Gym Members Exercise Dataset

**1. Tổng quan**

* **Tên dataset**: Gym Members Exercise Dataset
* **Tác giả**: Vala Khorasani
* **Ngày xuất bản**: 06/10/2024
* **Mô tả trên Kaggle**:  
  Dataset này được thiết kế để phân tích các mô hình tập luyện và hiệu suất của các thành viên phòng gym ở các cấp độ kinh nghiệm khác nhau. Nó cung cấp thông tin chi tiết về các đặc điểm thể chất, thói quen tập luyện, và các chỉ số sức khỏe, phù hợp cho các bài toán phân tích dữ liệu, học máy, và phát triển các ứng dụng liên quan đến thể dục.
* **Kích thước**:
  + Số dòng: **973 mẫu** (mỗi mẫu đại diện cho một thành viên phòng gym).
  + Số cột: **15 cột** (bao gồm các đặc trưng và nhãn).
* **Định dạng**: File CSV (gym\_members\_exercise\_tracking.csv).
* **Dung lượng**: Khoảng 82 KB.
* **Giấy phép**: CC0: Public Domain (dữ liệu có thể được sử dụng tự do mà không cần xin phép).

**2. Mục đích và ứng dụng**

Theo mô tả trên Kaggle, dataset này được tạo ra để:

* **Phân tích mô hình tập luyện**: Hiểu cách các yếu tố như tuổi tác, cân nặng, nhịp tim, và loại hình tập luyện ảnh hưởng đến hiệu suất và kinh nghiệm.
* **Xây dựng mô hình học máy**: Phù hợp cho các bài toán phân loại (ví dụ: dự đoán mức độ kinh nghiệm), hồi quy (dự đoán calo tiêu thụ), hoặc phân cụm (nhóm các thành viên theo đặc điểm).
* **Phát triển ứng dụng thể dục**: Hỗ trợ tạo các công cụ gợi ý kế hoạch tập luyện cá nhân hóa dựa trên dữ liệu người dùng.
* **Nghiên cứu sức khỏe**: Khám phá mối quan hệ giữa các chỉ số thể chất (BMI, tỷ lệ mỡ) và thói quen tập luyện.

Dataset này đặc biệt hữu ích cho các nhà phân tích dữ liệu, chuyên gia học máy, và các nhà phát triển ứng dụng thể dục muốn khám phá dữ liệu thực tế trong lĩnh vực gym.

**3. Cấu trúc dữ liệu**

Tập dữ liệu bao gồm **15 cột**, mỗi cột đại diện cho một thuộc tính liên quan đến thành viên phòng gym. Dưới đây là mô tả chi tiết từng cột, bao gồm kiểu dữ liệu và ý nghĩa:

| **Tên cột** | **Mô tả** | **Kiểu dữ liệu** | **Ví dụ giá trị** |
| --- | --- | --- | --- |
| Age | Tuổi của thành viên | Số nguyên | 56, 46, 32 |
| Gender | Giới tính của thành viên | Chuỗi (Male/Female) | Male, Female |
| Weight (kg) | Cân nặng của thành viên (kilogram) | Số thực | 88.3, 74.9, 68.1 |
| Height (m) | Chiều cao của thành viên (mét) | Số thực | 1.71, 1.53, 1.66 |
| Max\_BPM | Nhịp tim tối đa trong buổi tập (beats per minute) | Số nguyên | 180, 179, 167 |
| Avg\_BPM | Nhịp tim trung bình trong buổi tập | Số nguyên | 157, 151, 122 |
| Resting\_BPM | Nhịp tim nghỉ của thành viên | Số nguyên | 60, 66, 54 |
| Session\_Duration (hours) | Thời gian tập luyện mỗi buổi (giờ) | Số thực | 1.69, 1.30, 1.11 |
| Calories\_Burned | Lượng calo tiêu thụ trong buổi tập | Số thực | 1313.0, 883.0, 677.0 |
| Workout\_Type | Loại hình tập luyện (Yoga, HIIT, Cardio, Strength) | Chuỗi | Yoga, HIIT, Cardio |
| Fat\_Percentage | Tỷ lệ mỡ cơ thể (%) | Số thực | 12.6, 33.9, 33.4 |
| Water\_Intake (liters) | Lượng nước uống mỗi ngày (lít) | Số thực | 3.5, 2.1, 2.3 |
| Workout\_Frequency (days/week) | Tần suất tập luyện mỗi tuần (ngày/tuần) | Số nguyên | 4, 3 |
| Experience\_Level | Mức độ kinh nghiệm (1: Beginner, 2: Intermediate, 3: Advanced) | Số nguyên | 1, 2, 3 |
| BMI | Chỉ số khối cơ thể (tính bằng Weight / Height²) | Số thực | 30.20, 32.00, 24.71 |

**Nhận xét về cấu trúc:**

* **Đặc trưng đa dạng**: Bao gồm thông tin nhân khẩu học (Age, Gender), thể chất (Weight, Height, BMI, Fat\_Percentage), và tập luyện (Session\_Duration, Calories\_Burned, Workout\_Type).
* **Nhãn mục tiêu**: Experience\_Level là cột chính được sử dụng cho bài toán phân loại, với 3 giá trị (1, 2, 3). Trong một số mô hình (như XGBoost), nhãn này cần được chuyển thành [0, 1, 2].
* **Không có giá trị thiếu**: Dữ liệu đã được làm sạch, đảm bảo không cần xử lý giá trị thiếu.
* **Kiểu dữ liệu phù hợp**: Các cột số (như Weight, Calories\_Burned) là số thực hoặc số nguyên, trong khi các cột danh mục (Gender, Workout\_Type) là chuỗi, có thể cần mã hóa khi sử dụng trong học máy.

**4. Đặc điểm nổi bật**

Dựa trên phân tích từ trang Kaggle và các bài toán đã thực hiện với dataset này:

* **Kích thước vừa phải**: Với 973 mẫu, dataset đủ lớn để huấn luyện các mô hình học máy cơ bản như KNN, Random Forest, XGBoost, nhưng không quá lớn để gây khó khăn trong xử lý.
* **Không có giá trị thiếu hoặc trùng lặp**: Điều này giúp tiết kiệm thời gian tiền xử lý, cho phép tập trung vào phân tích và xây dựng mô hình.
* **Phân bố nhãn**:
  + Theo các phân tích trước đó, cần kiểm tra phân bố của Experience\_Level để xác định xem có mất cân bằng lớp hay không (ví dụ: nhiều Beginner hơn Advanced).
  + Một phân tích mẫu cho thấy số lượng mẫu của các lớp có thể tương đối cân bằng, nhưng điều này cần được xác nhận qua EDA.
* **Tính đa dạng**:
  + **Loại hình tập luyện**: Bao gồm 4 loại chính (Yoga, HIIT, Cardio, Strength), phản ánh các phong cách tập luyện phổ biến.
  + **Độ tuổi**: Từ trẻ (25) đến trung niên (56), phù hợp với nhiều nhóm đối tượng.
  + **Chỉ số sức khỏe**: Nhịp tim (Max\_BPM, Avg\_BPM, Resting\_BPM), BMI, và tỷ lệ mỡ cung cấp thông tin toàn diện về thể trạng.
* **Tính thực tế**: Các chỉ số như Calories\_Burned, Session\_Duration, và Workout\_Frequency phản ánh dữ liệu thực tế từ phòng gym, giúp mô hình có giá trị ứng dụng cao.

**5. Ví dụ dữ liệu**

Dưới đây là 5 dòng đầu tiên của dataset (trích từ thông tin bạn cung cấp và phù hợp với mô tả trên Kaggle):

**6. Các bài toán đã áp dụng**

Theo ghi chú trên Kaggle và các phân tích liên quan:

* **Phân loại mức độ kinh nghiệm**:
  + Sử dụng Experience\_Level làm nhãn, với các đặc trưng như Calories\_Burned, Session\_Duration, Age, Weight, Avg\_BPM, Resting\_BPM, Workout\_Frequency, Max\_BPM, Fat\_Percentage.
  + Các mô hình như Random Forest đạt accuracy 91.79% sau khi tối ưu siêu tham số (max\_depth=10, n\_estimators=200, min\_samples\_split=5).
* **Dự đoán calo tiêu thụ**:
  + Một số notebook (ví dụ: từ GitHub của sayande01) đã sử dụng dataset này để dự đoán Calories\_Burned bằng hồi quy, đạt RMSE khoảng 40.5.
* **EDA và trực quan hóa**:
  + Phân tích phân bố nhãn, tương quan giữa các đặc trưng (ví dụ: Calories\_Burned và Session\_Duration có tương quan cao).
  + Trực quan hóa như biểu đồ cột (Experience\_Level) và heatmap tương quan.

**7. Điểm mạnh và hạn chế**

**Điểm mạnh:**

* **Dữ liệu sạch**: Không có giá trị thiếu hoặc trùng lặp, tiết kiệm thời gian tiền xử lý.
* **Đặc trưng phong phú**: Bao gồm cả thông tin nhân khẩu học, thể chất, và tập luyện, phù hợp cho nhiều bài toán.
* **Ứng dụng thực tế**: Có thể dùng để xây dựng các công cụ gợi ý tập luyện hoặc phân tích sức khỏe.
* **Kích thước hợp lý**: 973 mẫu đủ để huấn luyện mô hình mà không quá nặng về tính toán.

**Hạn chế:**

* **Quy mô hạn chế**: Chỉ có 973 mẫu, có thể không đủ cho các mô hình học sâu hoặc khi cần phân tích chi tiết hơn.
* **Thiếu thông tin thời gian**: Không có cột liên quan đến thời điểm thu thập dữ liệu, gây khó khăn trong phân tích xu hướng theo thời gian.
* **Phân bố nhãn**: Nếu Experience\_Level không cân bằng (ví dụ: ít mẫu Advanced), có thể cần kỹ thuật như SMOTE để cải thiện hiệu suất mô hình.
* **Thiếu ngữ cảnh bổ sung**: Không có thông tin về cách dữ liệu được thu thập (ví dụ: thực tế hay tổng hợp), có thể ảnh hưởng đến độ tin cậy.

**8. Nguồn gốc và độ tin cậy**

* **Nguồn**: Được cung cấp bởi Vala Khorasani trên Kaggle, một nền tảng uy tín cho dữ liệu học máy.
* **Tính xác thực**: Mặc dù Kaggle không nêu rõ nguồn gốc thu thập (thực tế hay tổng hợp), các phân tích cho thấy dữ liệu có tính nhất quán cao (ví dụ: BMI khớp với công thức Weight / Height²).
* **Cộng đồng sử dụng**: Dataset đã được sử dụng trong các notebook Kaggle và dự án phân tích (ví dụ: phân tích bằng R trên blog crosshyou.info), cho thấy mức độ phổ biến và hữu ích.

[[A red blue and yellow squares

AI-generated content may be incorrect.](https://www.crosshyou.info/entry/2024/10/27/092127)](https://www.crosshyou.info/entry/2024/10/27/092127" \t "_blank)

**9. Kết luận**

"Gym Members Exercise Dataset" là một tập dữ liệu chất lượng, được thiết kế để hỗ trợ các bài toán phân tích và học máy trong lĩnh vực thể dục và sức khỏe. Với 973 mẫu, 15 cột, và các đặc trưng đa dạng, nó đặc biệt phù hợp cho các nhiệm vụ như phân loại mức độ kinh nghiệm, dự đoán calo tiêu thụ, hoặc phân tích thói quen tập luyện. Tuy có một số hạn chế về quy mô và ngữ cảnh, dataset này vẫn là một nguồn tài nguyên quý giá cho các nhà phân tích dữ liệu, chuyên gia học máy, và những người quan tâm đến lĩnh vực gym.