Big Data Course

Capstone Project   
Final Report

For students (instructor review required)

ⓒ2023 SAMSUNG. All rights reserved.

Samsung Electronics Corporate Citizenship Office holds the copyright of this document.

This document is a literary property protected by copyright law so reprint and reproduction without permission are prohibited.

To use this document other than the curriculum of Samsung Innovation Campus, you must receive written consent from copyright holder.

|  |
| --- |
| Chủ đề số 11: Phân tích dữ liệu chăm sóc sức khỏe từ các thiết bị đeo |

Date (19/08/2024)

TEAM BIGDATA CAPSTONE 01

Lê Nghĩa Dũng

Nguyễn Thị Tú Lan

Trần Thùy Ninh

Ngọ Quốc Minh

Hoàng Thị Kim Thoa

Trần Thị Yến Nhi

Nguyễn Thu Trang

Menu

1. Giới Thiệu

1.1. Thông tin cơ bản

1.2. Động lực và mục tiêu

1.3. Thành viên và phân công vai trò

1.4. Lịch trình và các sự kiện quan trọng

2. Thực hiện dự án

2.1. Mô tả kịch bản mô phỏng

2.2. Lựa chọn và mô tả bộ dữ liệu

2.3. Đường ống nhập dữ liệu

2.4. Xử lý chuyển đổi dữ liệu

2.5. Truy vấn và hiểu biết dữ liệu

3. Kết quả

3.1. Mã nhập dữ liệu và code

3.2. Mã chuyển đổi dữ liệu và code

3.3. Mô tả và mẫu chuyển đổi dữ liệu

3.4. Trực quan hóa dữ liệu của kết quả truy vấn

4.  Tác động dự kiến

4.1. Thành tựu và lợi ích

4.2. Cải tiến trong tương lai

5. Đánh giá và nhận xét của thành viên nhóm

6. Giáo viên đánh giá và nhận xét

1. Giới Thiệu

# Thông tin cơ bản

* Dự án tập trung vào việc phân tích dữ liệu chăm sóc sức khỏe từ các thiết bị đeo là gia tốc tuyến tính 03 trục, vận tốc góc 03 trục từ gia tốc kế và con quay hồi chuyển tích hợp để đưa ra dự đoán về tư thế vận động của con người cụ thể là 06 hoạt động (Đi bộ, Đi lên, Đi xuống, Ngồi, Đứng và Nằm). Dự án sử dụng công nghệ lưu trữ Big Data Hadoop và các công cụ phân tích, học máy từ Spark Mllib. Thông qua dự đoán về tư thế giúp đưa ra gợi ý và đánh giá về sức khỏe cho người dùng

# Động lực và mục tiêu

* Động lực : Thông qua việc phân tích có thể tăng cường sự hiểu biết về hoạt động và tư thế của con người trong các tình huống khác nhau có thể giúp cải thiện sức khỏe và sự an toàn. Với sự phổ biến của các thiết bị đeo, việc phân tích dữ liệu từ các cảm biến này sẽ cung cấp thông tin quan trọng về hành vi và chuyển động của con người, giúp phát hiện sớm các vấn đề sức khỏe hoặc tình trạng nguy hiểm.
* Mục tiêu :
  + Phân tích dữ liệu chăm sóc sức khỏe từ các thiết bị đeo từ bộ dữ liệu: Human Activity Recognition Using Smartphones Data Set
  + Sử dụng thành thạo hệ thống lưu trữ Big Data Hadoop và ứng dụng vào việc lưu trữ dữ liệu hoạt động từ các thiết bị đeo.
  + Ứng dụng Spark MLlib vào phân tích và dự đoán các hoạt động của con người.
  + Qua đó đưa ra đánh giá hiệu suất của các mô hình dự đoán.

# Thành viên và phân công vai trò

* Cả nhóm cùng thực hiện các nhiệm vụ cụ thể sau:
  + Tìm hiểu về đề tài phân tích dữ liệu chăm sóc sức khỏe từ các thiết bị đeo.
  + Thu thập dữ liệu từ bộ dữ liệu: Human Activity Recognition Using Smartphones Data Set.
  + Áp dụng thuật toán Select giảm chiều dữ liệu.
  + Sử dụng Spark chuyển dữ liệu về dạng Vector
  + Scaling dữ liệu --> Chọn thuật toán phù hợp cho vào mô hình
  + Xây dựng mô hình
  + Đánh giá mô hình

# Lịch trình và mốc thời gian

* Tuần 1:
  + Tìm hiểu về đề tài phân tích dữ liệu chăm sóc sức khỏe từ các thiết bị đeo.
  + Thu thập dữ liệu từ bộ dữ liệu: Human Activity Recognition Using Smartphones Data Set.
  + Áp dụng thuật toán SelectKBest giảm chiều dữ liệu.
* Tuần 2 :
  + Sử dụng Spark chuyển dữ liệu về dạng Vector
  + Scaling dữ liệu --> Chọn thuật toán phù hợp cho vào mô hình
* Tuần 3 :
  + Xây dựng mô hình
  + Đánh giá mô hình

1. Thực Hiện Dự Án

# 2.1. Mô tả kịch bản mô phỏng

## Môi trường Mô phỏng:

* + Thiết bị Đeo: Sử dụng smartphone (Samsung Galaxy S II) gắn trên thắt lưng của người dùng, với cảm biến gia tốc và con quay hồi chuyển được kích hoạt để thu thập dữ liệu.
  + Thử nghiệm: Thực hiện mô phỏng trong môi trường phòng thí nghiệm với 10 tình nguyện viên, mỗi người thực hiện 6 hoạt động khác nhau: Đi bộ, Đi lên cầu thang, Đi xuống cầu thang, Ngồi, Đứng, và Nằm.

## Tình huống và Tình trạng:

* + Tình huống: Các tình nguyện viên sẽ thực hiện các hoạt động cụ thể trong một khoảng thời gian xác định, ví dụ như mỗi hoạt động kéo dài 30 giây.
  + Tình trạng: Môi trường phòng thí nghiệm sẽ được kiểm soát để giảm thiểu sự can thiệp từ các yếu tố bên ngoài như tiếng ồn và sự di chuyển không mong muốn.

## Các Thông số và Điều kiện Thử nghiệm:

* + Dữ liệu Đầu vào:
    - Dữ liệu cảm biến gia tốc (trục X, Y, Z): Thu thập dữ liệu từ cảm biến gia tốc với tần số lấy mẫu 50Hz.
    - Dữ liệu con quay hồi chuyển (trục X, Y, Z): Thu thập dữ liệu từ cảm biến con quay hồi chuyển với tần số lấy mẫu 50Hz.
  + Tham số và Cài đặt:
    - Cửa sổ Thời gian: Sử dụng cửa sổ trượt rộng 2.56 giây với 50% trùng lặp (128 đọc mỗi cửa sổ) để phân tích dữ liệu.
    - Tiền xử lý Dữ liệu: Áp dụng bộ lọc Butterworth với tần số cắt 0.3 Hz để tách thành phần gia tốc cơ thể và trọng lực.

## Mục tiêu của Mô phỏng:

* + Đánh giá Hiệu suất:
    - Nhận diện chính xác: Xác định khả năng của hệ thống trong việc phân loại chính xác các hoạt động và tư thế từ dữ liệu cảm biến.
    - Đánh giá hiệu suất mô hình: Đánh giá các chỉ số hiệu suất như độ chính xác, độ nhạy và độ đặc hiệu của mô hình dự đoán.
  + Xác định Vấn đề: Phân tích lỗi:
    - Phát hiện các lỗi hoặc sai sót trong quá trình nhận diện hoạt động và tư thế.
    - Khắc phục vấn đề: Xác định các vấn đề trong quy trình tiền xử lý dữ liệu hoặc cấu hình mô hình dự đoán.

## Kết quả Dự kiến và Đánh giá:

* + Kết quả Mong đợi:
    - Độ chính xác cao: Dự đoán các hoạt động và tư thế với độ chính xác ít nhất 90%.
    - Nhận diện chính xác: Đạt được mức độ phân loại chính xác cho từng loại hoạt động.
  + Phương pháp Đánh giá:
    - Chỉ số Hiệu suất: Sử dụng các chỉ số như ma trận nhầm lẫn, độ chính xác, độ nhạy, và độ đặc hiệu để đánh giá hiệu suất của mô hình.
    - Phản hồi Người dùng: Thu thập phản hồi từ các tình nguyện viên về độ chính xác và tính hữu ích của hệ thống.

## Kế hoạch Thực hiện:

* + Quy trình Thực hiện:
    - Chuẩn bị: Đọc hiểu dữ liệu từ bộ dữ liệu đã có
    - Thu thập dữ liệu: Tổng hợp và xử lý dữ liệu từ bộ dữ liệu.
    - Phân tích dữ liệu: Áp dụng các phương pháp phân tích dữ liệu để nhận diện và phân loại các hoạt động từ dữ liệu thu thập được.
    - Đánh giá kết quả: So sánh kết quả mô phỏng với mục tiêu dự kiến và điều chỉnh hệ thống nếu cần thiết.

# 2.2. Lựa chọn và mô tả bộ dữ liệu

## Môi trường Mô phỏng:

* + Thiết bị Đeo: Sử dụng smartphone (Samsung Galaxy S II) gắn trên thắt lưng của người dùng, với cảm biến gia tốc và con quay hồi chuyển được kích hoạt để thu thập dữ liệu.
  + Thử nghiệm: Thực hiện mô phỏng trong môi trường phòng thí nghiệm với 10 tình nguyện viên, mỗi người thực hiện 6 hoạt động khác nhau: Đi bộ, Đi lên cầu thang, Đi xuống cầu thang, Ngồi, Đứng, và Nằm.

## Tình huống và Tình trạng:

* + Tình huống: Các tình nguyện viên sẽ thực hiện các hoạt động cụ thể trong một khoảng thời gian xác định, ví dụ như mỗi hoạt động kéo dài 30 giây.
  + Tình trạng: Môi trường phòng thí nghiệm sẽ được kiểm soát để giảm thiểu sự can thiệp từ các yếu tố bên ngoài như tiếng ồn và sự di chuyển không mong muốn.

## Các Thông số và Điều kiện Thử nghiệm:

* + Dữ liệu Đầu vào:
    - Dữ liệu cảm biến gia tốc (trục X, Y, Z): Thu thập dữ liệu từ cảm biến gia tốc với tần số lấy mẫu 50Hz.
    - Dữ liệu con quay hồi chuyển (trục X, Y, Z): Thu thập dữ liệu từ cảm biến con quay hồi chuyển với tần số lấy mẫu 50Hz.
  + Tham số và Cài đặt:
    - Cửa sổ Thời gian: Sử dụng cửa sổ trượt rộng 2.56 giây với 50% trùng lặp (128 đọc mỗi cửa sổ) để phân tích dữ liệu.
    - Tiền xử lý Dữ liệu: Áp dụng bộ lọc Butterworth với tần số cắt 0.3 Hz để tách thành phần gia tốc cơ thể và trọng lực.

## Mục tiêu của Mô phỏng:

* + Đánh giá Hiệu suất:
    - Nhận diện chính xác: Xác định khả năng của hệ thống trong việc phân loại chính xác các hoạt động và tư thế từ dữ liệu cảm biến.
    - Đánh giá hiệu suất mô hình: Đánh giá các chỉ số hiệu suất như độ chính xác, độ nhạy và độ đặc hiệu của mô hình dự đoán.
  + Xác định Vấn đề: Phân tích lỗi:
    - Phát hiện các lỗi hoặc sai sót trong quá trình nhận diện hoạt động và tư thế.
    - Khắc phục vấn đề: Xác định các vấn đề trong quy trình tiền xử lý dữ liệu hoặc cấu hình mô hình dự đoán.

## Kết quả Dự kiến và Đánh giá:

* + Kết quả Mong đợi:
    - Độ chính xác cao: Dự đoán các hoạt động và tư thế với độ chính xác ít nhất 90%.
    - Nhận diện chính xác: Đạt được mức độ phân loại chính xác cho từng loại hoạt động.
  + Phương pháp Đánh giá:
    - Chỉ số Hiệu suất: Sử dụng các chỉ số như ma trận nhầm lẫn, độ chính xác, độ nhạy, và độ đặc hiệu để đánh giá hiệu suất của mô hình.
    - Phản hồi Người dùng: Thu thập phản hồi từ các tình nguyện viên về độ chính xác và tính hữu ích của hệ thống.

## Kế hoạch Thực hiện:

* + Quy trình Thực hiện:
    - Chuẩn bị: Đọc hiểu dữ liệu từ bộ dữ liệu đã có.
    - Thu thập dữ liệu: Tổng hợp và xử lý dữ liệu từ bộ dữ liệu.
    - Phân tích dữ liệu: Áp dụng các phương pháp phân tích dữ liệu để nhận diện và phân loại các hoạt động từ dữ liệu thu thập được.
    - Đánh giá kết quả: So sánh kết quả mô phỏng với mục tiêu dự kiến và điều chỉnh hệ thống nếu cần thiết.

# 2.3. Đường ống nhập dữ liệu

* Nguồn dữ liệu được thu thập từ các thiết bị đeo sức khỏe và lưu trữ dưới dạng file CSV. Để nhập dữ liệu này vào hệ thống Hadoop, thực hiện các bước sau:
  + Tải dữ liệu : Dữ liệu ở dạng file CSV.
  + Tạo các thư mục trên HDFS để lưu trữ dữ liệu.
  + Tải dữ liệu lên HDFS.
  + Kiểm tra dữ liệu trên HDFS.

# 2.4. Xử lý chuyển đổi dữ liệu

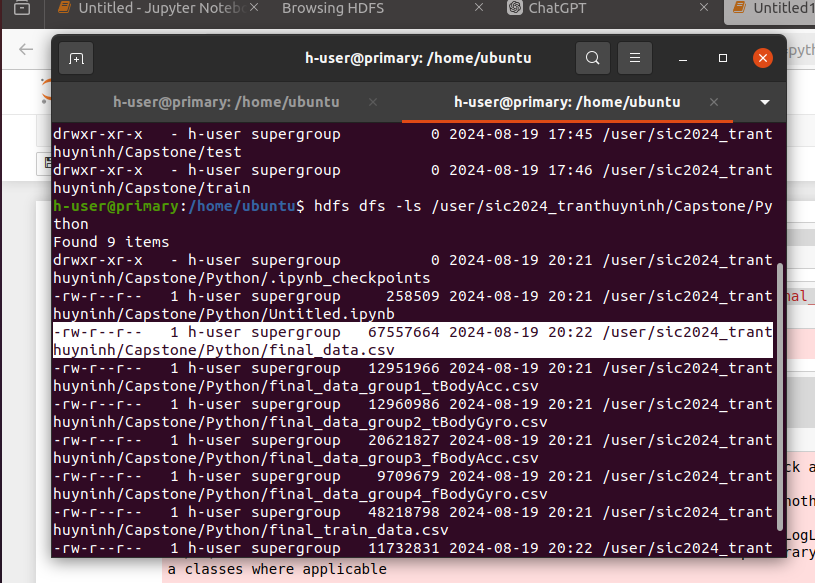
* Bộ Dataset đã được xử lý :
  + Làm sạch dữ liệu (Data Cleaning).
  + Biến đổi dữ liệu (Data Transformation).
  + Tích hợp dữ liệu (Data Integration).
  + Định dạng lại dữ liệu (Data Formatting).
  + Chuyển đổi kiểu dữ liệu (Data Type Conversion).
* Dữ liệu được lấy từ một nguồn duy nhất nên bước làm giàu dữ liệu sẽ không thực thi.

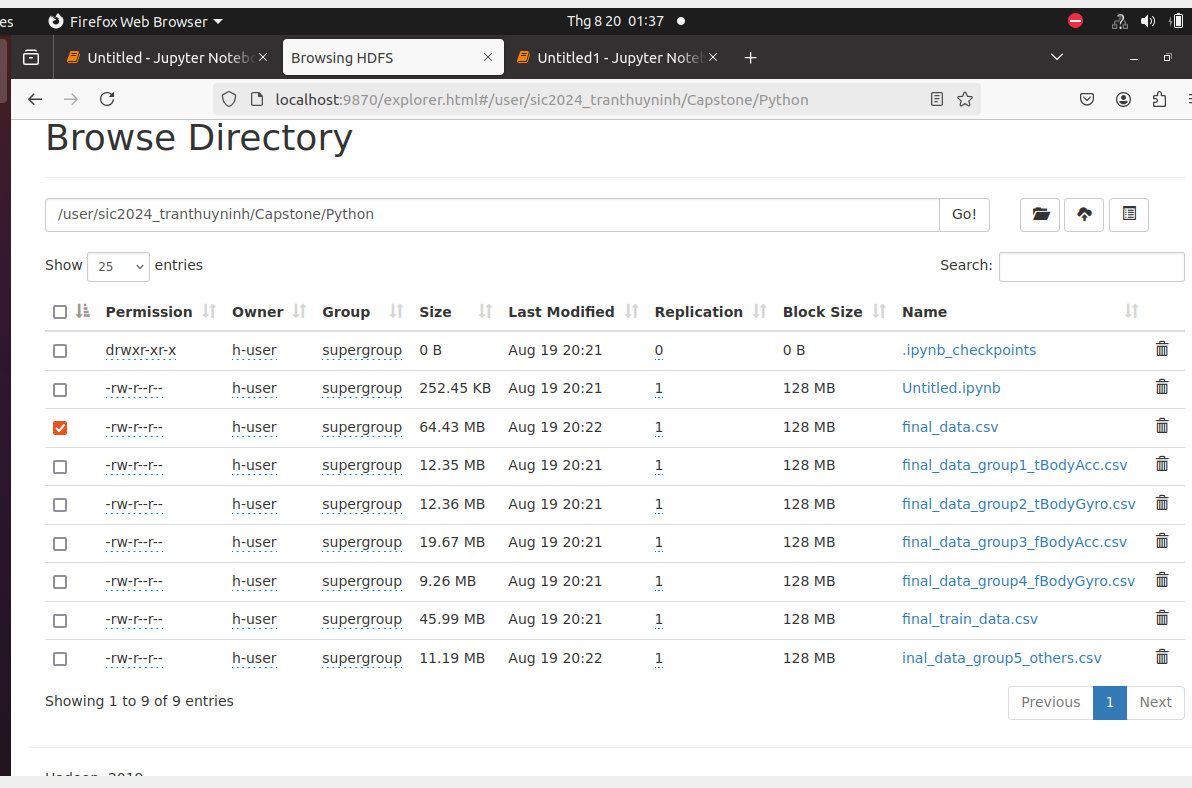
# 2.5. Truy vấn và hiểu biết dữ liệu

* Chuẩn bị dữ liệu.
* Khám phá dữ liệu và trực quan hóa kết quả phân tích:
  + Sự khác biệt về giá trị trung bình và độ lệch chuẩn của các biến giữa các hoạt động khác nhau
  + Phân tích sự tương quan giữa các biến gia tốc, tốc độ góc và các góc độ
  + Các đặc trưng nào trong dữ liệu có khả năng dự đoán tốt nhất hoạt động của một đối tượng?
  + Sự phân phối của các giá trị biến trong các hoạt động khác nhau trông như thế nào?
  + Số lần các hoạt động mà các tình nguyện viên thực hiện?
  + Đặc trưng nào có giá trị trung bình cao nhất?
  + Đối tượng tham gia hoạt động nhiều nhất

1. Kết quả

# 3.1. Mã nhập dữ liệu và code



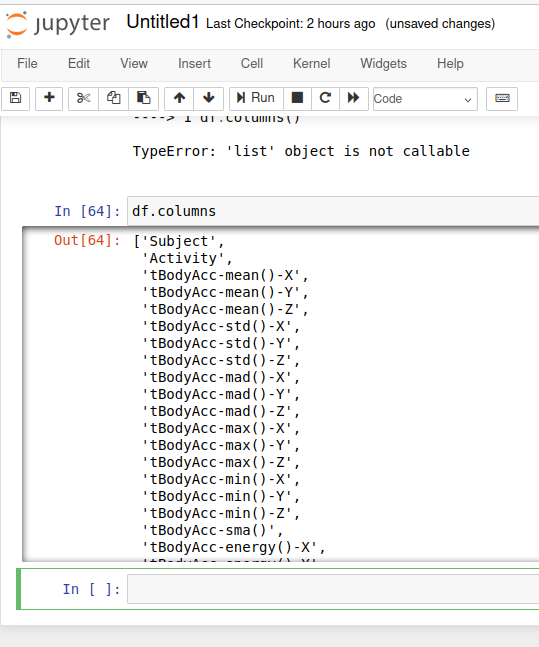


* Đọc dữ liệu từ HDFS:

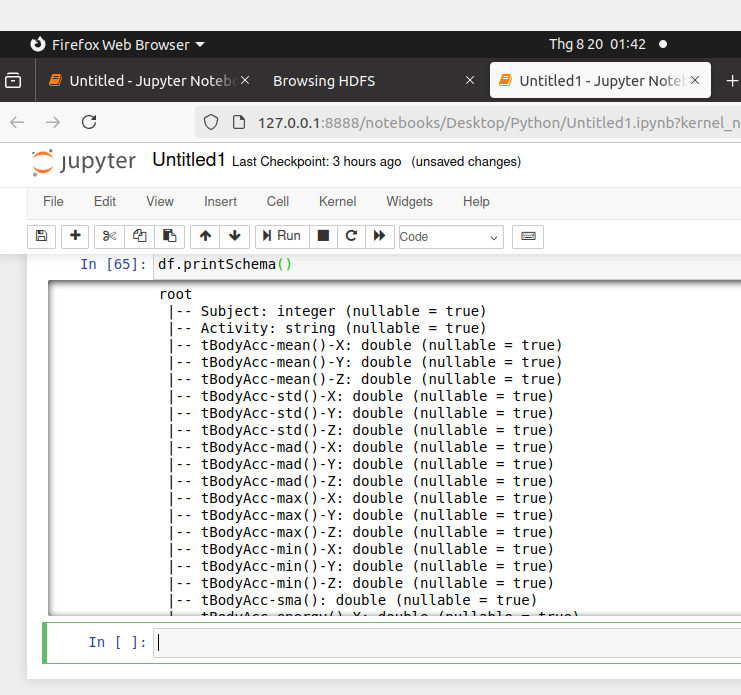
df = spark.read.csv('hdfs:///user/sic2024\_tranthuyninh/Capstone/Python/final\_data.csv', inferSchema=True, header=True)

# 3.2. Description and Sample of Transformed Datasets

* Các cột hiện tại trong df

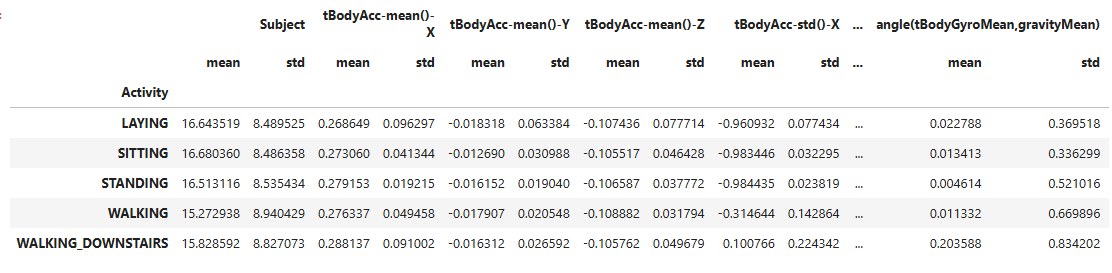


* Trường và thuộc tính các cột

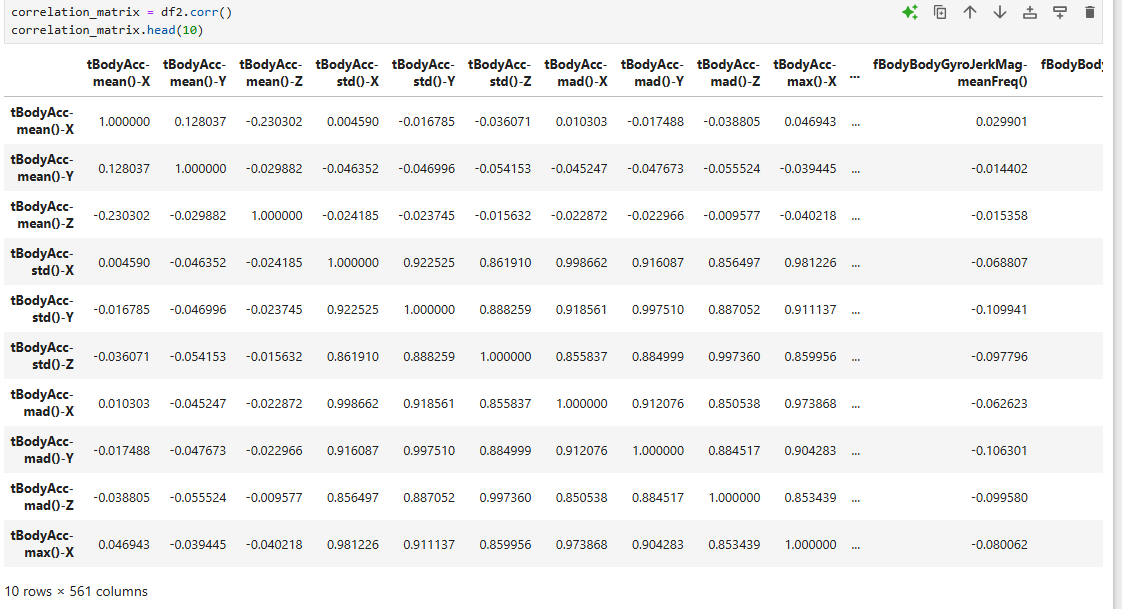


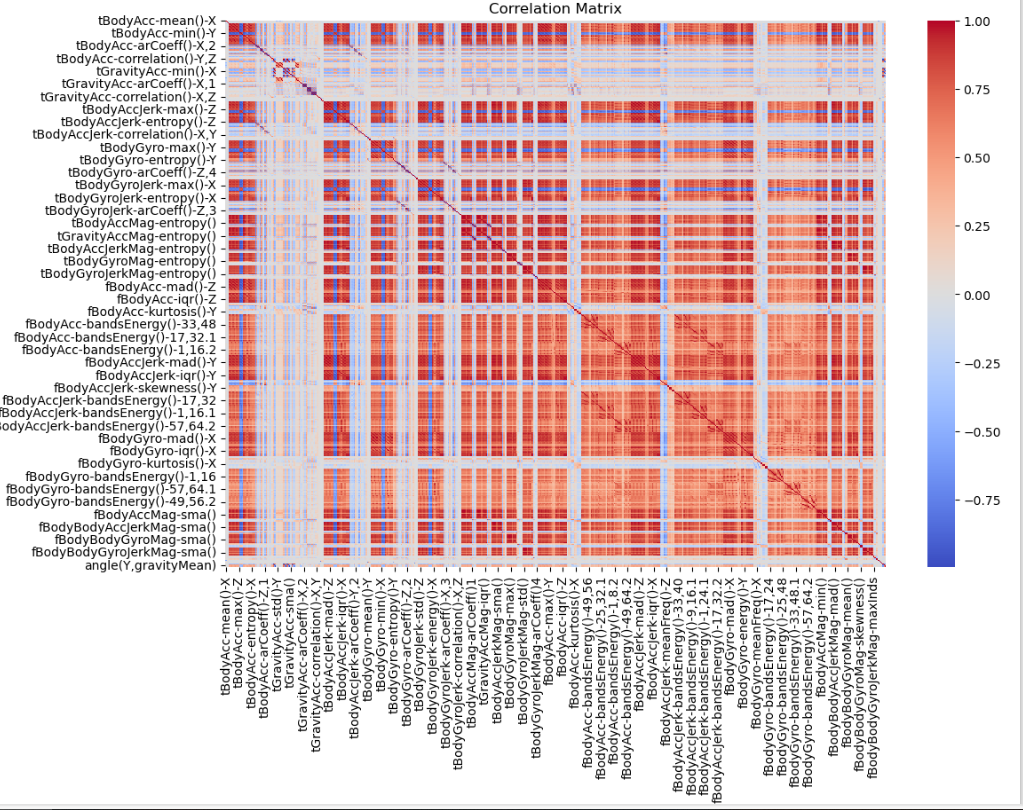
# 3.3. Data Visualization of Query Results

* Sự khác biệt về giá trị trung bình và độ lệch chuẩn của các biến giữa các hoạt động khác nhau:

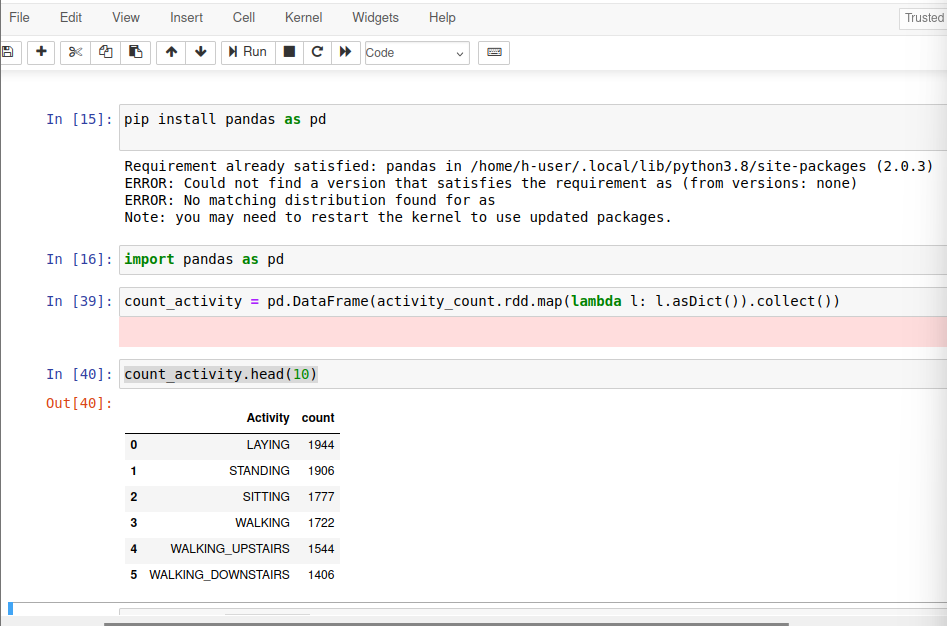


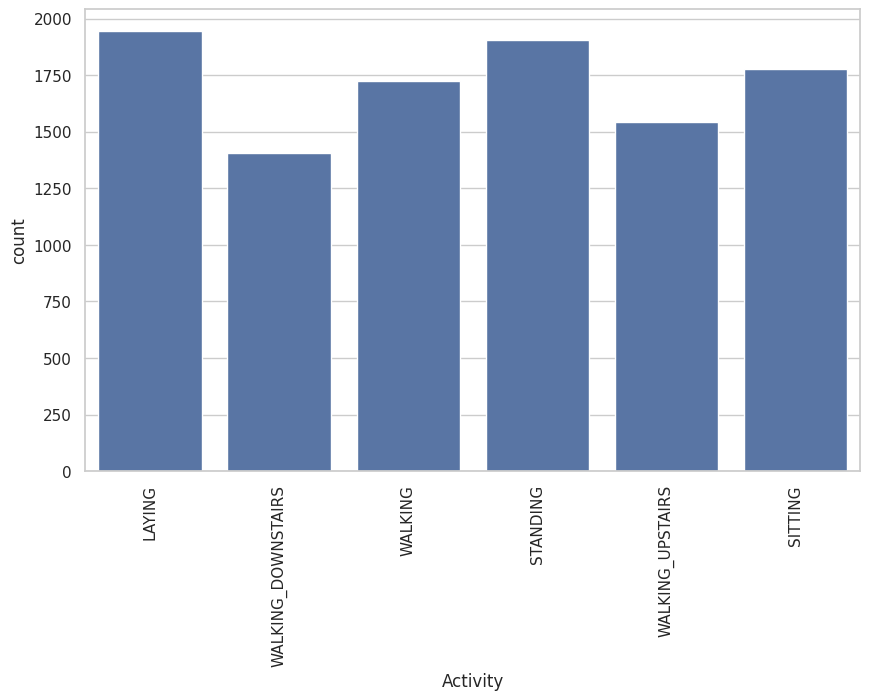
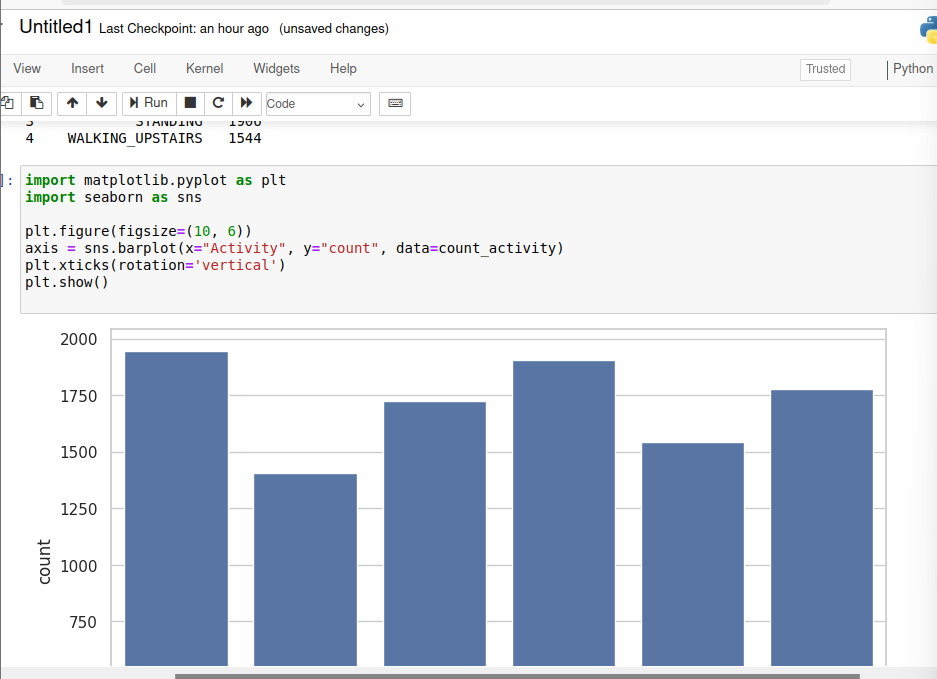
* Phân tích sự tương quan giữa các biến gia tốc, tốc độ góc và các góc độ



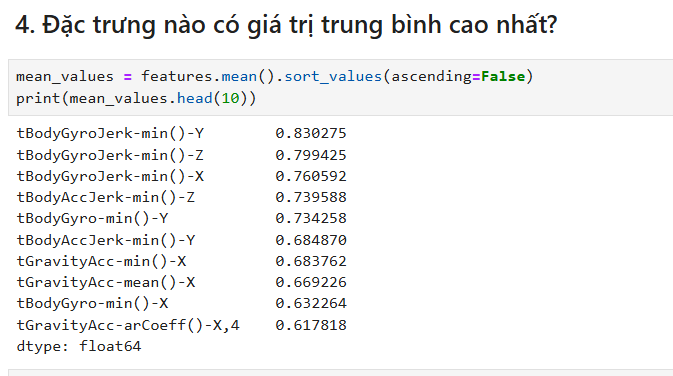


* Thống kê số lượng các hoạt động mà các tình nguyện viên đã thực hiện

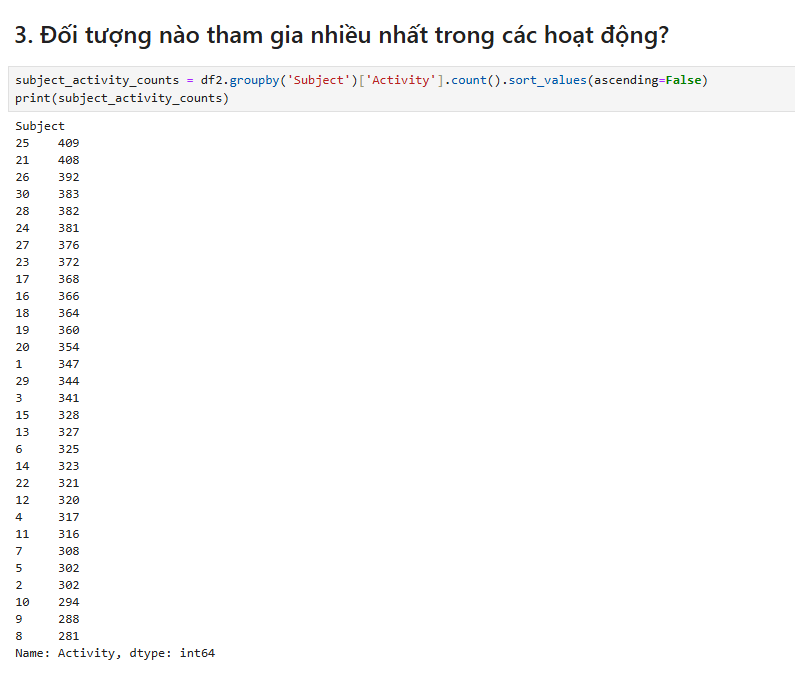




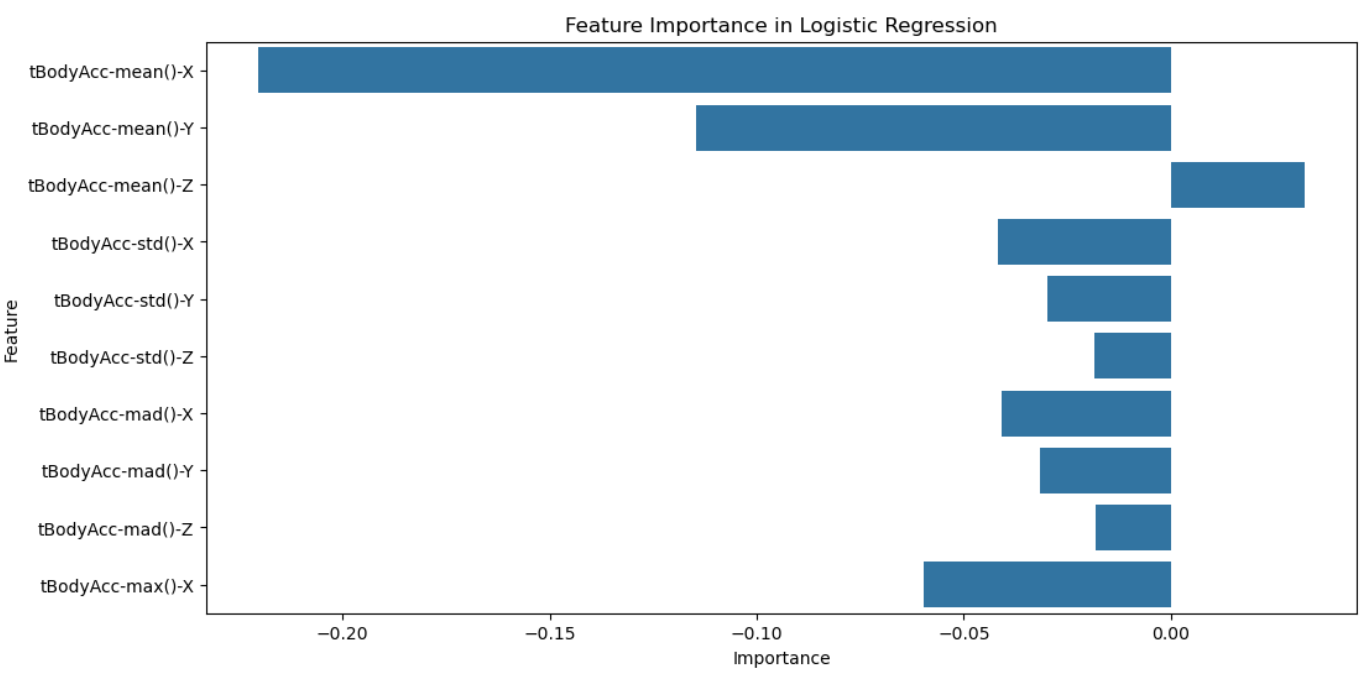
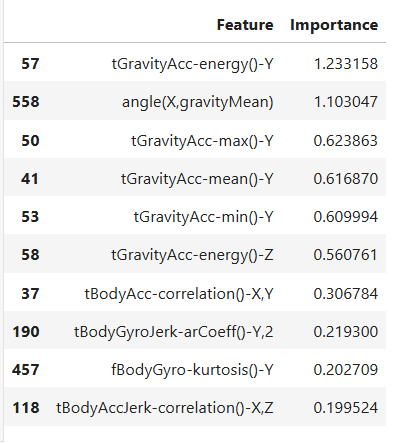
* Thống kê đặc trưng có giá trị trung bình cao nhất



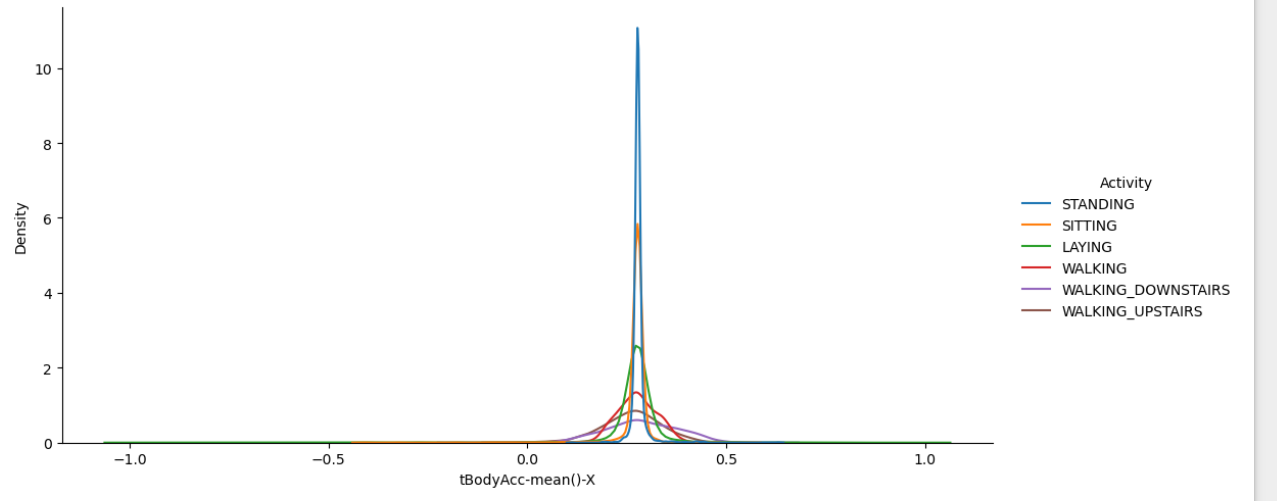
* Thống kê đối tượng tham gia hoạt động nhiều nhất

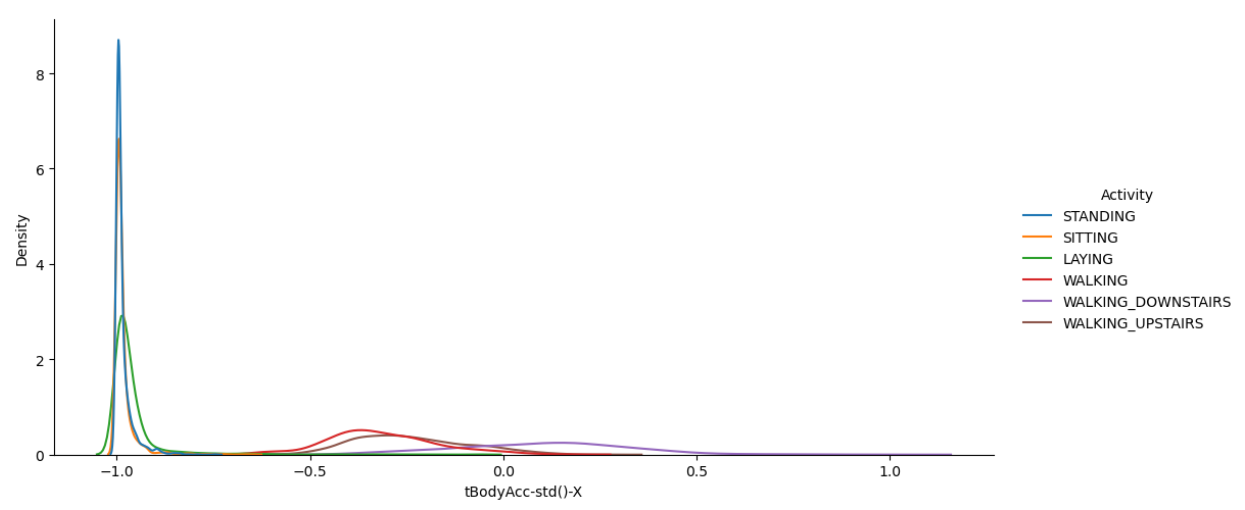


* Các đặc trưng nào trong dữ liệu có khả năng dự đoán tốt nhất hoạt động của một đối tượng?



* Sự phân phối của các giá trị biến trong các hoạt động khác nhau trông như thế nào?





# 3.5. Xây dựng mô hình

* LogisticRegression

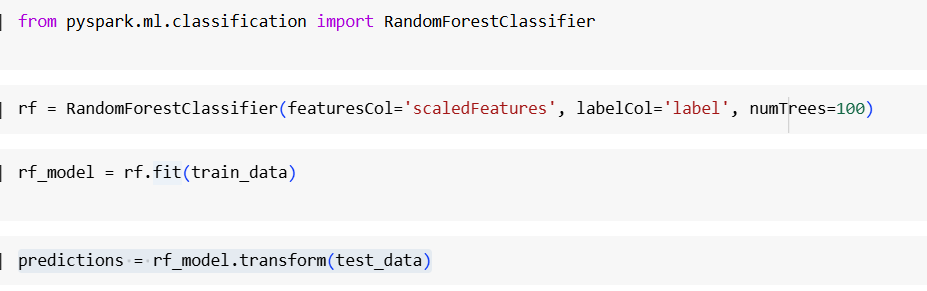
from pyspark.ml.classification import LogisticRegression

log\_reg = LogisticRegression(featuresCol='scaledFeatures', labelCol='label', maxIter=100, family="multinomial")

log\_reg\_model = log\_reg.fit(train\_data)

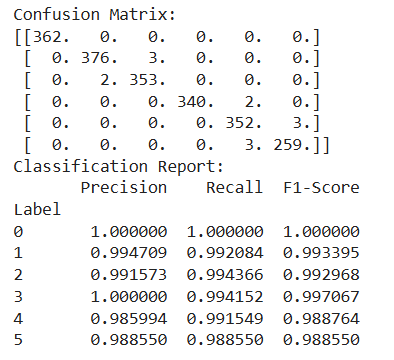
predictions = log\_reg\_model.transform(test\_data)

* RandomForest



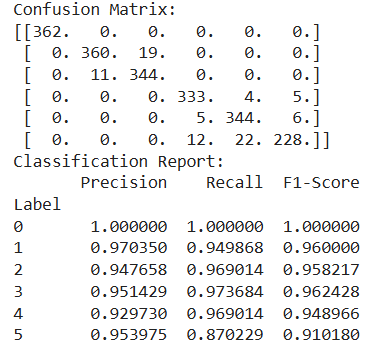
# 3.6. Đánh giá mô hình.

* LogisticRegression





* RandomForest







4. Projected Impact

4.1. Accomplishments and Benefits

* Về cơ bản nhóm đã hoàn thành những mục tiêu đã đề ra :
* Hiểu và phân tích được dữ liệu chăm sóc sức khỏe từ các thiết bị đeo từ bộ dữ liệu cho sẵn Human Activity Recognition Using Smartphones Data Set. Phân biệt và lọc được các dữ liệu đầu vào (chỉ số gia tốc tuyến tính 03 trục, vận tốc góc 03 trục) và dữ liệu đầu ra (dự đoán về tư thế, hoạt động của con người).
* Đã đẩy và lưu trữ dữ liệu dưới dạng file csv lên Hadoop thành công.
* Trong quá trình xây dựng các mô hình học máy như mô hình Hồi quy logistic và Random forest đã áp dụng được Spark MLlib (thư viện của Spark) vào việc xử lý dữ liệu lớn và dữ liệu phân tán.
* Qua việc sử dụng 02 mô hình học máy nêu trên, nhóm đã đưa đánh giá thông qua ma trận nhầm lẫn (Confusing Matrix) với tỷ lệ dự đoán đúng như đã trình bày tại mục 3.6.
* Lợi ích dự án mang lại:
* Dự án "Phân tích dữ liệu chăm sóc sức khỏe từ các thiết bị đeo" mang lại nhiều lợi ích tiềm năng cho người dùng và cộng đồng như cải thiện, dự đoán sớm các vấn đề sức khỏe, cung cấp dữ liệu chính xác cho các ứng dụng chăm sóc sức khỏe.
* Giúp sinh viên nắm được các kỹ năng về việc phân tích và xử lý các dữ liệu lớn, hiểu và nắm được kỹ năng sử dụng hệ thống lưu trữ dữ liệu lớn Hadoop, hiểu biết thêm về các thuật toán, kỹ năng làm việc nhóm trong quá trình làm dự án.

4.2. Future Improvements

* Hạn chế hiện tại : Bộ dữ liệu đã được cung cấp sẵn nên nhóm chưa thực hiện được các bước xử lý như lọc, làm sạch dữ liệu. Dự án về dữ liệu chăm sóc sức khỏe cần có nhiều nguồn dữ liệu nhưng nhóm vẫn chưa mở rộng dữ liệu được.
* Đề xuất cải thiện : Thu thập thêm dữ liệu thông qua các bộ dataset với nguồn đáng tin cậy, tạo các cuộc thử nghiệm thực tế, khảo sát, phỏng vấn…

5. Team Member Review and Comment 

|  |
| --- |
| <ATTACH A TEAM PICTURE HERE> |

|  |  |
| --- | --- |
| NAME | REVIEW and COMMENT |
| Ngọ Quốc Minh |  |
| Lê Nghĩa Dũng |  |
| Trần Thùy Ninh |  |
| Hoàng Thị Kim Thoa |  |
| Trần Thị Yến Nhi |  |
| Nguyễn Thu Trang |  |
| Nguyễn Thị Tú Lan |  |

6. Instructor Review and Comment

|  |  |  |
| --- | --- | --- |
| CATEGORY | SCORE | REVIEW and COMMENT |
| IDEA | \_\_/10 |  |
| APPLICATION | \_\_/30 |  |
| RESULT | \_\_/30 |  |
| PROJECT MANAGEMENT | \_\_/10 |  |
| PRESENTATION & REPORT | \_\_/20 |  |
| TOTAL | \_\_/100 |  |