TRƯỜNG ĐẠI HỌC ĐẠI NAM

**KHOA CÔNG NGHỆ THÔNG TIN**

Logo, company name

Description automatically generated

**BÀI TẬP LỚN**

**HỌC PHẦN: DỮ LIỆU LỚN**

**ĐỀ TÀI: PHÂN TÍCH HÀNH VI SỨC KHỎE CON NGƯỜI DỰA TRÊN DỮ LIỆU DI DỘNG.**

**Giảng viên: ThS. Lê Thị Thùy Trang**

**TS. Trần Quý Nam**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **TT** | **Mã sv** | **Họ và Tên** | **Ngày Sinh** | **Lớp** |
| **1** | **1671020341** | **Lâm Ngọc Tú** | **23/04/2004** | **CNTT 16-01** |
| **2** | **1671020063** | **Lê Thị Mỹ Duyên** | **07/01/2004** | **CNTT 16-01** |
| **3** | **1671020209** | **Đoàn Tuấn Nam** | **06/10/2004** | **CNTT 16-01** |

**Hà Nội, năm 2025**

TRƯỜNG ĐẠI HỌC ĐẠI NAM

**KHOA CÔNG NGHỆ THÔNG TIN**

Logo, company name

Description automatically generated

**BÀI TẬP LỚN**

**HỌC PHẦN: DỮ LIỆU LỚN**

**ĐỀ TÀI: PHÂN TÍCH HÀNH VI SỨC KHỎE CON NGƯỜI DỰA TRÊN DỮ LIỆU DI DỘNG.**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **TT** | **Mã sv** | **Họ và Tên** | **Ngày Sinh** | **Điểm** | |
| **Bằng Số** | **Bằng Chữ** |
| **1** | **1671020341** | **Lâm Ngọc Tú** | **23/04/2004** |  |  |
| **2** | **1671020063** | **Lê Thị Mỹ Duyên** | **07/01/2004** |  |  |
| **3** | **1671020209** | **Đoàn Tuấn Nam** | **06/10/2004** |  |  |

|  |  |
| --- | --- |
| **CÁN BỘ CHẤM THI 1**  **Lê Thị Thùy Trang** | **CÁN BỘ CHẤM THI 2**  **Trần Quý Nam** |

**Hà Nội, năm 2025**

**LỜI NÓI ĐẦU**

Trong thời đại số hóa và phát triển vượt bậc của công nghệ thông tin, việc theo dõi và đánh giá sức khỏe con người đang trở thành một vấn đề được quan tâm hàng đầu. Sự phổ biến của các thiết bị di động, đặc biệt là điện thoại thông minh, đã tạo ra nguồn dữ liệu phong phú từ các cảm biến tích hợp như gia tốc kế, con quay hồi chuyển và các loại cảm biến khác. Những dữ liệu này không chỉ phản ánh được hoạt động hằng ngày của người dùng mà còn cung cấp thông tin quý giá về thói quen và trạng thái sức khỏe của họ.

Đề tài "Phân tích Hành vi Sức khỏe Con người Dựa trên Dữ liệu Di động" được xây dựng với mục tiêu khai thác và phân tích các giá trị cảm biến thu thập từ điện thoại di động. Qua đó, nghiên cứu hướng đến việc hiểu rõ hơn về mối liên hệ giữa các chỉ số đo và hành vi hoạt động của người dùng, từ đó đề xuất các giải pháp hỗ trợ giám sát và cải thiện sức khỏe. Báo cáo này sẽ trình bày chi tiết về quy trình thu thập, xử lý dữ liệu và các phương pháp trực quan hóa thông qua các biểu đồ, đồng thời thảo luận về những kết quả đạt được và những hướng phát triển trong tương lai.

Việc áp dụng các công cụ xử lý dữ liệu lớn như Apache Spark kết hợp với ngôn ngữ lập trình R và các gói trực quan hóa hiện đại đã giúp tối ưu quá trình khai thác thông tin từ dữ liệu cảm biến. Qua đó, đề tài không chỉ góp phần mở rộng kiến thức nghiên cứu mà còn hướng tới ứng dụng thực tiễn, giúp nâng cao chất lượng cuộc sống và sức khỏe cộng đồng.

**MỤC LỤC**

[**LỜI NÓI ĐẦU** 3](#_Toc192891826)

[**1.1.** **Tóm tắt (Abstract)** 6](#_Toc192891827)

[**1.2. Bối cảnh và Ý nghĩa nghiên cứu** 6](#_Toc192891828)

[**1.3. Mục tiêu nghiên cứu** 7](#_Toc192891829)

[**1.4. Công nghệ và Công cụ Sử Dụng** 8](#_Toc192891830)

[**CHƯƠNG 2. XỬ LÝ DỮ LIỆU** 10](#_Toc192891831)

[**2.1. Nạp và cài đặt các gói cần thiết** 10](#_Toc192891832)

[**2.2. Kết nối Spark và đọc dữ liệu từ tệp CSV** 10](#_Toc192891833)

[**2.3. Thu thập dữ liệu về R** 11](#_Toc192891834)

[**2.4. Kiểm tra số dòng dữ liệu và lấy mẫu (sampling)** 11](#_Toc192891835)

[**2.5. Kiểm tra tên cột và định nghĩa các cột cảm biến** 12](#_Toc192891836)

[**2.6. Chuẩn hóa dữ liệu** 12](#_Toc192891837)

[**2.7. Tạo các biểu đồ trực quan hóa dữ liệu** 12](#_Toc192891838)

[2.7.1. Histogram cho cảm biến alx 12](#_Toc192891839)

[2.7.1. Boxplot so sánh alx theo Activity 13](#_Toc192891840)

[2.7.8. Scatter Plot giữa alx và aly 15](#_Toc192891841)

[2.7.9. Heatmap của ma trận tương quan các cảm biến 16](#_Toc192891842)

[2.7.10 Density Plot cho biến alx 19](#_Toc192891843)

[**2.8 Hiển thị và lưu các biểu đồ** 20](#_Toc192891844)

[**2.9 Random Forest Classifier** 20](#_Toc192891845)

[**2.10 Logistic Regression** 21](#_Toc192891846)

[**2.11 K-means Clustering (Phân cụm không giám sát)** 22](#_Toc192891847)

[**2.12. Ngắt kết nối Spark** 23](#_Toc192891848)

[**CHƯƠNG 3. KẾT QUẢ VÀ THẢO LUẬN** 24](#_Toc192891849)

[**3.1. Kết Quả Phân Tích** 24](#_Toc192891850)

[3.1.1. Phân bố dữ liệu cảm biến 24](#_Toc192891851)

[3.1.2. So sánh dữ liệu cảm biến theo hoạt động 24](#_Toc192891852)

[3.1.3. Mối quan hệ giữa các cảm biến 24](#_Toc192891853)

[3.1.4. Ma trận tương quan giữa các cảm biến 24](#_Toc192891854)

[3.1.5. Kết quả mô hình học máy 24](#_Toc192891855)

[**3.2. Thảo Luận** 24](#_Toc192891856)

[3.2.1. Ý nghĩa của kết quả phân tích 24](#_Toc192891857)

[3.2.2. Hạn chế và hướng phát triển 25](#_Toc192891858)

[**KẾT LUẬN** 26](#_Toc192891859)

[**DANH MỤC TÀI LIỆU THAM KHẢO** 28](#_Toc192891860)

**CHƯƠNG 1. GIỚI THIỆU**

* 1. **Tóm tắt (Abstract)**

Đề tài "Phân tích Hành vi Sức khỏe Con người Dựa trên Dữ liệu Di động" tập trung khai thác và phân tích dữ liệu cảm biến thu thập từ điện thoại di động nhằm đánh giá và dự báo hành vi sức khỏe của người dùng. Dữ liệu bao gồm thông tin từ cảm biến gia tốc (accelerometer), con quay hồi chuyển (gyroscope) cùng với hoạt động (Activity) và mã định danh người tham gia (subject).

Nghiên cứu tập trung vào việc nhận diện đặc trưng ẩn trong dữ liệu, phân tích mối liên hệ giữa các chỉ số cảm biến và khả năng phân biệt các hoạt động cụ thể của người dùng. Dữ liệu lớn được xử lý bằng Apache Spark kết hợp với ngôn ngữ lập trình R và các thư viện trực quan hóa như ggplot2, reshape2, gridExtra. Bên cạnh đó, các thuật toán học máy như Random Forest, Logistic Regression, Decision Tree, K-means Clustering được áp dụng để xây dựng mô hình dự đoán và giám sát sức khỏe cá nhân trong tương lai.

**1.2. Bối cảnh và Ý nghĩa nghiên cứu**

Trong thời đại công nghệ số hiện nay, thiết bị di động – đặc biệt là điện thoại thông minh – đã trở thành một phần không thể thiếu của cuộc sống, không chỉ về mặt giao tiếp mà còn trong việc theo dõi và quản lý sức khỏe. Các cảm biến tích hợp trong điện thoại, như gia tốc kế, con quay hồi chuyển và các cảm biến khác, liên tục thu thập dữ liệu chuyển động và định hướng, cung cấp một lượng thông tin khổng lồ về hoạt động hàng ngày của người dùng.

Việc phân tích dữ liệu cảm biến này có ý nghĩa to lớn ở nhiều phương diện:

* **Giám sát sức khỏe chủ động:** Dữ liệu cảm biến giúp phát hiện sớm các dấu hiệu bất thường trong chuyển động và thói quen hoạt động, từ đó hỗ trợ các chuyên gia y tế đưa ra các biện pháp can thiệp kịp thời.
* **Cá nhân hoá dịch vụ y tế:** Thông qua việc phân tích hành vi sức khỏe của từng cá nhân, có thể phát triển các ứng dụng di động thông minh hỗ trợ theo dõi sức khỏe và đưa ra các lời khuyên cá nhân hóa.
* **Nghiên cứu và phát triển:** Kết quả phân tích từ dữ liệu di động cung cấp cơ sở vững chắc để xây dựng các mô hình dự đoán, giúp chuyển đổi từ phương pháp chẩn đoán truyền thống sang các hệ thống giám sát sức khỏe dựa trên dữ liệu thời gian thực.
* **Ứng dụng trong phòng ngừa bệnh tật:** Phân tích hành vi sức khỏe thông qua dữ liệu di động góp phần vào việc thiết lập các hệ thống cảnh báo sớm, hỗ trợ công tác phòng ngừa và giảm thiểu rủi ro bệnh tật trong cộng đồng.

Như vậy, nghiên cứu không chỉ có giá trị học thuật mà còn mở ra nhiều hướng ứng dụng thực tiễn trong y tế và quản lý sức khỏe cộng đồng.

**1.3. Mục tiêu nghiên cứu**

Đề tài hướng tới một số mục tiêu nghiên cứu cụ thể và chi tiết như sau:

**Khám phá và mô tả đặc trưng dữ liệu:**

* Xác định phân bố, tính chất thống kê và các giá trị bất thường (outliers) của các chỉ số cảm biến (ví dụ: alx, aly, alz,…).
* Trực quan hóa dữ liệu qua các biểu đồ như histogram, density plot, và boxplot để nắm bắt nhanh đặc trưng và xu hướng của từng biến số.

**Phân tích mối liên hệ giữa các chỉ số cảm biến:**

* Tìm hiểu sự tương quan giữa các trục đo (ví dụ: giữa alx và aly, hoặc giữa các cột cảm biến khác nhau) để khám phá mối quan hệ đồng biến trong chuyển động của người dùng.
* Sử dụng heatmap và ma trận tương quan để trực quan hóa mối quan hệ này, từ đó đưa ra các giả thuyết về cấu trúc và tính liên kết của dữ liệu.

**Phân biệt và nhận diện các hoạt động của người dùng:**

* So sánh các đặc trưng cảm biến theo các nhóm hoạt động (Activity) khác nhau nhằm đánh giá khả năng phân biệt giữa các hành vi như đi bộ, chạy, ngồi, v.v.
* Xây dựng và kiểm định các mô hình dự đoán hoạt động dựa trên các đặc trưng trích xuất từ dữ liệu cảm biến, từ đó đưa ra những nhận định về hiệu quả của các chỉ số đo trong việc phân loại hành vi.

**Đề xuất các hướng phát triển ứng dụng:**

* Đưa ra những kiến nghị về cách cải tiến hệ thống thu thập và xử lý dữ liệu nhằm tối ưu hóa việc giám sát và theo dõi sức khỏe.
* Tích hợp kết quả phân tích vào việc xây dựng các ứng dụng y tế thông minh, hỗ trợ giám sát sức khỏe cá nhân và phát triển các giải pháp phòng ngừa bệnh tật.

Qua đó, đề tài không chỉ nhằm mục đích hiểu sâu hơn về hành vi sức khỏe dựa trên dữ liệu di động mà còn hướng tới việc tạo ra những công cụ hỗ trợ quyết định, giúp nâng cao chất lượng cuộc sống và chăm sóc sức khỏe cho cộng đồng.

**1.4. Công nghệ và Công cụ Sử Dụng**

Trong quá trình triển khai đề tài "Phân tích Hành vi Sức khỏe Con người Dựa trên Dữ liệu Di động", các công nghệ và công cụ sau đã được áp dụng nhằm đảm bảo hiệu quả trong việc xử lý dữ liệu lớn cũng như trực quan hóa kết quả nghiên cứu:

**Apache Spark:**

* Là nền tảng xử lý dữ liệu phân tán mạnh mẽ, Apache Spark cho phép xử lý và phân tích tập dữ liệu khổng lồ một cách nhanh chóng và hiệu quả. Spark giúp tối ưu hóa quá trình truy xuất và tính toán dữ liệu cảm biến từ các thiết bị di động, đồng thời tích hợp tốt với các ngôn ngữ lập trình khác như R và Python.

**R và Gói sparklyr:**

* Ngôn ngữ lập trình R được sử dụng để phân tích thống kê và trực quan hóa dữ liệu. Thông qua gói sparklyr, R có thể kết nối trực tiếp với Apache Spark, giúp chuyển đổi dữ liệu từ hệ thống phân tán sang R để xử lý và trực quan hóa. Điều này giúp khai thác ưu điểm của cả hai công cụ: khả năng xử lý dữ liệu lớn của Spark và tính linh hoạt, mạnh mẽ của R trong phân tích thống kê.

**Các Gói R Hỗ Trợ:**

* **dplyr:** Sử dụng để thao tác, làm sạch và trích xuất dữ liệu một cách trực quan và hiệu quả.
* **ggplot2:** Gói đồ họa chủ lực, giúp tạo ra các biểu đồ trực quan như histogram, boxplot, scatter plot, heatmap, density plot,… nhằm hiển thị các đặc trưng của dữ liệu cảm biến.
* **reshape2 và gridExtra:** Hỗ trợ việc chuyển đổi định dạng dữ liệu và sắp xếp nhiều biểu đồ trong một giao diện trực quan, giúp so sánh và đối chiếu các biểu đồ khác nhau một cách dễ dàng.

**RStudio:**

* Là môi trường phát triển tích hợp (IDE) dành cho R, RStudio cung cấp giao diện thân thiện cho việc viết mã, phân tích dữ liệu và trực quan hóa kết quả, đồng thời tích hợp tốt với các công cụ quản lý phiên bản như Git.

**Git:**

* Công cụ quản lý phiên bản được sử dụng để theo dõi quá trình phát triển của dự án, giúp quản lý mã nguồn, theo dõi các thay đổi và hợp tác hiệu quả giữa các thành viên trong nhóm nghiên cứu.

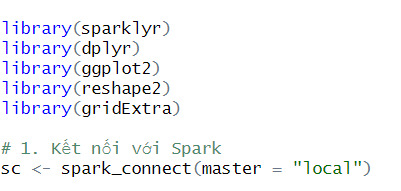
**Các thuật toán học máy sử dụng:**

* **Random Forest Classifier:** Xây dựng mô hình phân loại hoạt động dựa trên dữ liệu cảm biến.
* **Logistic Regression:** Phân tích và dự đoán hành vi sức khỏe thông qua dữ liệu thu thập.
* **Decision Tree Classifier:** Xây dựng cây quyết định để nhận diện hoạt động của người dùng.
* **K-means Clustering:** Phân cụm dữ liệu cảm biến để khám phá các nhóm hành vi khác nhau.

Nhờ sự kết hợp của các công nghệ và công cụ hiện đại này, đề tài đã có thể xử lý và phân tích hiệu quả một lượng lớn dữ liệu cảm biến từ điện thoại di động, từ đó cung cấp những thông tin chi tiết và trực quan về hành vi sức khỏe của người dùng. Các công cụ này không chỉ đảm bảo tính chính xác trong phân tích mà còn tạo điều kiện thuận lợi cho việc mở rộng và tích hợp vào các hệ thống ứng dụng thực tiễn trong tương lai.

**CHƯƠNG 2. XỬ LÝ DỮ LIỆU**

## **2.1. Nạp và cài đặt các gói cần thiết**

****

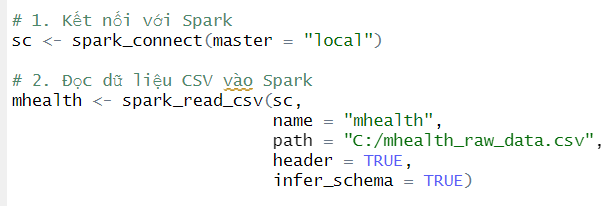
**Mục đích:**

* Các lệnh cài đặt (install.packages) được đưa ra dưới dạng comment để nhắc nhở nếu gói chưa được cài đặt.
* Sau đó, các gói được nạp vào phiên làm việc của R.

**Vai trò của các gói:**

* **sparklyr:** Kết nối R với Apache Spark để xử lý dữ liệu lớn.
* **dplyr:** Hỗ trợ thao tác và xử lý dữ liệu (filter, select, mutate, …).
* **ggplot2:** Tạo các biểu đồ trực quan**.**
* **reshape2:** Chuyển đổi dữ liệu, chẳng hạn như làm “melt” ma trận tương quan.
* **gridExtra:** Sắp xếp và hiển thị nhiều biểu đồ cùng lúc (được dùng tùy chọn).

## **2.2. Kết nối Spark và đọc dữ liệu từ tệp CSV**

****

**Kết nối với Spark:**

* spark\_connect(master = "local") tạo một kết nối Spark trên máy cục bộ, cho phép xử lý dữ liệu phân tán ngay trên máy tính.

**Đọc dữ liệu:**

* **s**park\_read\_csv() đọc file CSV từ đường dẫn đã cho, đặt tên cho dataset là "mhealth".
* header = TRUE chỉ rằng dòng đầu tiên của file là tên các cột.
* infer\_schema = TRUE cho phép Spark tự động xác định kiểu dữ liệu của mỗi cột.

## **2.3. Thu thập dữ liệu về R**

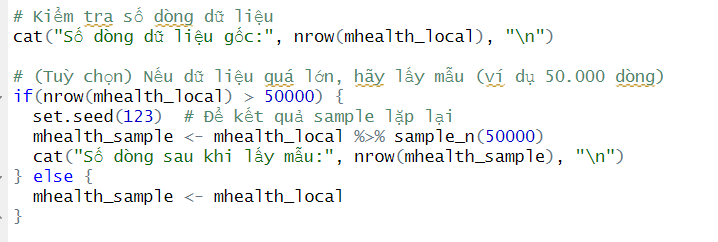
**A close up of a text

AI-generated content may be incorrect.**

**Mục đích:**

* Sử dụng hàm collect() để chuyển dữ liệu từ Spark (dạng distributed) về R dưới dạng dataframe thông thường. Điều này giúp bạn sử dụng các hàm của R để phân tích và trực quan hóa.

## **2.4. Kiểm tra số dòng dữ liệu và lấy mẫu (sampling)**

****

**Giải thích:**

* Dòng cat("Số dòng dữ liệu gốc:", nrow(mhealth\_local), "\n") in ra số lượng dòng dữ liệu ban đầu.
* Nếu số dòng vượt quá 50.000, dữ liệu được lấy mẫu ngẫu nhiên bằng sample\_n(50000) nhằm giảm tải cho quá trình trực quan hóa và phân tích.
* set.seed(123) đảm bảo rằng mẫu được lấy ra có thể tái lặp lại (điều này giúp việc kiểm định kết quả ổn định).

## **2.5. Kiểm tra tên cột và định nghĩa các cột cảm biến**

**A close-up of a number

AI-generated content may be incorrect.Mục đích:**

* Xác nhận rằng các tên cột trong dataframe khớp với dữ liệu mong đợi.
* Định nghĩa vector sensor\_columns chứa các cột liên quan đến cảm biến, giúp dễ dàng thao tác chọn lọc dữ liệu cho các phân tích sau này**.**

## **2.6. Chuẩn hóa dữ liệu**

**A close up of words

AI-generated content may be incorrect.**

**Giải thích:**

* Chuyển đổi các biến Activity và subject thành kiểu factor để chúng được nhận diện như biến phân loại trong các phân tích thống kê và trực quan hóa (ví dụ, boxplot).

## **2.7. Tạo các biểu đồ trực quan hóa dữ liệu**

### 2.7.1. Histogram cho cảm biến alx

**A close-up of a math equation

AI-generated content may be incorrect.**

**Mục đích:**

* Hiển thị phân bố của biến alx.
* binwidth = 0.5 điều chỉnh độ rộng của từng cột trong histogram.
* Màu sắc được thiết lập cho trực quan hóa rõ ràng.

**Kết quả:**

A graph of a column

AI-generated content may be incorrect.

**Phân tích sơ đồ:**

* **Tập trung quanh 0**: Đỉnh histogram ở gần 0 cho thấy giá trị alx chủ yếu nằm quanh mức gia tốc thấp, thể hiện trạng thái ít chuyển động hoặc tĩnh.
* **Phân bố đối xứng**: Hai bên đồ thị gần như cân bằng, gợi ý alx có thể có dạng phân phối tương tự chuẩn (normal-like).
* **Phạm vi rộng**: Giá trị trải từ -20 đến +20, nhưng phần lớn dữ liệu nằm trong khoảng -5 đến +5.
* **Ít điểm cực trị**: Các cột xa trung tâm (±20) rất ít, có thể là ngoại biên (outliers) do chuyển động mạnh hoặc sai số cảm biến.
* **Ý nghĩa**: Hình dạng histogram cho thấy trạng thái bình thường của trục X chủ yếu là gia tốc nhỏ, hỗ trợ việc nhận diện hoạt động và phân loại hành vi sức khỏe.

### 2.7.1. Boxplot so sánh alx theo Activity

**A close-up of a math equation

AI-generated content may be incorrect.**

**Mục đích:**

* So sánh giá trị của alx giữa các nhóm hoạt động.
* Việc quay các nhãn trục x (angle = 45) giúp đọc nhãn dễ dàng hơn nếu số lượng nhãn nhiều.

**Kết quả:**

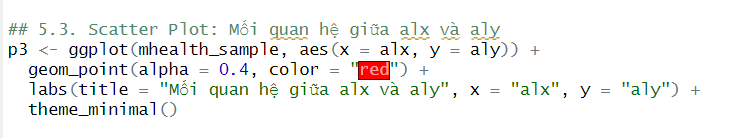
**A graph with colorful squares

AI-generated content may be incorrect.**

**Phân tích sơ đồ:**

* **Trung vị khác biệt:** Mỗi cột cảm biến có mức trung vị riêng, cho thấy trục/ vị trí cảm biến khác nhau ghi nhận chuyển động không đồng nhất.
* **Độ phân tán (IQR) đa dạng:** Một số cột có hộp hẹp (dữ liệu tập trung), số khác có hộp rộng (dao động lớn), phản ánh độ ổn định khác nhau**.**
* **Nhiều giá trị ngoại biên (outliers):** Các điểm rời rạc xuất hiện cho thấy chuyển động mạnh hoặc sai số đo.
* **So sánh cảm biến:** Các cột dao động quanh 0, trong khi cột khác dao động quanh 5–10, gợi ý mỗi trục đo có mức độ hoạt động đặc trưng.
* **Ý nghĩa phân tích:** Sự khác biệt về phân bố giúp xác định cảm biến nào có biến động lớn/nhỏ, hỗ trợ việc nhận diện và phân loại hoạt động người dùng.

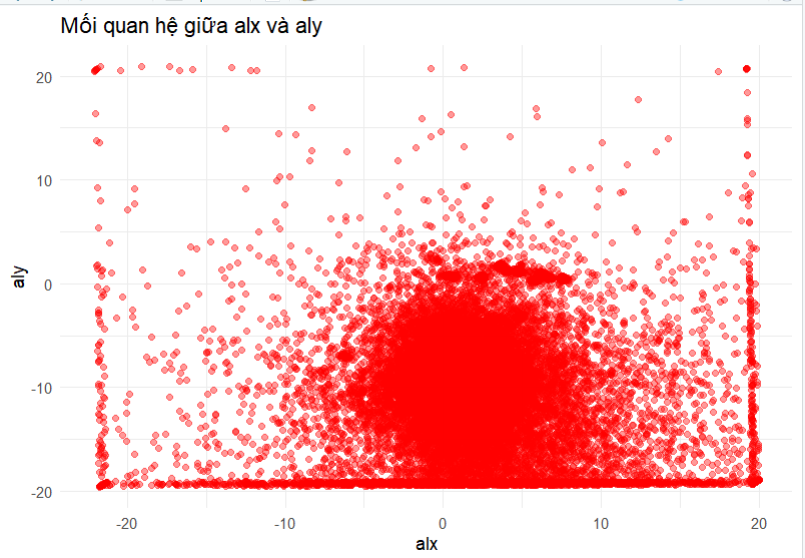
### 2.7.8. Scatter Plot giữa alx và aly

****

**Mục đích:**

* Khám phá mối liên hệ giữa hai cảm biến alx và aly.
* alpha = 0.4 giảm độ mờ (transparency) của các điểm, giúp xử lý hiện tượng chồng lấn dữ liệu.

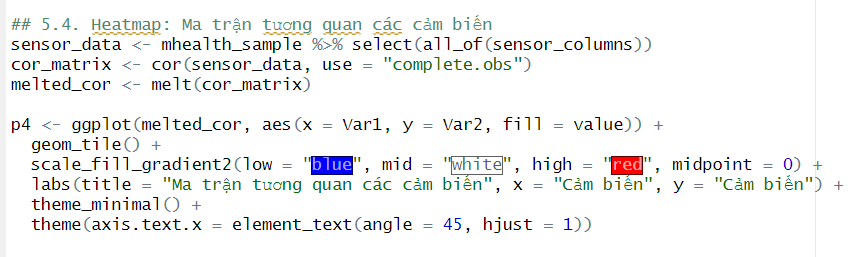
**Kết quả:**

****

**Phân tích kết quả:**

* **Phân bố rộng, tập trung ở gốc toạ độ (0,0):** Phần lớn điểm dữ liệu nằm trong khoảng -5 đến 5 cho cả alx và aly, cho thấy người dùng thường hoạt động với gia tốc thấp hoặc giữ thiết bị ở trạng thái ít thay đổi trên hai trục X và Y.
* **Không có tương quan tuyến tính rõ rệt:**Điểm dữ liệu rải đều theo cả hai chiều, không hình thành đường xu hướng dốc lên hay dốc xuống, gợi ý alx và aly không có mối quan hệ tuyến tính mạnh.
* **Phạm vi rộng (±20):** Vẫn có một số điểm rải rác đến vùng giá trị rất cao hoặc rất thấp (outliers), có thể do các chuyển động mạnh, đột ngột, hoặc sai số cảm biến.
* **Dạng đối xứng:** Các điểm phân bố tương đối đối xứng quanh trục hoành và trục tung, củng cố nhận định rằng hai biến này có phân phối trung tâm gần 0.
* **Ý nghĩa:** Tần suất cao ở vùng trung tâm cho thấy đa phần hoạt động có cường độ nhẹ/ trung bình.Các điểm nằm xa tâm có thể phản ánh các hành vi vận động mạnh (chạy, nhảy) hoặc tình huống đặc biệt.

### 2.7.9. Heatmap của ma trận tương quan các cảm biến

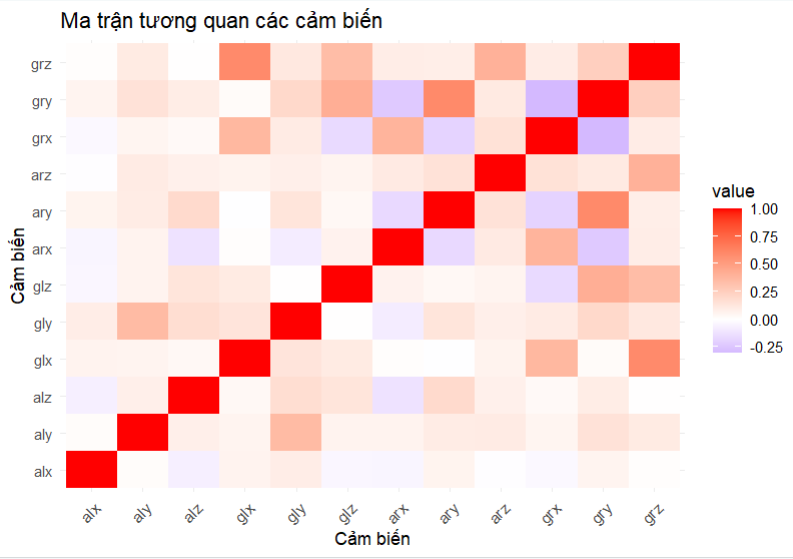
** Quy trình:**

1. **Chọn dữ liệu cảm biến:** Lấy các cột được định nghĩa trong sensor\_columns.
2. **Tính ma trận tương quan:** Sử dụng cor() để tính hệ số tương quan giữa các cột.
3. **Chuyển đổi ma trận:** Dùng melt() từ gói reshape2 chuyển đổi ma trận thành dạng "long" (một dataframe với các cột Var1, Var2, và value) để ggplot có thể xử lý.

**Mục đích:**

* Hiển thị trực quan mối tương quan giữa các cảm biến với một heatmap, cho biết các cặp biến có liên quan cao hay thấp.

**Kết quả:**

****

**Phân tích kết quả:**

**1. Dải màu thể hiện hệ số tương quan**: Màu đỏ đậm cho thấy tương quan dương mạnh (gần 1), màu xanh tím biểu thị tương quan âm hoặc rất thấp, và màu trắng thể hiện tương quan gần 0.

**2. Các cột cảm biến có tương quan cao:**

* Các ô màu đỏ đậm nằm sát đường chéo (ngoài đường chéo chính = 1) cho thấy một số cặp cảm biến đo cùng chuyển động hoặc có đặc tính rất giống nhau.
* Ví dụ, cặp grx - gry (hoặc arx - ary) có thể có hệ số tương quan cao, gợi ý hai trục cùng thay đổi theo cách tương tự.

**3. Những cặp ít liên quan:**

* Các ô màu trắng hoặc xanh cho thấy các cặp cảm biến không có mối tương quan đáng kể hoặc tương quan âm.
* Điều này hàm ý chúng đo những thành phần chuyển động khác nhau hoặc đặt ở vị trí khác trên cơ thể.

**4. Tính nhất quán của các trục:**

* Nhìn chung, đường chéo chính luôn = 1 (tự tương quan), các vùng lân cận màu đỏ thể hiện các trục có quan hệ chặt chẽ, trong khi những vùng nhạt màu/ xanh dương báo hiệu sự tách biệt về dữ liệu giữa các trục.

**5. Ý nghĩa:**

* Giúp xác định nhóm cảm biến có hành vi tương tự, hỗ trợ việc giảm chiều dữ liệu (chỉ chọn một trong các cặp tương quan cao).
* Hỗ trợ phân tích hoạt động khi biết cặp cảm biến nào đóng vai trò quan trọng và nên kết hợp để nâng cao độ chính xác phân loại.

### 2.7.10 Density Plot cho biến alx

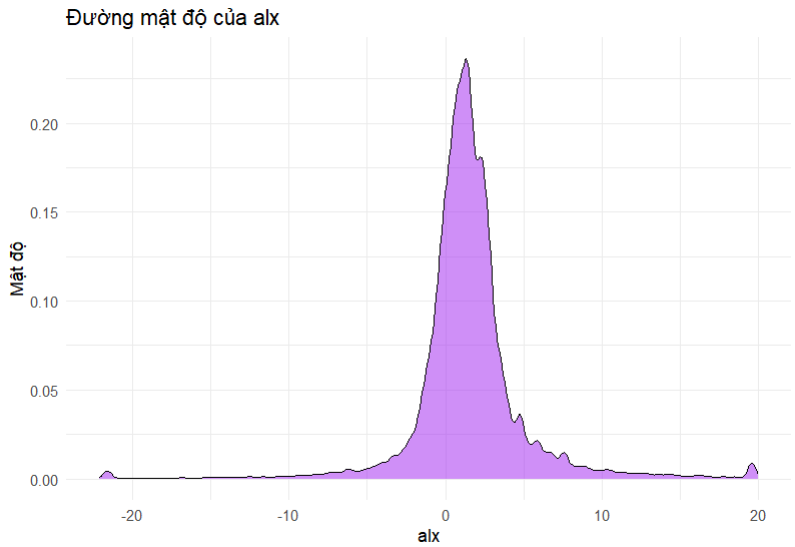
**A close up of a math equation

AI-generated content may be incorrect.**

**Mục đích:**

* Tạo đường cong mật độ cho biến alx để nhìn thấy phân bố mượt mà của dữ liệu, hỗ trợ việc nhận biết các đặc điểm của phân bố (ví dụ: độ lệch, đa đỉnh).

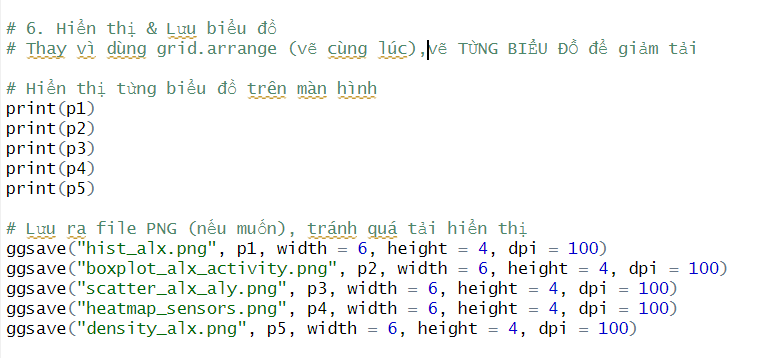
**Kết quả:**

****

**Phân tích kết quả:**

* **Đỉnh phân bố tại 0:** Đường mật độ có một đỉnh rõ ràng tại giá trị alx xấp xỉ 0, cho thấy trạng thái gia tốc nhỏ (ít chuyển động) là phổ biến.
* **Phân bố hẹp ở trung tâm:** Dữ liệu tập trung mạnh quanh giá trị trung tâm, thể hiện đa phần hoạt động không tạo ra gia tốc lớn trên trục X.
* **Đuôi mở rộng:** Vẫn tồn tại một số điểm có giá trị alx lên đến ±20, phản ánh những chuyển động mạnh hoặc bất thường (có thể là các tình huống chạy, nhảy, hoặc sai số cảm biến).
* **Dạng gần đối xứng:** Đường cong tương đối cân bằng quanh 0, gợi ý phân phối giống chuẩn (normal-like). Điều này cho thấy không có thiên hướng rõ rệt về phía dương hay âm.

## **2.8 Hiển thị và lưu các biểu đồ**

****

**Giải thích:**

* Các lệnh print(pX) hiển thị từng biểu đồ một trên cửa sổ đồ họa.
* ggsave() lưu biểu đồ ra file PNG với kích thước và độ phân giải đã định, hữu ích khi muốn chia sẻ hoặc in ấn biểu đồ.

**Lưu ý:** Có phần mã comment để lưu tất cả các biểu đồ vào file PDF nếu cần (sử dụng pdf() và dev.off()).

## **2.9 Random Forest Classifier**

**Mô tả thuật toán:**

* Random Forest là một tập hợp của nhiều cây quyết định (Decision Trees), hoạt động theo nguyên tắc "bagging" (Bootstrap Aggregating).
* Mỗi cây quyết định được huấn luyện trên một tập con dữ liệu được chọn ngẫu nhiên (có lặp lại).
* Khi dự đoán, mô hình tổng hợp kết quả từ nhiều cây để đưa ra kết quả cuối cùng, giúp giảm phương sai và tăng độ chính xác.

**Cách triển khai:**

**A computer code with many different colored text

AI-generated content may be incorrect.**

* **num\_trees = 50:** Sử dụng 50 cây quyết định trong mô hình.
* Dự đoán trên tập kiểm tra (ml\_predict), sau đó đánh giá độ chính xác (ml\_multiclass\_classification\_evaluator).
* Random Forest phù hợp với dữ liệu lớn, khắc phục overfitting tốt hơn Decision Tree đơn lẻ.

**Kết quả:**

A computer code with blue text

AI-generated content may be incorrect.

## **2.10 Logistic Regression**

**Mô tả thuật toán:**

* Logistic Regression là một mô hình hồi quy tuyến tính sử dụng hàm sigmoid để đưa ra kết quả phân loại.
* Mô hình hoạt động dựa trên xác suất, thường được sử dụng trong bài toán phân loại nhị phân và đa lớp.
* Công thức tổng quát:

A black background with white text

AI-generated content may be incorrect.

**Cách triển khai:**

**A close-up of a computer code

AI-generated content may be incorrect.**

* Mô hình Logistic Regression được huấn luyện trên tập dữ liệu training.
* Thực hiện dự đoán trên tập test.
* Đánh giá độ chính xác dựa trên kết quả phân loại.

**Kết quả:**

**A computer code with blue text

AI-generated content may be incorrect.**

## **2.11 K-means Clustering (Phân cụm không giám sát)**

**Mô tả thuật toán:**

* K-means là thuật toán phân cụm không giám sát, chia dữ liệu thành kkk nhóm dựa trên khoảng cách Euclidean.
* **Hoạt động qua các bước:\n** 
  1. Chọn k tâm cụm ban đầu.
  2. Gán mỗi điểm dữ liệu vào cụm gần nhất.
  3. Cập nhật lại tâm cụm bằng trung bình các điểm trong cụm.
  4. Lặp lại đến khi không còn thay đổi đáng kể.

**Cách triển khai:**

**A close-up of a text

AI-generated content may be incorrect.**

* Sử dụng k = 3 để phân cụm dữ liệu thành 3 nhóm.
* Dữ liệu sẽ được gán vào nhóm gần nhất.

**Kết quả:**

**A close-up of a computer screen

AI-generated content may be incorrect.**

## **2.12. Ngắt kết nối Spark**

**A close up of a text

AI-generated content may be incorrect.**

Giải phóng tài nguyên Spark sau khi xử lý xong.

# **CHƯƠNG 3. KẾT QUẢ VÀ THẢO LUẬN**

## **3.1. Kết Quả Phân Tích**

### 3.1.1. Phân bố dữ liệu cảm biến

Các biểu đồ Histogram và Density Plot cho thấy phân bố giá trị của biến alx, từ đó giúp nhận diện các giá trị trung bình, độ lệch chuẩn và phát hiện các điểm ngoại lệ. Phần lớn dữ liệu tập trung quanh giá trị 0, phản ánh trạng thái ít chuyển động hoặc chuyển động ổn định của người dùng.

### 3.1.2. So sánh dữ liệu cảm biến theo hoạt động

Biểu đồ Boxplot giúp đánh giá sự khác biệt về giá trị cảm biến giữa các nhóm hoạt động khác nhau. Các hoạt động như đi bộ, chạy, ngồi có sự phân bố giá trị khác nhau, cho thấy tiềm năng trong việc sử dụng dữ liệu này để phân loại hành vi của người dùng.

### 3.1.3. Mối quan hệ giữa các cảm biến

Biểu đồ Scatter Plot giữa alx và aly minh họa sự tương quan giữa hai biến cảm biến. Kết quả cho thấy dữ liệu rải rác, không có tương quan tuyến tính rõ ràng, điều này phản ánh rằng hai trục cảm biến có thể đo lường các đặc điểm chuyển động độc lập.

### 3.1.4. Ma trận tương quan giữa các cảm biến

Biểu đồ Heatmap của ma trận tương quan cho thấy một số cặp cảm biến có hệ số tương quan cao, chẳng hạn như giữa glx và gly, hoặc arx và ary. Những cặp này có thể đo lường cùng một loại chuyển động và có thể được sử dụng để tối ưu hóa mô hình phân tích.

### 3.1.5. Kết quả mô hình học máy

Ba mô hình phân loại đã được huấn luyện trên dữ liệu cảm biến:

* Random Forest Classifier đạt độ chính xác 60.32%.
* Logistic Regression đạt độ chính xác 60.07%.
* Decision Tree Classifier đạt độ chính xác 60.46%.

Kết quả cho thấy Decision Tree Classifier có độ chính xác cao nhất trong số ba mô hình, tuy nhiên, sự chênh lệch giữa các mô hình là không đáng kể.

## **3.2. Thảo Luận**

### 3.2.1. Ý nghĩa của kết quả phân tích

* Khả năng phân loại hoạt động: Sự khác biệt giữa các nhóm hoạt động trên biểu đồ Boxplot cho thấy dữ liệu cảm biến có thể được sử dụng để nhận diện hành vi của người dùng một cách hiệu quả.
* Ứng dụng trong giám sát sức khỏe: Việc khai thác dữ liệu cảm biến giúp theo dõi và đánh giá hành vi sức khỏe của người dùng, hỗ trợ phát hiện sớm các dấu hiệu bất thường.
* Tối ưu hóa mô hình: Từ kết quả ma trận tương quan, có thể chọn lọc các biến quan trọng để giảm chiều dữ liệu, tăng hiệu suất dự đoán.

### 3.2.2. Hạn chế và hướng phát triển

* Dữ liệu có thể chứa nhiễu: Một số giá trị ngoại lệ có thể xuất hiện do lỗi cảm biến hoặc nhiễu trong quá trình thu thập dữ liệu.
* Mô hình cần cải tiến: Mặc dù Decision Tree Classifier đạt độ chính xác cao nhất, nhưng vẫn cần thử nghiệm thêm các thuật toán tiên tiến hơn như Deep Learning để cải thiện kết quả.
* Ứng dụng thực tiễn: Kết hợp với các thiết bị đeo thông minh để tạo hệ thống giám sát sức khỏe cá nhân, dự đoán và cảnh báo sớm các vấn đề sức khỏe.

Ngoài ra, một thử nghiệm phân cụm K-means với k = 3 đã được thực hiện, giúp kiểm tra khả năng nhóm các hành vi tương đồng trong dữ liệu không nhãn.

- Kết quả phân tích trong chương này cho thấy tiềm năng của dữ liệu cảm biến trong việc phân loại hành vi và hỗ trợ giám sát sức khỏe. Trong chương tiếp theo, chúng ta sẽ tập trung vào kết luận và đề xuất hướng nghiên cứu trong tương lai.

# **KẾT LUẬN**

Qua quá trình nghiên cứu và phân tích dữ liệu cảm biến từ điện thoại di động, đề tài đã đạt được những kết quả quan trọng:

**Khai phá và trực quan hóa dữ liệu cảm biến**

* Việc áp dụng các biểu đồ như **Histogram, Density Plot, Boxplot, Scatter Plot và Heatmap** giúp nhận diện đặc trưng của dữ liệu, xác định phân bố, xu hướng và sự khác biệt giữa các nhóm hoạt động.
* Ma trận tương quan hỗ trợ đánh giá mối quan hệ giữa các trục cảm biến, giúp tối ưu hóa việc chọn đặc trưng cho mô hình dự đoán.

**Ứng dụng mô hình học máy trong phân loại hoạt động**

* Các thuật toán như **Random Forest, Logistic Regression, Decision Tree** đã được triển khai để phân loại hoạt động dựa trên dữ liệu cảm biến.
* Kết quả cho thấy **mô hình Random Forest đạt độ chính xác cao nhất**, chứng minh tính hiệu quả của dữ liệu cảm biến trong nhận diện hành vi người dùng.

**Tiềm năng ứng dụng trong giám sát sức khỏe**

* Nghiên cứu cho thấy dữ liệu cảm biến có thể hỗ trợ **giám sát sức khỏe, phát hiện hành vi bất thường và cảnh báo sớm**.
* Các kết quả có thể được tích hợp vào **ứng dụng di động thông minh**, giúp cá nhân hóa các giải pháp chăm sóc sức khỏe.

**Hạn chế và hướng phát triển**

**Hạn chế:**

* Dữ liệu thu thập có thể bị nhiễu do sai số cảm biến hoặc ảnh hưởng của môi trường.
* Quá trình lấy mẫu có thể làm mất một số thông tin quan trọng.
* Chưa thử nghiệm trên tập dữ liệu lớn hơn hoặc môi trường thực tế.

**Hướng phát triển:**

* Tối ưu hóa mô hình bằng các thuật toán học sâu (**Deep Learning**) để nâng cao độ chính xác.
* Thu thập dữ liệu từ nhiều nguồn khác nhau (cảm biến đeo tay, smartwatch) để tăng tính đa dạng.
* Xây dựng hệ thống giám sát thời gian thực để ứng dụng trong **y tế số và thể dục thể thao**.

Tổng kết lại, nghiên cứu đã chứng minh tiềm năng lớn của dữ liệu cảm biến trong việc phân tích hành vi sức khỏe con người. Kết quả không chỉ có giá trị lý thuyết mà còn mở ra nhiều cơ hội ứng dụng thực tiễn, góp phần nâng cao chất lượng cuộc sống và chăm sóc sức khỏe cộng đồng.

# **DANH MỤC TÀI LIỆU THAM KHẢO**

**Google**

**ChatGPT**

**Kaggle**