Big Data Course

Capstone Project   
Action Plan

For students (instructor’s 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.

|  |  |
| --- | --- |
| Course | Big Data Course |
| Team Name | Nhóm BigData Capstone 01 |
| Team Leader/  Members | Lê Nghĩa Dũng  Nguyễn Thị Tú Lan  Trần Thùy Ninh  Ngọ Quốc Minh  Hoàng Thị Kim Thoa  Trần Thị Yến Nhi  Nguyễn Thu Trang |
| Project Title | Chủ đề số 11: Phân tích dữ liệu chăm sóc sức khỏe từ các thiết bị đeo |
| Goal |  |
| * Thu thập, xử lý, và phân tích khối lượng lớn dữ liệu sức khỏe từ các thiết bị đeo như đồng hồ thông minh và vòng tay theo dõi sức khỏe. Điều này bao gồm dữ liệu về hoạt động thể chất, giấc ngủ, nhịp tim, và các chỉ số khác từ hàng nghìn người dùng. * Sử dụng các công cụ và kỹ thuật Big Data để xử lý dữ liệu lớn, bao gồm việc làm sạch, phân loại, và trích xuất thông tin hữu ích. * Áp dụng các thuật toán học máy hoặc phân tích dữ liệu tiên tiến để khám phá các mẫu hình sức khỏe, xu hướng hành vi, hoặc dự đoán nguy cơ sức khỏe dựa trên dữ liệu thu thập được. * Đưa ra các khuyến nghị hoặc dự đoán về sức khỏe cá nhân hóa, từ đó cải thiện sức khỏe và lối sống cho người dùng. | |
| Abstract |  |
| Dự án này nhằm mục đích phân tích dữ liệu sức khỏe thu thập từ các thiết bị đeo thông minh như đồng hồ thông minh và vòng đeo tay. Dữ liệu bao gồm các chỉ số gia tốc tuyến tính 03 trục, vận tốc góc 03 trục. Dữ liệu được lưu trữ trong hệ thống Hadoop và phân tích bằng Spark MLlib. Các bước thực hiện bao gồm phân tích xu hướng vận động, xây dựng và đánh giá mô hình dự đoán tư thế, hoạt động của con người . Dựa trên kết quả phân tích, dự án cũng đề xuất các biện pháp cải thiện các điểm hạn chế. | |
| Method |  |
| **Lưu trữ Dữ liệu trên Hadoop:** Sử dụng Hadoop để lưu trữ và quản lý dữ liệu thu thập từ các thiết bị đeo.  **Phân Tích Xu Hướng Tư Thế, HOẠT ĐỘNG:** Xác định các xu hướng hoạt động từ dữ liệu thu được từ các thiết bị đo như gia tốc kế, con quay hồi chuyển.  **Dự Đoán Chỉ Số Sức Khỏe:** Xây dựng mô hình dự đoán cho các chỉ số chăm sóc sức khỏe quan trọng dựa trên dữ liệu lịch sử. | |

|  |  |
| --- | --- |
| Data |  |
| Thu thập dữ liệu từ bộ dữ liệu: Human Activity Recognition Using Smartphones Data Set. | |
| Expected  Outcome |  |
| * 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. | |
| Role by  Member |  |
| Lê Nghĩa Dũng  Nguyễn Thị Tú Lan  Trần Thùy Ninh  Ngọ Quốc Minh  Hoàng Thị Kim Thoa  Trần Thị Yến Nhi  Nguyễn Thu Trang | |

|  |  |
| --- | --- |
| Schedule  Summary |  |
| 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 | |
| Comment &  Assessment |  |
| * 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.      * 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… | |