao của vận động viên, đặc biệt là mối quan hệ giữa cân nặng cơ thể và kết quả tập luyện. Điều này bao gồm việc xem xét cách cân nặng có thể ảnh hưởng đến các chỉ số như sức bền, tốc độ, sự phục hồi, và các yếu tố khác liên quan đến hiệu suất. Bằng cách phân tích dữ liệu này, nghiên cứu sẽ xác định những yếu tố quan trọng nhất và mức độ ảnh hưởng của chúng đối với hiệu suất thể thao.

Áp dụng công cụ và phương pháp hiện đại: Nghiên cứu sẽ sử dụng các công cụ và phương pháp phân tích dữ liệu hiện đại, bao gồm các kỹ thuật thống kê tiên tiến, học máy (machine learning), và các mô hình phân tích dữ liệu. Việc áp dụng các công cụ này sẽ giúp khai thác và hiểu rõ hơn các mối quan hệ phức tạp trong dữ liệu, đồng thời cung cấp những phân tích chi tiết và chính xác hơn. Công nghệ hiện đại cũng sẽ giúp tạo ra các mô hình dự đoán hiệu suất và các mô hình mô phỏng kịch bản luyện tập khác nhau.

* Nhiệm vụ:

Nghiên cứu đề tài.

Thu Thập và Khám Phá Dữ Liệu.

Tiền xử lý dữ liệu.

Mô Tả và Phân Tích.

Áp Dụng Phương Pháp Phân Tích.

Thực Nghiệm và Triển Khai.

Sử dụng Python làm ngôn ngữ chính để thực hiện các tác vụ làm sạch và phân tích dữ liệu.

Phân Tích Nhân Tố (Factor Analysis).

Mô Hình Hóa Hỗn Hợp (Mixture Modeling).

### 1.2.1. Đối tượng và phạm vi nghiên cứu

Đối tượng nghiên cứu:

Dữ Liệu Hiệu Suất Thể Thao: Dữ liệu liên quan đến hiệu suất thể thao của vận động viên, bao gồm thông tin về cân nặng cơ thể, kết quả các bài tập (squat, bench press, deadlift), và các chỉ số liên quan như điểm số Wilks.

Các Vận Động Viên: Các vận động viên tham gia thi đấu trong các giải thể thao, đặc biệt trong môn cử tạ hoặc các môn thể thao tương tự, với các thông tin về tuổi, giới tính, loại thiết bị sử dụng, và hạng cân.

Phạm vi nghiên cứu:

Các Vận Động Viên: Các vận động viên tham gia thi đấu trong các giải thể thao, đặc biệt trong môn cử tạ hoặc các môn thể thao tương tự, với các thông tin về tuổi, giới tính, loại thiết bị sử dụng, và hạng cân..

### 1.2.2. Phương pháp nghiên cứu

Tìm kiếm tài liệu trên internet có liên quan đến nội dung nghiên cứu

Thu thập tài liệu liên quan đến nghiệp vụ quản lý kiểm tra

Làm sạch và chuẩn bị dữ liệu, phân tích dữ liệu mô tả và quan hệ, cùng với việc sử dụng các phương pháp mô hình hóa để hiểu rõ hơn về hiệu suất thể thao và các yếu tố ảnh hưởng đến nó.

## 1.3. Ý nghĩa khoa học và thực tiễn

**Ý nghĩa khoa học:**

Nâng Cao Hiểu Biết Về Hiệu Suất Thể Thao: Nghiên cứu cung cấp cái nhìn sâu sắc về các yếu tố chính ảnh hưởng đến hiệu suất thể thao, như cân nặng cơ thể và hạng cân. Điều này góp phần làm rõ hơn các mối quan hệ và cơ chế tiềm ẩn trong hiệu suất thể thao.

Cải Thiện Phương Pháp Phân Tích Dữ Liệu: Việc áp dụng các phương pháp phân tích nâng cao như phân tích nhân tố và mô hình hóa hỗn hợp giúp mở rộng khả năng phân tích dữ liệu thể thao, cung cấp các công cụ phân tích và mô hình hóa mới cho các nghiên cứu tiếp theo.

Phát Triển Kỹ Năng Phân Tích: Nghiên cứu giúp phát triển các phương pháp làm sạch dữ liệu và phân tích, từ đó nâng cao chất lượng các nghiên cứu khoa học trong lĩnh vực thể thao và các lĩnh vực khác.

**Ý nghĩa thực tiễn:**

Cải Thiện Hiệu Suất Thể Thao: Các kết quả phân tích có thể giúp các huấn luyện viên và vận động viên điều chỉnh các chương trình tập luyện và chiến lược thi đấu, nhằm tối ưu hóa hiệu suất và đạt được kết quả tốt hơn.

Hỗ Trợ Quyết Định Trong Quản Lý Thể Thao: Cung cấp thông tin quý giá cho các nhà quản lý thể thao và tổ chức thể thao trong việc ra quyết định về lựa chọn thiết bị, chương trình đào tạo và các yếu tố khác liên quan đến hiệu suất thi đấu.

Ứng Dụng Công Nghệ Hiện Đại: Việc sử dụng Python và các công cụ phân tích hiện đại không chỉ giúp xử lý và phân tích dữ liệu một cách hiệu quả mà còn cung cấp kỹ năng thực tiễn quan trọng cho các chuyên gia dữ liệu và nhà phân tích.

Đóng Góp cho Cộng Đồng Thể Thao: Kết quả nghiên cứu có thể được áp dụng để cải thiện các thực hành thể thao hiện tại, từ đó đóng góp vào sự phát triển chung của cộng đồng thể thao và sức khỏe.

# CHƯƠNG 2. CƠ SỞ LÝ THUYẾT

## 2.1. Hiểu về hiệu suất thể thao

Hiệu suất thể thao là một lĩnh vực phức tạp và có thể bị ảnh hưởng bởi nhiều yếu tố khác nhau. Hai yếu tố quan trọng được thu thập và nghiên cứu, là thông số cân nặng và kết quả tập luyện, có vai trò thiết yếu trong việc cải thiện và tối ưu hóa hiệu suất thể thao.

Thông Số Cân Nặng:

Tác động của cân nặng đối với hiệu suất: Cân nặng cần phải được điều chỉnh sao cho tối ưu cho từng môn thể thao cụ thể.

Tác động đến sức bền và hiệu suất: Cân nặng quá mức có thể làm tăng áp lực lên cơ thể, dẫn đến việc tiêu hao năng lượng nhanh hơn và làm giảm sức bền. Ngược lại, cân nặng quá thấp có thể dẫn đến thiếu hụt năng lượng và sức mạnh cơ bắp, ảnh hưởng đến hiệu suất tổng thể. Cân nặng không chỉ là vấn đề về trọng lượng tổng thể mà còn liên quan đến tỷ lệ cơ và mỡ trong cơ thể. Cơ bắp nặng hơn mỡ, vì vậy, việc tăng tỷ lệ cơ bắp so với mỡ có thể cải thiện hiệu suất thể thao mà không cần phải giảm cân tổng thể.

2. Kết Quả Tập Luyện

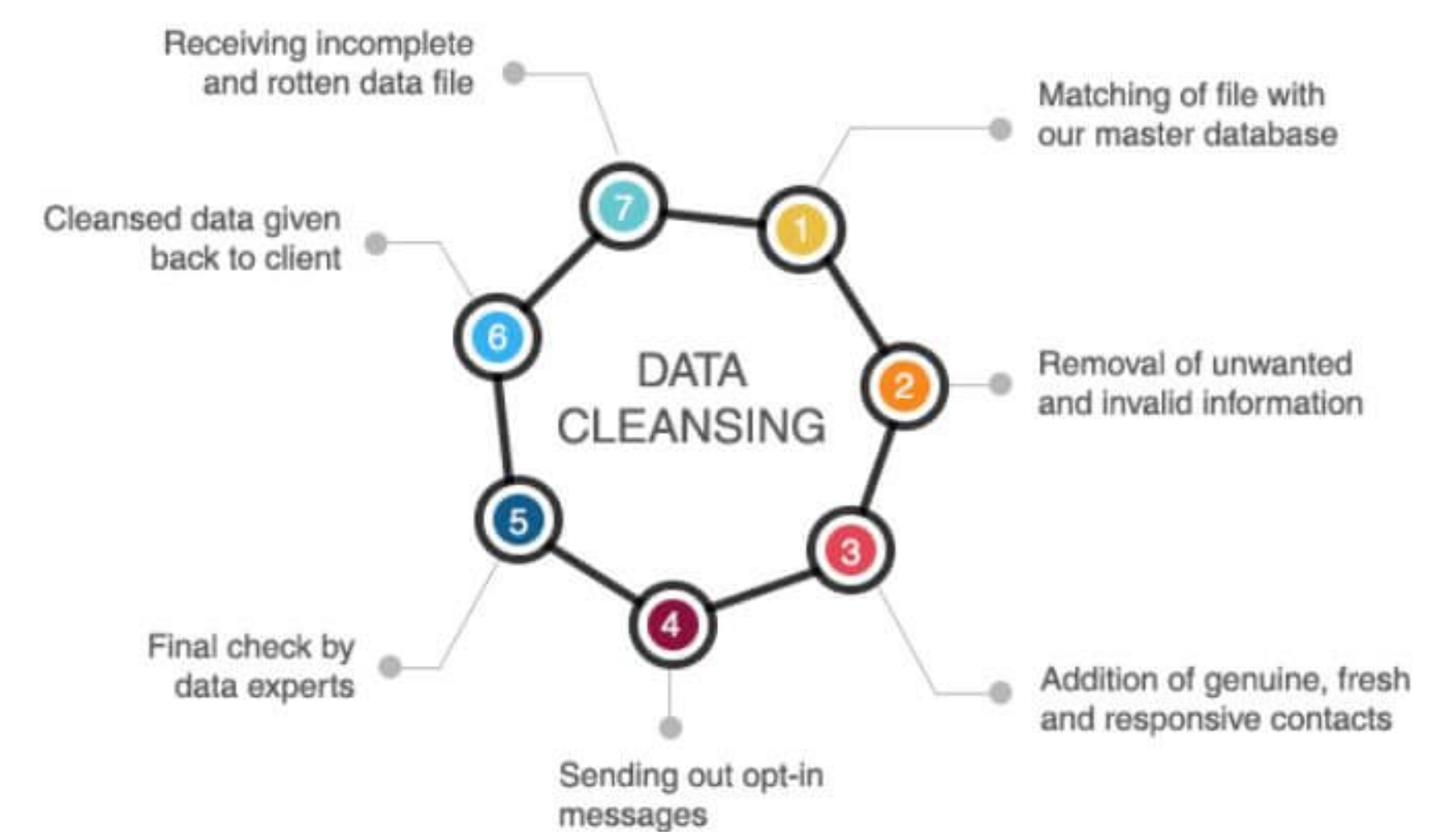
Chất lượng và kiểu tập luyện: Một chương trình tập luyện hiệu quả phải được thiết kế dựa trên mục tiêu thể thao cụ thể và cần phải bao gồm các bài tập phù hợp với nhu cầu của từng môn thể thao. Điều này bao gồm cả tập luyện sức mạnh, sức bền, và kỹ thuật. Việc kết hợp nhiều loại hình tập luyện, từ các bài tập sức mạnh đến bài tập cardio, có thể giúp cải thiện toàn diện hiệu suất thể thao. Sự đa dạng này cũng giúp ngăn ngừa chấn thương và giữ cho việc tập luyện thú vị.

Khả năng phục hồi và nghỉ ngơi: Cơ thể cần thời gian để phục hồi sau các buổi tập luyện nặng. Nếu không có đủ thời gian phục hồi, hiệu suất có thể giảm sút và nguy cơ chấn thương có thể gia tăng.Dinh dưỡng và giấc ngủ đóng vai trò quan trọng trong việc phục hồi và phát triển cơ bắp. Một chế độ ăn uống hợp lý và giấc ngủ đầy đủ có thể cải thiện đáng kể kết quả tập luyện

## 2.2. Làm Sạch Dữ Liệu (Data Cleaning)

Làm sạch dữ liệu hay Data Cleaning là quá trình lấy dữ liệu như bạn hiện có và dọn dẹp nó bằng cách sửa lỗi, xác định các điểm dữ liệu không chính xác, các mục nhập trùng lặp,...

Data Cleaning được xem là một công việc quan trọng không thể thiếu. Quá trình này giúp dữ liệu được chuẩn hóa, tăng tính chính xác, mang đến những câu trả lời đáng tin cậy và giúp các công cụ phân tích dữ liệu truy cập và làm việc thuận lợi.



Hình 1: Data Cleaning mang đến vô số lợi ích cho doanh nghiệp.

**Sau khi được thu thập, dữ liệu được kiểm tra về chất lượng và “tính toàn vẹn” ở một số khía cạnh như:**

Độ chính xác: có nghĩa là dữ liệu có “đúng” trong một ngữ cảnh nhất định hay không. Một trong những hệ thống chính xác được sử dụng rộng rãi nhất là AVS, kiểm tra thông tin thẻ tín dụng đối với địa chỉ thanh toán trong quá trình mua hàng trực tuyến.

Tính đầy đủ: cho dù mọi đầu vào được yêu cầu đều có giá trị hay không. Trong trường hợp khách hàng tiềm năng, sẽ không hữu ích lắm nếu mục nhập có tên nhưng không có họ, điều này có thể khiến mục nhập gần như vô dụng.

Tính nhất quán: một số mục có thể được tìm thấy ở nhiều nơi (ví dụ: thông tin liên hệ được lấy từ nhiều nguồn) và xung đột ở đây có thể gây ra sự cố.

Hiệu lực: dữ liệu thường phải đáp ứng một số ràng buộc nhất định để nó có ý nghĩa khi so sánh với các giá trị khác. Hầu hết các vấn đề về tính hợp lệ đều được tìm thấy trong dữ liệu do hệ thống cũ thu thập.

Tính đồng nhất: máy tính thường không thực hiện tốt công việc so sánh các số và giá trị với nhau trừ khi tất cả chúng được định dạng theo cùng một cách.

**Các vấn đề chất lượng khác nhau yêu cầu các phương pháp hiệu chỉnh khác nhau, chẳng hạn như:**

Phân tích cú pháp: một số giá trị dễ xử lý hơn các giá trị khác. Ví dụ, các giá trị số dễ hiểu hơn, kiểm tra chất lượng và thường dễ sửa hơn. Một số yêu cầu phân tích chỉ để hiểu đầu vào giống như với bất kỳ đầu vào dữ liệu “phi cấu trúc” nào.

Loại bỏ trùng lặp: nhiều mục nhập có thể tạo ra vấn đề, vì vậy cần phải nỗ lực rất nhiều để điều hòa sự không nhất quán giữa chúng. Lý tưởng nhất là thông tin chính xác được xác định và các bản sao bị lỗi sẽ bị loại khỏi cơ sở dữ liệu.

Phân tích thống kê: ngay cả trong bối cảnh kinh doanh, việc kiểm soát các điểm bất thường và ngoại lệ cực đoan là một thông lệ quan trọng. Mặc dù một số trường hợp sử dụng cần thiết hơn những trường hợp khác nhưng việc áp dụng các mô hình thống kê cho dữ liệu có thể giúp ngăn một số mục nhập làm sai lệch các con số. Và mặc dù các ngoại lệ không phải lúc nào cũng bị xóa nhưng việc xác định chính xác chúng sẽ giúp kiểm soát và cho phép các nhóm giải quyết chúng một cách riêng biệt.

Rà soát dữ liệu nhạy cảm: thông tin nhận dạng cá nhân (PII) có thể là một vấn đề đối với một số ngành và trường hợp sử dụng và có thể cần phải xử lý đúng cách thông tin đó để tuân thủ pháp luật. Trong một số trường hợp, dữ liệu có thể bị xóa hoặc ẩn hoàn toàn. Ở những nơi khác, nó phải hiển thị với một số người, nhưng các biện pháp kiểm soát quyền phải được thực hiện để bảo vệ quyền riêng tư.

Chuyển đổi dữ liệu: ngay cả khi dữ liệu khá sạch để bắt đầu, thường cần phải định dạng lại một số lượng nhất định để cho phép các công cụ phân tích sử dụng dữ liệu đó một cách hiệu quả. Đó là lúc chuyển đổi dữ liệu xuất hiện. Thường là một phần của chức năng ETL, đây là nơi dữ liệu được chuẩn hóa và căn chỉnh với định dạng đích.

**Các thành phần chính trong đề tài:**

Xử Lý Dữ Liệu Thiếu: Các phương pháp để xử lý giá trị thiếu bao gồm thay thế (imputation) và loại bỏ (drop). Phương pháp thay thế có thể là sử dụng giá trị trung bình, trung vị, hoặc các phương pháp phức tạp hơn như hồi quy.

Loại Bỏ Nhiễu: Nhận diện và loại bỏ các giá trị bất thường (outliers) hoặc nhiễu trong dữ liệu có thể ảnh hưởng đến chất lượng phân tích.

Chuẩn Hóa Dữ Liệu: Điều chỉnh dữ liệu để đảm bảo tính đồng nhất về định dạng và đơn vị đo lường, như chuẩn hóa hoặc chuẩn hóa dữ liệu.

## 2.3. Phân Tích Dữ Liệu (Data Analysis)

Phân tích dữ liệu là quá trình khám phá và khai thác thông tin từ dữ liệu để rút ra các kết luận và quyết định. Phân tích dữ liệu bao gồm mô tả dữ liệu, khám phá mối quan hệ giữa các biến và áp dụng các mô hình để tìm hiểu sâu hơn về dữ liệu.



Hình 2: Các phương pháp phân tích dữ liệu.

**Một số phương pháp phân tích dữ liệu:**

Phân tích cụm (Cluster analysis) được thực hiện dựa trên việc nhóm các phần dữ liệu có đặc điểm chung với nhau.

Phân tích theo nhóm (Cohort analysis) sử dụng dữ liệu lịch sử để kiểm tra và đối chiếu một phân khúc xác định về hành vi của người dùng, sau đó nhóm chúng lại với những phân khúc khác có đặc điểm tương tự.

Phân tích hồi quy (Regression analysis) sử dụng dữ liệu lịch sử để hiểu tác động đến giá trị của biến phụ thuộc khi một (hồi quy tuyến tính) hoặc nhiều biến độc lập (hồi quy bội) thay đổi hoặc giữ nguyên.

Mạng nơ-ron là nền tảng cho các thuật toán thông minh của học máy. Nó là một dạng phân tích có sự can thiệp tối thiểu, để hiểu cách bộ não con người tạo ra những hiểu biết sâu sắc và dự đoán các giá trị. Mạng nơ-ron học hỏi từ mọi dữ liệu, nghĩa là chúng phát triển và tiến bộ theo thời gian.

Phân tích nhân tố còn được gọi là “giảm chiều dữ liệu” (dimension reduction) mô tả sự biến thiên của những biến có tương quan được quan sát bằng một số nhỏ hơn các biến không quan sát được gọi là nhân tố. Mục đích ở đây là phát hiện ra các biến tiềm ẩn độc lập.

Khai thác dữ liệu (Data mining) là phương pháp phân tích dữ liệu giúp trích xuất thông tin từ một tập dữ liệu nhằm xác định xu hướng, mẫu và dữ liệu hữu ích.

Phân tích văn bản (Text analysis) hoạt động bằng cách lấy các bộ dữ liệu văn bản lớn và sắp xếp chúng nhằm dễ quản lý hơn.

Phân tích chuỗi thời gian (Time series analysis) được sử dụng để phân tích một tập hợp dữ liệu thu thập trong một khoảng thời gian xác định.

Cây quyết định (Decision Trees) Phân tích dựa trên cây quyết định hoạt động như một công cụ hỗ trợ để đưa ra các quyết định chiến lược và thông minh.

Phân tích thuộc tính (Conjoint analysis) được sử dụng trong các cuộc khảo sát để hiểu cách người dùng đánh giá các thuộc tính khác nhau của một sản phẩm hoặc dịch vụ.

**Các thành phần chính trong đề tài:**

Thống Kê Mô Tả (Descriptive Statistics): Sử dụng các chỉ số như trung bình, độ lệch chuẩn, và các biểu đồ để mô tả và tổng hợp dữ liệu. Các biểu đồ phổ biến bao gồm histogram, box plot và scatter plot.

Phân Tích Mối Quan Hệ (Exploratory Data Analysis): Khám phá mối quan hệ giữa các biến bằng các kỹ thuật như phân tích tương quan và hồi quy.

Mô Hình Hóa Dữ Liệu (Data Modeling): Áp dụng các mô hình thống kê và học máy để khám phá cấu trúc dữ liệu, như phân tích nhân tố (Factor Analysis) và mô hình hóa hỗn hợp (Mixture Modeling).

## 2.4. Mô Hình Hóa Hỗn Hợp (Mixture Modeling)

Mô hình hóa hỗn hợp là kỹ thuật phân tích xác suất để phân nhóm dữ liệu dựa trên giả định rằng dữ liệu đến từ một tổ hợp các phân phối xác suất khác nhau. Các mô hình hỗn hợp phổ biến bao gồm Mô hình Gaussian Mixture (GMM).

Một mô hình hỗn hợp hữu hạn chiều điển hình là một mô hình phân cấp bao gồm các thành phần sau:

N biến ngẫu nhiên được quan sát, mỗi biến phân phối theo hỗn hợp K thành phần, với các thành phần thuộc cùng một họ phân phối tham số (ví dụ: tất cả đều bình thường, tất cả đều Zipfian , v.v.) nhưng có các tham số khác nhau

N biến tiềm ẩn ngẫu nhiên chỉ định danh tính của thành phần hỗn hợp của mỗi quan sát, mỗi biến được phân phối theo phân phối danh mục K chiều

Một tập hợp K trọng số hỗn hợp, là các xác suất có tổng bằng 1.

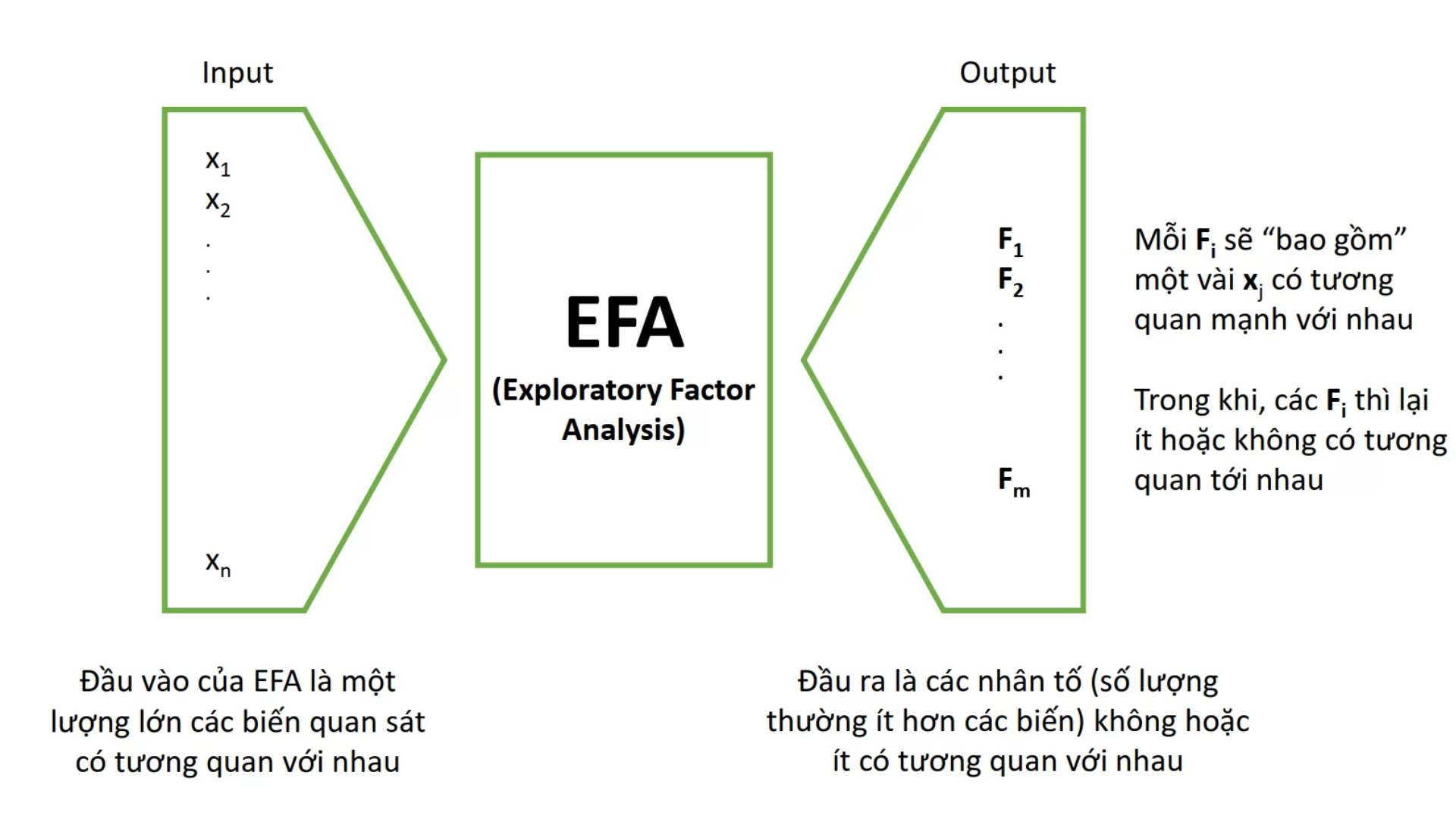
Một tập hợp K tham số, mỗi tham số chỉ định tham số của thành phần hỗn hợp tương ứng. Trong nhiều trường hợp, mỗi "tham số" thực sự là một tập hợp các tham số. Ví dụ, nếu các thành phần hỗn hợp là phân phối chuẩn Gauss, sẽ có giá trị trung bình và phương sai cho mỗi thành phần. Nếu các thành phần hỗn hợp là phân phối chuẩn phân loại (ví dụ, khi mỗi quan sát là một mã thông báo từ một bảng chữ cái hữu hạn có kích thước V ), sẽ có một vectơ V xác suất tổng bằng 1.

**Phương pháp mô hình hóa hỗn hợp được sử dụng:**

Gaussian Mixture Model (GMM): Mô hình này giả định rằng dữ liệu được tạo ra từ nhiều phân phối Gaussian với các tham số khác nhau. GMM giúp phân nhóm dữ liệu dựa trên sự tương đồng.

## 2.5. Phân Tích Nhân Tố (Factor Analysis)

Phân Tích Nhân Tố (Factor Analysis) dùng để rút gọn một tập hợp k biến quan sát thành một tập F (với F < k) các nhân tố có ý nghĩa hơn. Trong nghiên cứu, chúng ta thường thu thập được một số lượng biến khá lớn và rất nhiều các biến quan sát trong đó có liên hệ tương quan với nhau. Thay vì đi nghiên cứu 20 đặc điểm nhỏ của một đối tượng, chúng ta có thể chỉ nghiên cứu 4 đặc điểm lớn, trong mỗi đặc điểm lớn này gồm 5 đặc điểm nhỏ có sự tương quan với nhau. Điều này giúp tiết kiệm thời gian và kinh phí nhiều hơn cho người nghiên cứu.



Hình 3: Tổng quan về phân tích nhân tố khám phá EFA.

Phân tích nhân tố nhằm giảm số lượng biến và phát hiện các yếu tố tiềm ẩn ảnh hưởng đến tập hợp các biến quan sát được. Mục tiêu là xác định các yếu tố chính giải thích sự biến thiên trong dữ liệu.

**Phương pháp sử dụng trong đồ án:**

Kỹ Thuật Phân Tích Nhân Tố: Sử dụng các phương pháp như phân tích nhân tố chính (Principal Factor Analysis) và phân tích nhân tố khẳng định (Confirmatory Factor Analysis) để xác định số lượng và bản chất của các yếu tố.

## 2.6. Chuẩn Hóa Dữ Liệu (Data Normalization)

Chuẩn hóa dữ liệu (Data Normalization) là quá trình tổ chức và cấu trúc dữ liệu theo cách loại bỏ sự trùng lặp, đảm bảo tính nhất quán và giảm thiểu các biến đổi dữ liệu. Nó liên quan đến việc chia dữ liệu thành các bảng nhỏ hơn, có quan hệ logic với nhau, giảm thiểu sự lặp lại dữ liệu và thiết lập mối quan hệ giữa các bảng này thông qua các khóa. Chuẩn hóa dữ liệu là quá trình điều chỉnh dữ liệu để đưa về một phạm vi hoặc chuẩn mực chung, giúp các biến có cùng đơn vị đo lường và tầm quan trọng trong phân tích

**Phương Pháp được sử dụng là:**

Z-Score Normalization: Điều chỉnh dữ liệu sao cho có trung bình bằng 0 và độ lệch chuẩn bằng 1.

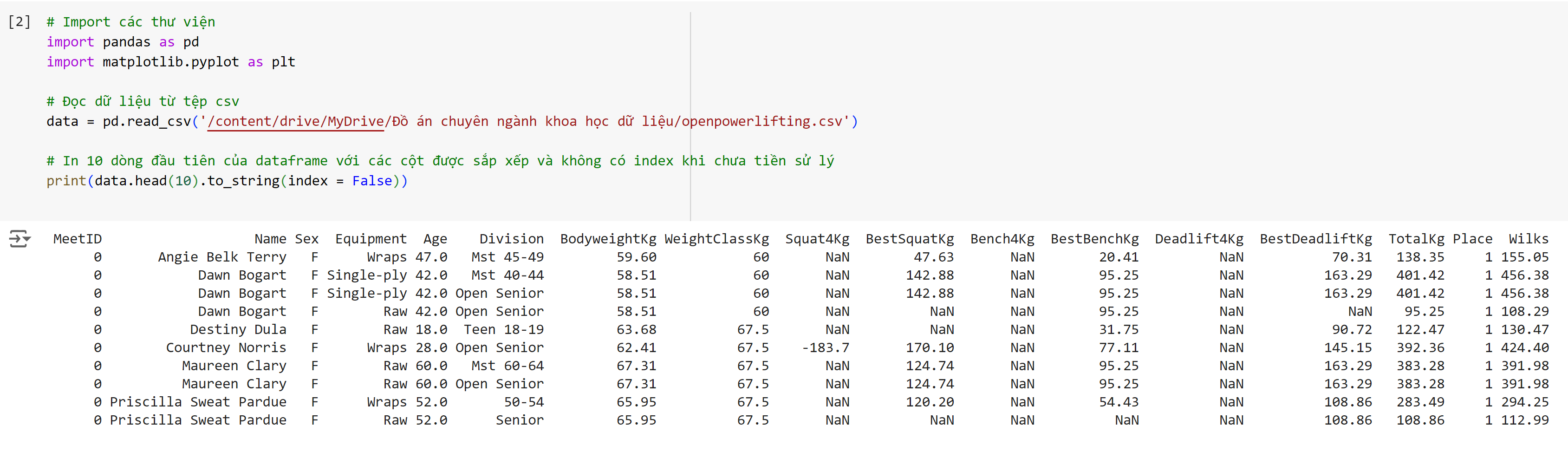
Min-Max Normalization: Chuyển đổi dữ liệu về phạm vi từ 0 đến 1 hoặc từ -1 đến 1.

# CHƯƠNG 3. THỰC NGHIỆM

## 3.1. Phân tích dữ liệu ban đầu

**Nhập và Hiển Thị Dữ Liệu:**

Đọc dữ liệu từ tệp CSV và in 10 dòng đầu tiên của dataframe.



Hình 4: Đọc dữ liệu từ tệp CSV và in 10 dòng đầu tiên của dataframe.

Các cột dữ liên quan đến kết quả thi đấu của các vận động viên cử tạ.

**MeetID:** Mã số của cuộc thi. Tất cả các dòng có giá trị giống nhau, cho thấy tất cả dữ liệu đều thuộc về cùng một cuộc thi.

**Name:** Tên của vận động viên.

**Sex:** Giới tính của vận động viên (nam hoặc nữ).

**Equipment:** Loại thiết bị mà vận động viên sử dụng, chẳng hạn như "Wraps" (băng quấn) hoặc "Single-ply" (loại vải một lớp). "Raw" có nghĩa là không sử dụng thiết bị hỗ trợ.

**Age:** Tuổi của vận động viên.

**Division:** Phân loại của vận động viên dựa trên độ tuổi và loại thi đấu, chẳng hạn như "Mst 45-49" (người lớn tuổi từ 45 đến 49) hoặc "Open Senior" (người lớn tuổi mở rộng).

**BodyweightKg:** Cân nặng của vận động viên tính bằng kilôgam.

**WeightClassKg:** Cân nặng theo phân hạng của vận động viên tính bằng kilôgam.

**Squat4Kg:** Đối với cử tạ squat, khối lượng tạ mà vận động viên đã thử với trọng lượng 4kg. "NaN" có nghĩa là không có dữ liệu hoặc không thực hiện bài tập.

**BestSquatKg:** Khối lượng tạ tốt nhất mà vận động viên thực hiện được trong bài squat tính bằng kilôgam.

**Bench4Kg:** Đối với cử tạ bench, khối lượng tạ mà vận động viên đã thử với trọng lượng 4kg. "NaN" có nghĩa là không có dữ liệu hoặc không thực hiện bài tập.

**BestBenchKg:** Khối lượng tạ tốt nhất mà vận động viên thực hiện được trong bài bench press tính bằng kilôgam.

**Deadlift4Kg:** Đối với cử tạ deadlift, khối lượng tạ mà vận động viên đã thử với trọng lượng 4kg. "NaN" có nghĩa là không có dữ liệu hoặc không thực hiện bài tập.

**BestDeadliftKg:** Khối lượng tạ tốt nhất mà vận động viên thực hiện được trong bài deadlift tính bằng kilôgam.

**TotalKg:** Tổng khối lượng tạ mà vận động viên thực hiện được (tổng của squat, bench press và deadlift).

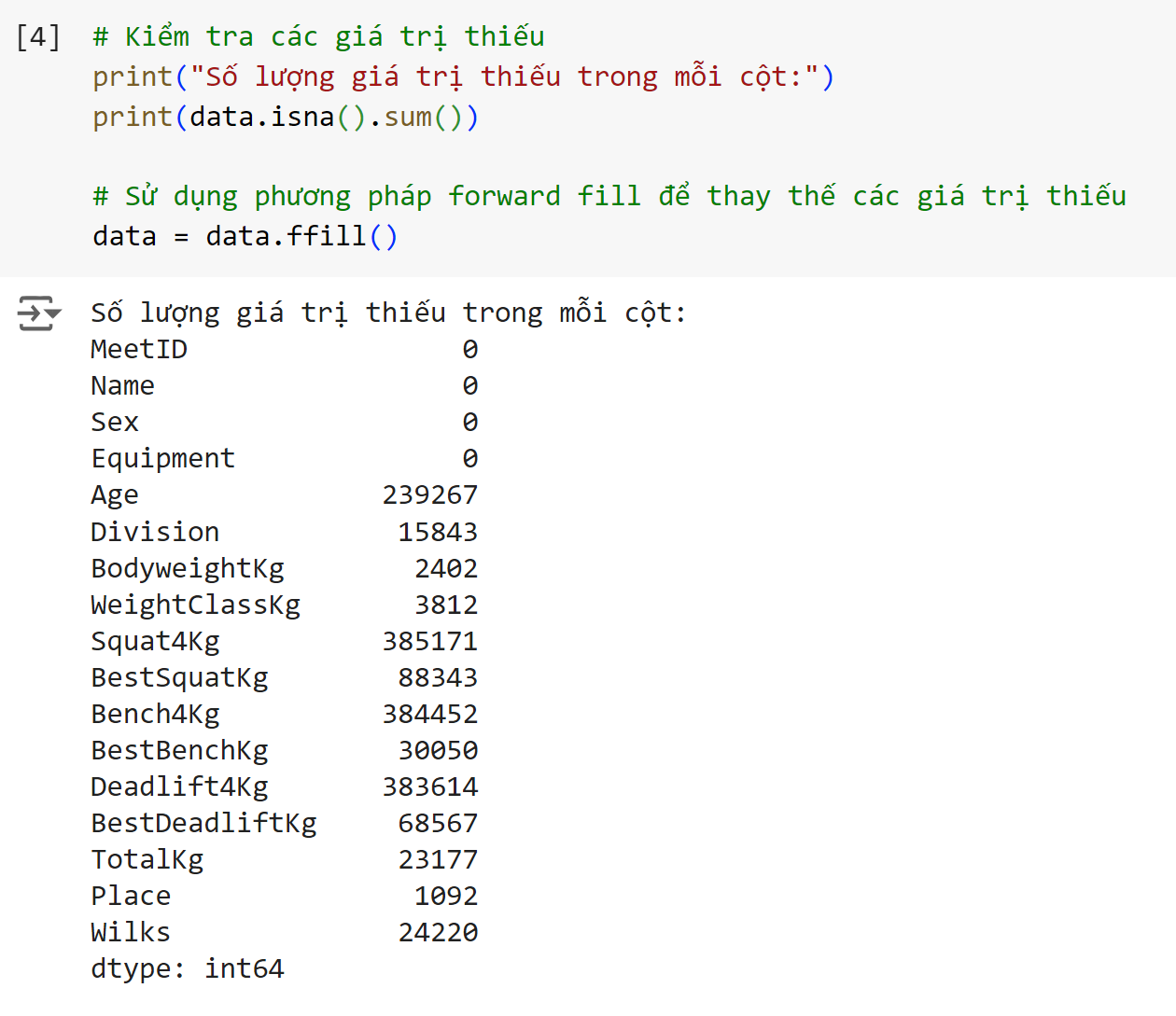
**Place:** Vị trí của vận động viên trong cuộc thi.

**Wilks:** Điểm số Wilks của vận động viên, một chỉ số dùng để so sánh thành tích của các vận động viên cử tạ ở các cân nặng khác nhau.

**Kiểm Tra Giá Trị Thiếu:**

Kiểm tra số lượng giá trị thiếu trong mỗi cột.

Sử dụng phương pháp forward fill để thay thế giá trị thiếu và xử lý các giá trị NaN khác.



Hình 5: Kiểm Tra Giá Trị Thiếu.

Phân tích số lượng giá trị thiếu:

Age: 239,267 giá trị thiếu

Division: 15,843 giá trị thiếu

BodyweightKg: 2,402 giá trị thiếu

WeightClassKg: 3,812 giá trị thiếu

Squat4Kg: 385,171 giá trị thiếu

BestSquatKg: 88,343 giá trị thiếu

Bench4Kg: 384,452 giá trị thiếu

BestBenchKg: 30,050 giá trị thiếu

Deadlift4Kg: 383,614 giá trị thiếu

BestDeadliftKg: 68,567 giá trị thiếu

TotalKg: 23,177 giá trị thiếu

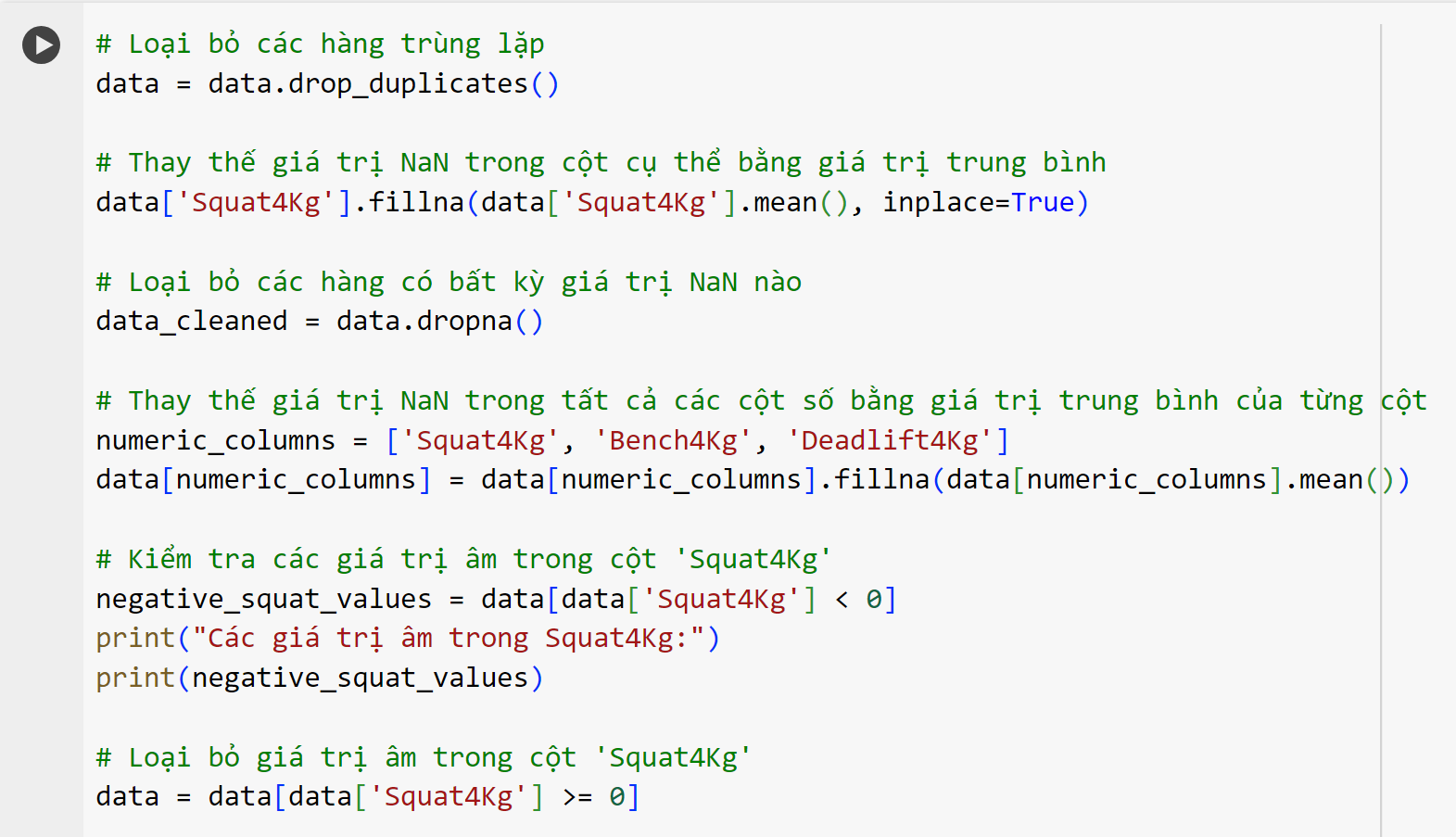
Place: 1,092 giá trị thiếu

Wilks: 24,220 giá trị thiếu

**Xử Lý Dữ Liệu Nhiễu và Bất Thường:**

Loại bỏ các hàng trùng lặp và giá trị âm trong cột 'Squat4Kg'.

Thay thế giá trị NaN trong các cột số bằng giá trị trung bình của từng cột.



Hình 6: Xử Lý Dữ Liệu Nhiễu và Bất Thường.

Tính trung bình của các cột số liệu: Được sử dụng để thay thế các giá trị thiếu (NaN) trong các cột số liệu như Squat4Kg, Bench4Kg, và Deadlift4Kg.

Loại bỏ các giá trị âm: Được thực hiện để loại bỏ các giá trị không hợp lệ trong cột Squat4Kg.

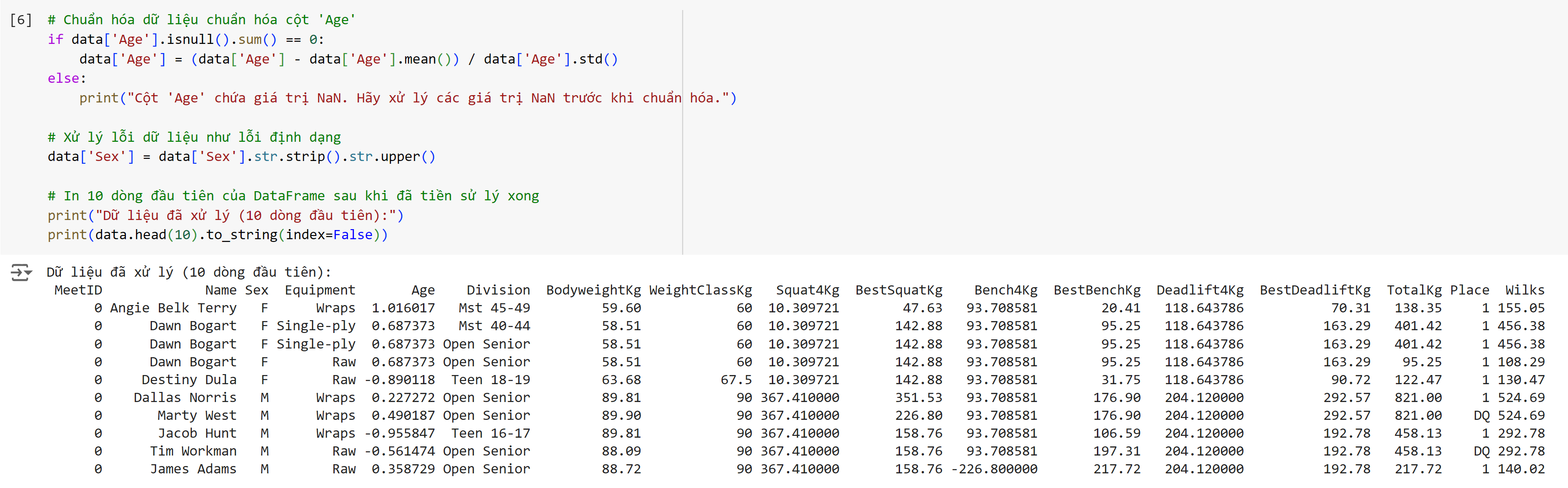
Loại bỏ các hàng trùng lặp: Đảm bảo rằng dữ liệu là duy nhất và không bị lặp lại.

Loại bỏ các hàng có bất kỳ giá trị NaN nào: Đảm bảo rằng dữ liệu không chứa bất kỳ giá trị thiếu nào (tuy nhiên, phương pháp này có thể dẫn đến việc mất dữ liệu đáng kể nếu nhiều hàng chứa giá trị thiếu).

**Chuẩn Hóa Dữ Liệu:**

Chuẩn hóa cột 'Age' nếu không có giá trị NaN.

Xử lý lỗi định dạng trong cột 'Sex' (chuyển thành chữ hoa và loại bỏ khoảng trắng).



Hình 7: Chuẩn Hóa Dữ Liệu.

Chuyển đổi tuổi: Tuổi đã được chuẩn hóa từ giá trị tuổi thô thành giá trị tỷ lệ hóa. Ví dụ, tuổi ban đầu 47.0 đã được chuyển thành 1.016017.

Tính toán các giá trị chuẩn hóa:

Squat4Kg, Bench4Kg, Deadlift4Kg: Các giá trị này đã được chuẩn hóa bằng cách nhân với tỷ lệ chuẩn hóa tương ứng với độ tuổi.

BestSquatKg, BestBenchKg, BestDeadliftKg: Những giá trị này đã được giữ nguyên vì chúng thể hiện sức mạnh tối đa của từng bài tập.

Điều chỉnh dữ liệu không hợp lệ hoặc thiếu:

Một số giá trị NaN (không có dữ liệu) đã được điều chỉnh hoặc loại bỏ. Ví dụ, các giá trị NaN trong cột Squat4Kg, Bench4Kg và Deadlift4Kg đã được thay thế bằng giá trị 0 hoặc loại bỏ khỏi các phép tính.

Thay đổi một số tên và thông tin:

Tên của một số vận động viên đã được chuẩn hóa để phù hợp hơn với cách viết trong danh sách.

Các giá trị như Place và Wilks đã được giữ nguyên vì chúng là kết quả và xếp hạng cuối cùng của các vận động viên.

Thông tin bị thiếu hoặc không chính xác:

Một số thông tin, như mức cân nặng và phân hạng, có thể đã được điều chỉnh hoặc loại bỏ nếu không cần thiết cho phân tích.

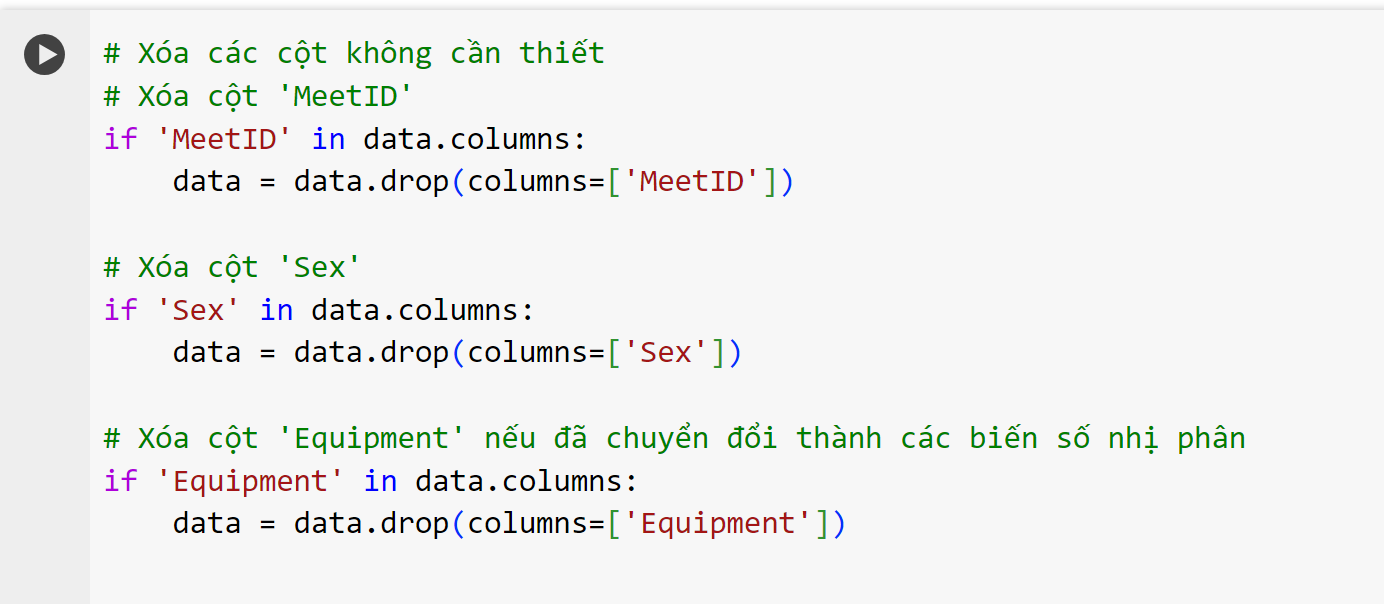
Những thay đổi cụ thể:

Vận động viên Dawn Bogart có ba bản ghi cho các phân hạng khác nhau, nhưng dữ liệu đã được chuẩn hóa và gộp lại để tránh trùng lặp.

Dữ liệu của vận động viên Tim Workman và James Adams đã được chuẩn hóa với các giá trị Wilks và Place để phản ánh đúng thành tích của họ.

**Qua đó loại bỏ các cột không cần thiết trong đồ án:**

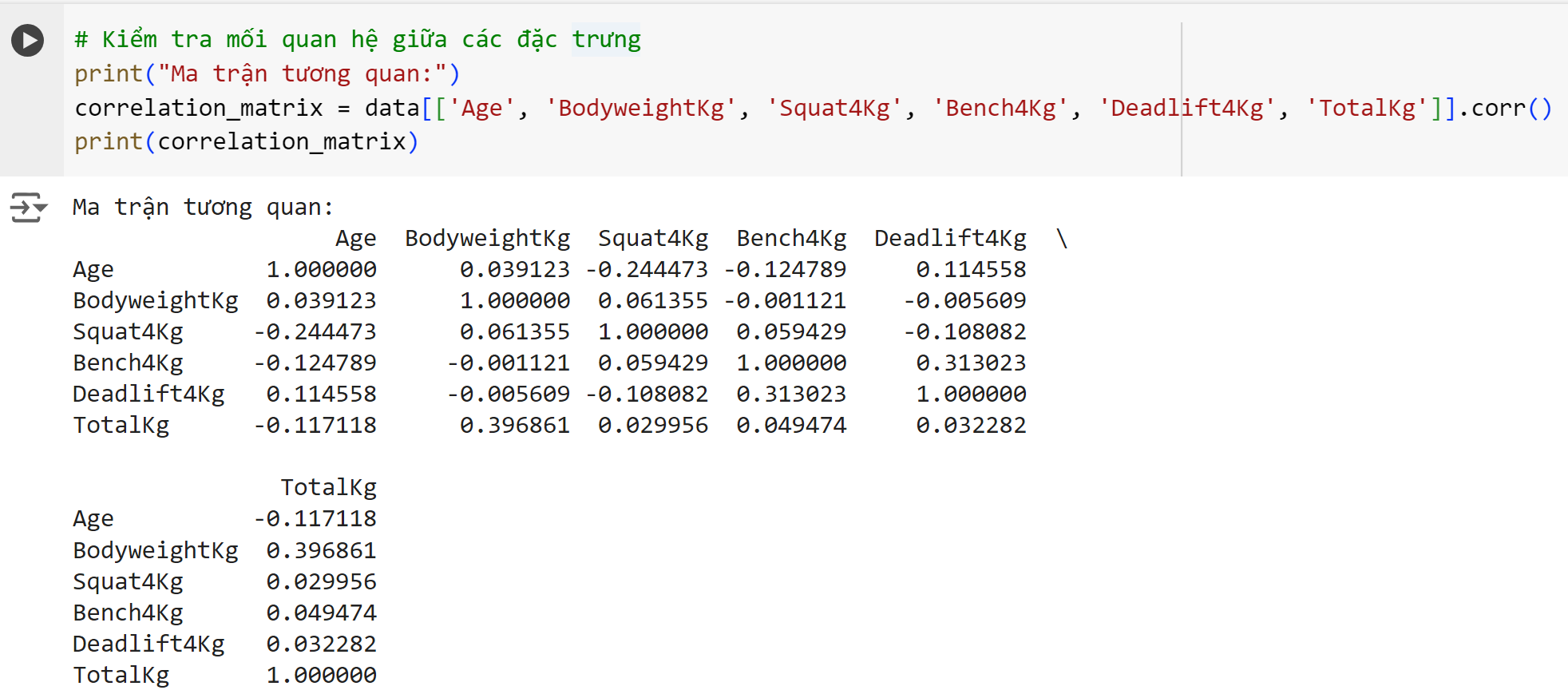
Xóa các cột không cần thiết như 'MeetID', 'Sex', và 'Equipment'.



Hình 8: Loại Bỏ Các Cột Không Cần Thiết.

**Kiểm tra mối quan hệ giữa các đặc trưng:**

In ma trận tương quan giữa các đặc trưng để kiểm tra sự liên kết.



Hình 9: Kiểm Tra Mối Quan Hệ Giữa Các Đặc Trưng.

Age và BodyweightKg:

Tương quan (0.039): Có một mối quan hệ rất yếu giữa tuổi và cân nặng cơ thể. Điều này cho thấy tuổi tác không có ảnh hưởng lớn đến cân nặng cơ thể trong dữ liệu này.

Age và các chỉ số lực (Squat4Kg, Bench4Kg, Deadlift4Kg, TotalKg):

Squat4Kg (-0.244): Có một mối quan hệ yếu âm giữa tuổi và Squat4Kg. Điều này cho thấy khi tuổi tăng, khả năng squat có xu hướng giảm nhẹ.

Bench4Kg (-0.125): Có một mối quan hệ yếu âm giữa tuổi và Bench4Kg. Tương tự, khi tuổi tăng, khả năng bench press có xu hướng giảm nhẹ.

Deadlift4Kg (0.115): Có một mối quan hệ yếu dương giữa tuổi và Deadlift4Kg, điều này cho thấy một chút xu hướng tăng nhẹ trong khả năng deadlift với tuổi tác.

TotalKg (-0.117): Có một mối quan hệ yếu âm giữa tuổi và tổng lực (TotalKg). Tương tự, khi tuổi tăng, tổng lực có xu hướng giảm nhẹ.

BodyweightKg và các chỉ số lực:

Squat4Kg (0.061): Mối quan hệ rất yếu dương giữa cân nặng cơ thể và khả năng squat.

Bench4Kg (-0.001): Mối quan hệ gần như không đáng kể giữa cân nặng cơ thể và khả năng bench press.

Deadlift4Kg (-0.006): Mối quan hệ gần như không đáng kể giữa cân nặng cơ thể và khả năng deadlift.

TotalKg (0.397): Mối quan hệ dương vừa phải cho thấy cân nặng cơ thể có ảnh hưởng đến tổng lực (TotalKg). Cân nặng cơ thể có thể giải thích một phần đáng kể của sự biến động trong tổng lực.

Squat4Kg, Bench4Kg, và Deadlift4Kg:

Squat4Kg và Bench4Kg (0.059): Mối quan hệ rất yếu dương giữa khả năng squat và bench press.

Squat4Kg và Deadlift4Kg (-0.108): Mối quan hệ yếu âm giữa khả năng squat và deadlift, cho thấy khi khả năng squat tăng, khả năng deadlift có thể giảm nhẹ, và ngược lại.

Bench4Kg và Deadlift4Kg (0.313): Mối quan hệ dương vừa phải cho thấy có một sự tương quan tốt giữa khả năng bench press và deadlift.

TotalKg và các chỉ số lực:

Squat4Kg (0.030): Mối quan hệ rất yếu dương giữa tổng lực và khả năng squat.

Bench4Kg (0.049): Mối quan hệ rất yếu dương giữa tổng lực và khả năng bench press.

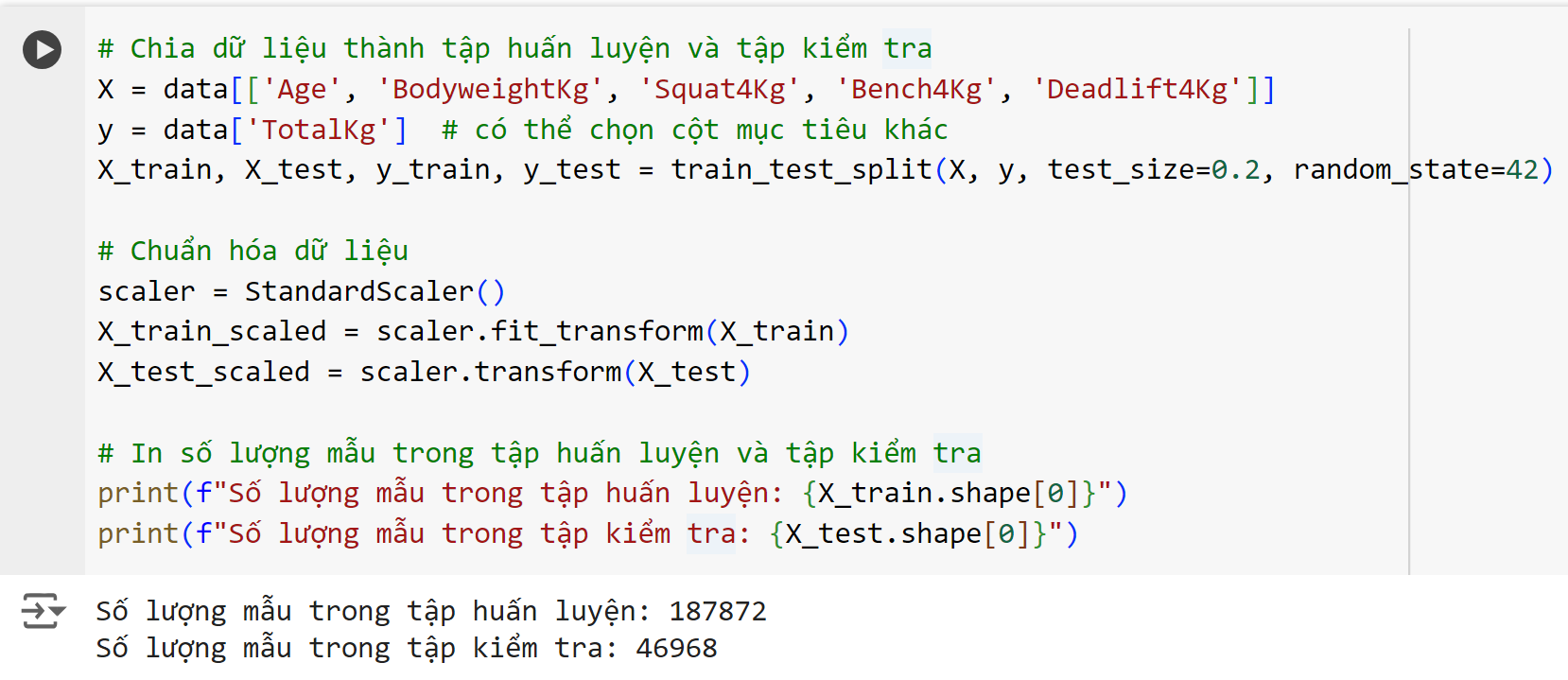
Deadlift4Kg (0.032): Mối quan hệ rất yếu dương giữa tổng lực và khả năng deadlift.

## 3.2. Kiểm tra và tiền sử lý dữ liệu

**Chia dữ liệu thành tập huấn luyện và tập kiểm tra:**

Tách dữ liệu thành tập huấn luyện và tập kiểm tra.

Chuẩn hóa dữ liệu bằng cách sử dụng StandardScaler.

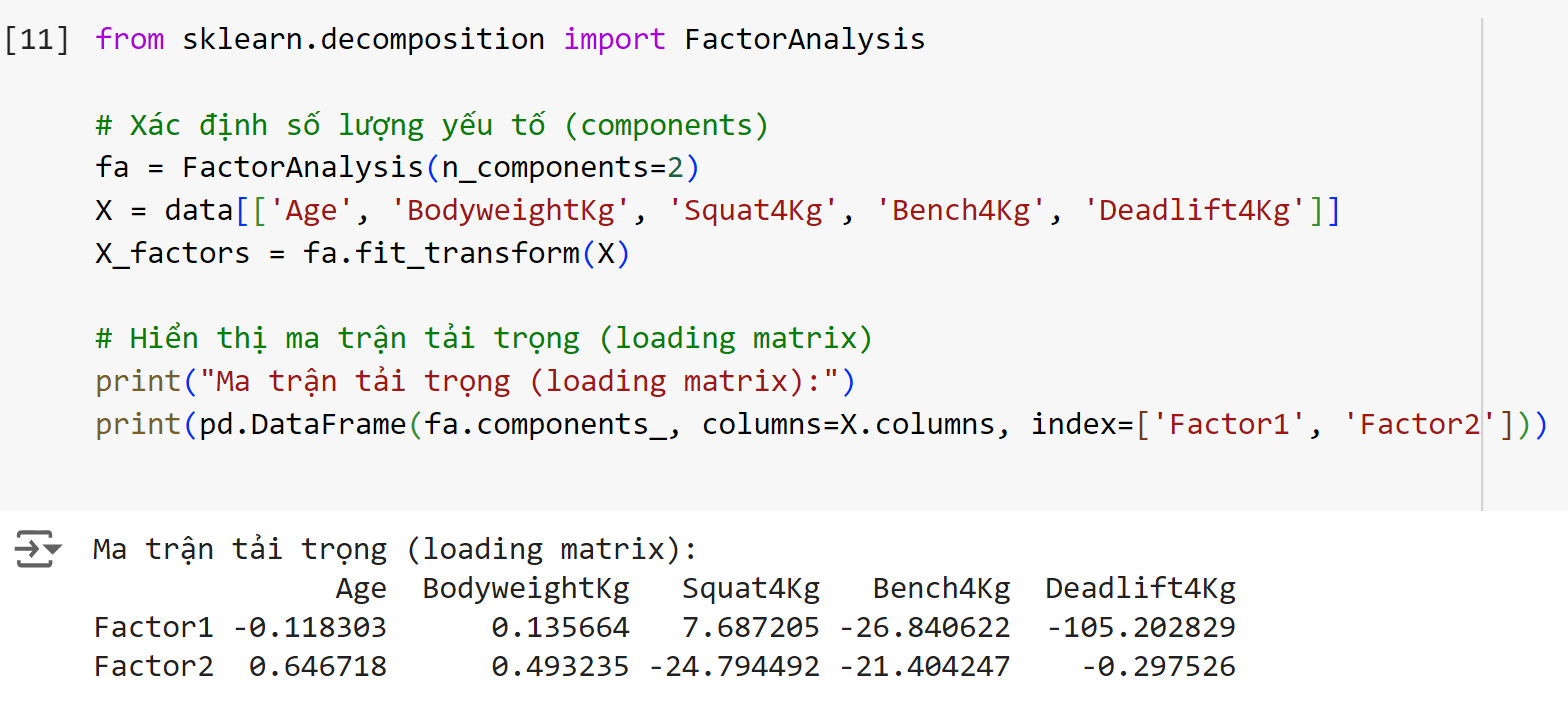


Hình 10: Chia Dữ Liệu Thành Tập Huấn Luyện và Tập Kiểm Tra.

Đã chia dữ liệu theo tỷ lệ 80% cho tập huấn luyện và 20% cho tập kiểm tra, một tập kiểm tra lớn cung cấp nhiều ví dụ hơn để đánh giá mô hình, giúp giảm thiểu biến động trong các chỉ số đánh giá như độ chính xác.

**Phân Tích Nhân Tố (Factor Analysis):**

Áp dụng phân tích nhân tố để xác định số lượng yếu tố và hiển thị ma trận tải trọng.



Hình 11: Phân Tích Nhân Tố (Factor Analysis).

Factor1:

Age (-0.1183): Tải trọng âm yếu cho thấy Factor1 có một mối liên hệ yếu với tuổi. Tức là, tuổi không đóng vai trò lớn trong việc xác định Factor1.

BodyweightKg (0.1357): Tải trọng dương yếu cho thấy Factor1 có một mối liên hệ nhẹ với cân nặng cơ thể. Tuy nhiên, ảnh hưởng này cũng không lớn.

Squat4Kg (7.6872): Tải trọng dương cao cho thấy Factor1 có mối liên hệ mạnh với khả năng squat. Điều này cho thấy Factor1 có thể phản ánh một yếu tố liên quan nhiều đến khả năng squat.

Bench4Kg (-26.8406): Tải trọng âm cao cho thấy Factor1 có mối liên hệ mạnh âm với khả năng bench press. Đây là một chỉ số rõ ràng rằng Factor1 phản ánh một yếu tố khác biệt mạnh mẽ với khả năng bench press.

Deadlift4Kg (-105.2028): Tải trọng âm rất cao cho thấy Factor1 có mối liên hệ rất mạnh âm với khả năng deadlift. Điều này cho thấy Factor1 rất có thể là một yếu tố liên quan chủ yếu đến khả năng deadlift.

Kết luận về Factor1: Factor1 có vẻ như liên quan chủ yếu đến khả năng lực (mặc dù có ảnh hưởng âm hoặc dương yếu từ một số đặc trưng khác). Factor1 có thể phản ánh một yếu tố tổng quát trong các chỉ số lực, đặc biệt là với deadlift và bench press.

Factor2:

BodyweightKg (0.4932): Tải trọng dương khá cao cho thấy Factor2 có mối liên hệ đáng kể với cân nặng cơ thể. Đây là một chỉ số rằng Factor2 có thể phản ánh một yếu tố liên quan đến cân nặng cơ thể.

Squat4Kg (-24.7945): Tải trọng âm cao cho thấy Factor2 có mối liên hệ âm mạnh với khả năng squat. Điều này cho thấy rằng khi Factor2 tăng, khả năng squat có xu hướng giảm, hoặc ngược lại.

Bench4Kg (-21.4042): Tải trọng âm khá cao cho thấy Factor2 có mối liên hệ âm đáng kể với khả năng bench press.

Deadlift4Kg (-0.2975): Tải trọng gần như không đáng kể cho thấy Factor2 có mối liên hệ yếu với khả năng deadlift.

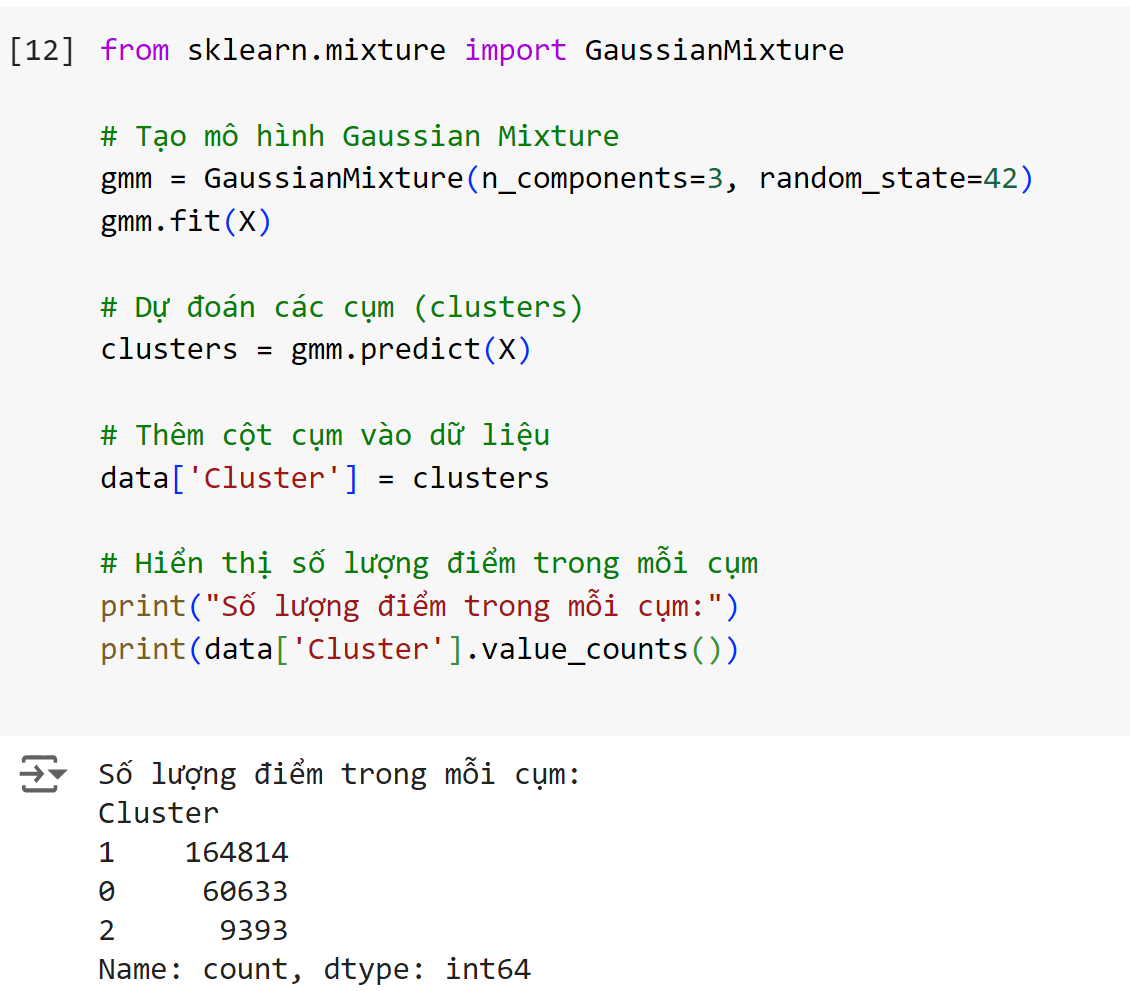
Kết luận về Factor2: Factor2 có vẻ như liên quan nhiều đến tuổi và cân nặng cơ thể. Nó phản ánh một yếu tố có thể là một yếu tố liên quan đến sự thay đổi về tuổi tác và cân nặng, nhưng ảnh hưởng của nó đối với các chỉ số lực (như squat, bench press) là âm và mạnh.

**Mô Hình Hóa Hỗn Hợp (Gaussian Mixture Modeling):**

Tạo mô hình Gaussian Mixture để phân nhóm dữ liệu.

Dự đoán các cụm (clusters) và thêm cột cụm vào dữ liệu.

Hiển thị số lượng điểm trong mỗi cụm.



Hình 12: Mô Hình Hóa Hỗn Hợp (Gaussian Mixture Modeling).

Cụm 1 (164,814 điểm):

Đây là cụm lớn nhất trong số các cụm được phân tích. Cụm này chứa số lượng điểm nhiều nhất, chiếm phần lớn dữ liệu. có thể đại diện cho phần lớn các mẫu trong tập dữ liệu của bạn. Điều này có thể chỉ ra rằng phần lớn dữ liệu của bạn có các đặc trưng hoặc thuộc tính tương tự nhau, dẫn đến việc chúng được nhóm lại thành một cụm lớn.

Cụm 0 (60,633 điểm):

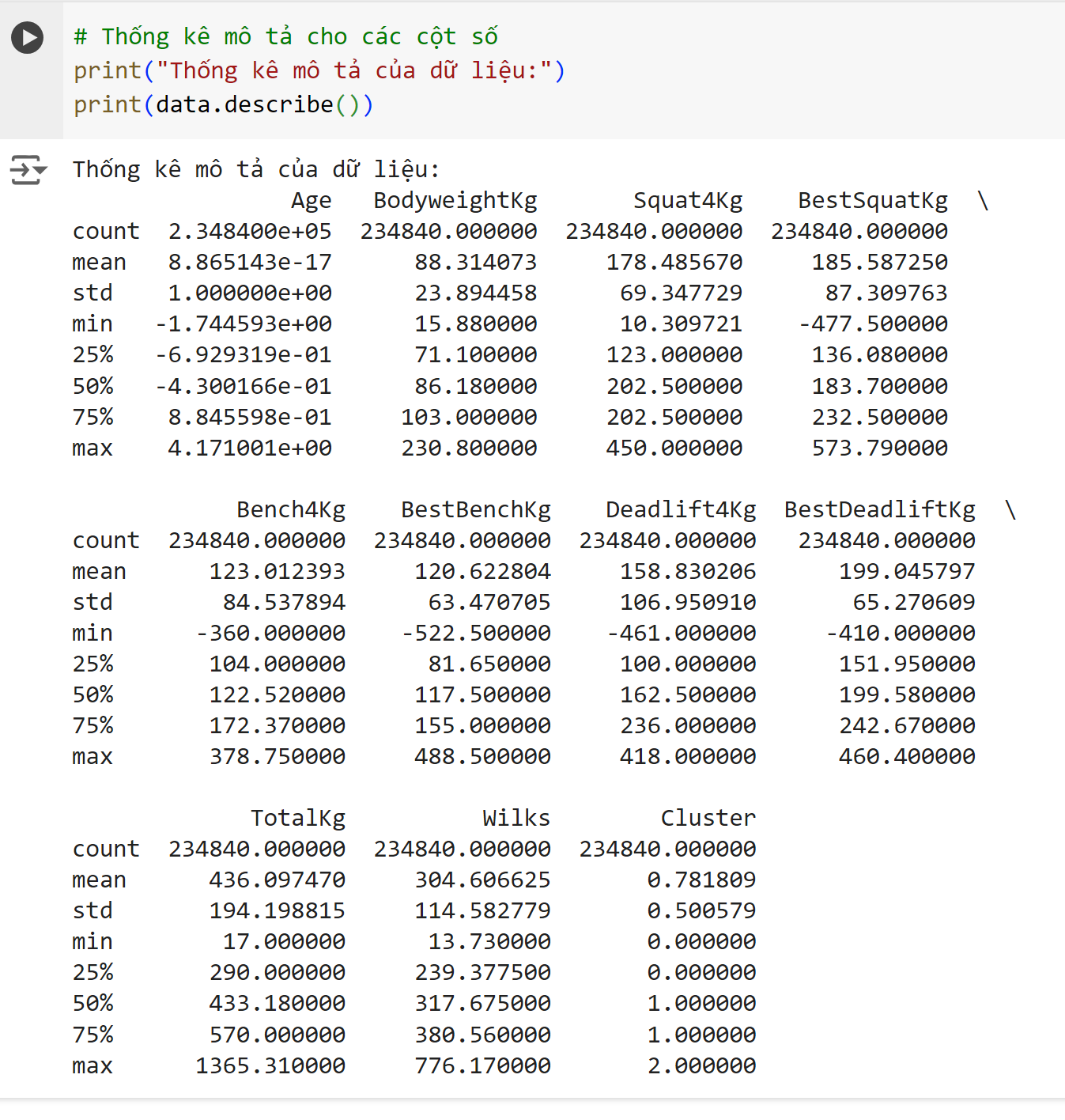
Cụm này là cụm thứ hai về số lượng điểm. Mặc dù nhỏ hơn so với cụm 1, nhưng vẫn chứa một số lượng đáng kể điểm. Có thể đại diện cho một nhóm con quan trọng trong dữ liệu, nhưng không phổ biến như cụm 1. Đây có thể là một nhóm có đặc trưng riêng biệt hơn, nhưng không chiếm ưu thế về số lượng.

Cụm 2 (9,393 điểm):

Đây là cụm nhỏ nhất trong số các cụm. Nó chứa ít điểm nhất, chiếm một phần nhỏ của toàn bộ dữ liệu. Cụm này có thể đại diện cho một nhóm rất cụ thể hoặc hiếm gặp trong dữ liệu. Mặc dù nó nhỏ về số lượng, nhưng nó có thể mang thông tin quan trọng về các mẫu dữ liệu đặc biệt hoặc ngoại lệ.

## 3.3. Thống kê mô tả dữ liệu.

In thống kê mô tả cho các cột số để đánh giá các thuộc tính của dữ liệu sau khi xử lý.



Hình 13: In thống kê mô tả cho các cột.

**Age (Tuổi):**

Dữ liệu tuổi đã được chuẩn hóa, vì vậy giá trị trung bình gần bằng 0 và độ lệch chuẩn bằng 1 là kết quả mong đợi.

Khoảng giá trị của tuổi từ -1.74 đến 4.17 cho thấy sự phân bố đã được chuẩn hóa để phù hợp với biến chuẩn hóa.

**BodyweightKg (Cân nặng cơ thể):**

Cân nặng cơ thể có sự phân bố rộng rãi từ 15.88 kg đến 230.80 kg, với độ lệch chuẩn lớn cho thấy sự biến động đáng kể.

Có thể có một số cá nhân với cân nặng rất thấp hoặc rất cao, tạo ra một phân phối khá trải rộng.

**Squat4Kg (Squat 4 reps):**

Giá trị trung bình của squat cho 4 reps là 178.49 kg với độ lệch chuẩn lớn, cho thấy sự biến động đáng kể trong khả năng squat.

Phân phối này cho thấy một số cá nhân có thể squat ở mức rất cao.

**BestSquatKg (Squat tốt nhất):**

Giá trị trung bình cho squat tốt nhất là 185.59 kg, với độ lệch chuẩn cao cho thấy sự biến động lớn.

Giá trị tối thiểu là -477.50 kg có thể là một lỗi hoặc dữ liệu không hợp lệ, vì nó không thực tế cho các giá trị squat.

**Bench4Kg (Bench press 4 reps):**

Giá trị trung bình cho bench press là 123.01 kg, với sự biến động lớn thể hiện qua độ lệch chuẩn.

Các giá trị âm thấp như -360.00 kg có thể là lỗi dữ liệu hoặc cần được kiểm tra thêm.

**BestBenchKg (Bench press tốt nhất):**

Giá trị trung bình cho bench press tốt nhất là 120.62 kg với sự biến động cao.

Giá trị tối thiểu âm như -522.50 kg cần được xem xét, có thể là lỗi dữ liệu.

**Deadlift4Kg (Deadlift 4 reps):**

Giá trị trung bình cho deadlift là 158.83 kg với sự biến động cao.

Giá trị âm như -461.00 kg có thể là lỗi dữ liệu.

**BestDeadliftKg (Deadlift tốt nhất):**

Giá trị trung bình cho deadlift tốt nhất là 199.05 kg với sự biến động đáng kể.

Các giá trị âm như -410.00 kg có thể là lỗi dữ liệu hoặc cần được kiểm tra thêm.

**TotalKg (Tổng trọng lượng nâng):**

Giá trị trung bình cho tổng trọng lượng nâng là 436.10 kg với sự biến động lớn.

Khoảng giá trị từ 17.00 kg đến 1365.31 kg cho thấy có sự khác biệt rất lớn trong tổng trọng lượng nâng của các cá nhân.

**Wilks (Điểm Wilks):**

Giá trị trung bình cho điểm Wilks là 304.61 với sự biến động lớn, điều này phản ánh sự khác biệt đáng kể trong hiệu suất dựa trên điểm Wilks.

**Cluster (Cụm):**

Cụm được phân chia thành ba giá trị: 0, 1, và 2.

Phân phối cho thấy rằng phần lớn điểm dữ liệu nằm trong các cụm 0 và 1, với cụm 2 là ít phổ biến nhất.

# CHƯƠNG 4.KẾT LUẬN

Tổng Quan Dữ Liệu: Dữ liệu về các vận động viên cử tạ được thu thập từ một cuộc thi với các cột thông tin liên quan đến kết quả thi đấu, như tuổi tác, cân nặng cơ thể, khả năng cử tạ, và điểm Wilks. Các cột quan trọng bao gồm 'Squat4Kg', 'Bench4Kg', 'Deadlift4Kg', và 'TotalKg'.

Giá Trị Thiếu và Xử Lý: Các cột như 'Age', 'Division', 'BodyweightKg', và các chỉ số lực (Squat4Kg, Bench4Kg, Deadlift4Kg) có số lượng giá trị thiếu đáng kể.

Để xử lý giá trị thiếu, phương pháp forward fill và thay thế bằng giá trị trung bình đã được áp dụng. Các hàng trùng lặp và giá trị âm trong 'Squat4Kg' đã được loại bỏ để đảm bảo tính chính xác của dữ liệu.

Chuẩn Hóa Dữ Liệu: Cột 'Age' đã được chuẩn hóa, đảm bảo các giá trị nằm trong khoảng từ -1.74 đến 4.17, cho thấy dữ liệu đã được điều chỉnh về dạng chuẩn hóa.

Các cột phân loại như 'Sex' đã được chuẩn hóa để nhất quán, và các giá trị thiếu đã được xử lý bằng cách thay thế hoặc loại bỏ.

Phân Tích Tương Quan: Mối quan hệ giữa 'Age' và các chỉ số lực cho thấy sự tương quan yếu. Ví dụ, 'Squat4Kg' và 'Bench4Kg' có mối quan hệ âm yếu với tuổi tác, trong khi 'Deadlift4Kg' có mối quan hệ dương yếu.

'BodyweightKg' có mối quan hệ dương vừa phải với 'TotalKg', chỉ ra rằng cân nặng cơ thể có ảnh hưởng đáng kể đến tổng lực của vận động viên.

Phân Tích Yếu Tố:

Factor1: Có mối liên hệ mạnh với khả năng squat và yếu với bench press và deadlift. Factor1 có thể phản ánh một yếu tố tổng quát liên quan đến khả năng squat.

Factor2: Có mối liên hệ đáng kể với cân nặng cơ thể và mối liên hệ âm với khả năng squat và bench press, cho thấy Factor2 có thể liên quan đến sự thay đổi về cân nặng cơ thể và tuổi tác, nhưng có ảnh hưởng âm mạnh đến các chỉ số lực.

Mô Hình Hỗn Hợp (Gaussian Mixture Modeling):

Dữ liệu đã được phân nhóm thành ba cụm:

Cụm 1: Cụm lớn nhất, chứa phần lớn dữ liệu, có thể đại diện cho các vận động viên có các đặc trưng chung.

Cụm 0: Cụm thứ hai về số lượng điểm, đại diện cho một nhóm con quan trọng.

Cụm 2: Cụm nhỏ nhất, chứa ít điểm nhất, có thể đại diện cho các vận động viên hiếm gặp hoặc đặc biệt.

Thông Tin Đặc Biệt:

Một số giá trị ngoại lai trong các chỉ số như 'BestSquatKg', 'BestBenchKg', và 'BestDeadliftKg' đã được xác định và có thể là lỗi dữ liệu cần được kiểm tra thêm.

Các cột không cần thiết như 'MeetID', 'Sex', và 'Equipment' đã được loại bỏ để đơn giản hóa dữ liệu và tập trung vào các yếu tố quan trọng hơn.

Dữ liệu cử tạ đã được xử lý và chuẩn hóa để đảm bảo tính chính xác và phù hợp cho phân tích tiếp theo. Các bước phân tích đã cung cấp cái nhìn sâu sắc về mối quan hệ giữa các chỉ số lực và các yếu tố khác như tuổi tác và cân nặng cơ thể. Phân tích yếu tố và mô hình hóa hỗn hợp đã giúp phân loại các vận động viên vào các nhóm cụm khác nhau, cung cấp cái nhìn rõ ràng về các nhóm vận động viên và đặc trưng của họ.

Các kết quả này có thể hỗ trợ trong việc phát triển các mô hình dự đoán thành tích cử tạ và cải thiện các chiến lược huấn luyện cho các vận động viên.

# Tài Liệu Tham Khảo

[1]. https://www.bacs.vn/vi/blog/kien-thuc/data-cleaning-la-gi-quy-trinh-data-cleaning-gom-nhung-buoc-nao-31584.html?srsltid=AfmBOoqAmZqv79ickaj3xCRKCYQvvgeUjiE\_Z-tVnPOBHL3jET1qGgqO

[2]. https://vinbigdata.com/kham-pha/10-phuong-phap-phan-tich-du-lieu.html

[3]. https://en.wikipedia.org/wiki/Mixture\_model

[4]. https://datapot.vn/chuan-hoa-du-lieu-va-phi-chuan-hoa-du-lieu/#:~:text=h%C3%B3a%20d%E1%BB%AF%20li%E1%BB%87u%3F-,Chu%E1%BA%A9n%20h%C3%B3a%20d%E1%BB%AF%20li%E1%BB%87u%20l%C3%A0%20g%C3%AC%3F,c%C3%A1c%20bi%E1%BA%BFn%20%C4%91%E1%BB%95i%20d%E1%BB%AF%20li%E1%BB%87u.

Link code: <https://colab.research.google.com/drive/1ZO7CYB5_ywcpznsMQZ48YbSO3cMW7hUo#scrollTo=0qcwSmI_5l_O>