

MỤC LỤC

TÓM LƯỢC	1
ABSTRACT	2
PHẦN 1 - GIỚI THIỆU	3
1. ĐẶT VẤN ĐỀ	3
2. MỤC TIÊU ĐỀ TÀI	3
3. ĐỐI TƯỢNG VÀ PHẠM VI NGHIÊN CỨU	3
4. PHƯƠNG PHÁP NGHIÊN CỨU	3
5. NỘI DUNG NGHIÊN CỨU	4
6. BỐ CỤC BÀI BÁO CÁO	4
PHẦN 2 - NỘI DUNG	5
I. ĐẶC TẢ YÊU CẦU	5
1. Mục tiêu	5
2. Phạm vi sản phẩm	5
4. Các chức năng của sản phẩm	5
5. Đặc điểm người sử dụng	5
6. Môi trường vận hành	5
7. Các ràng buộc về thực thi và thiết kế	6
II. THIẾT KẾ GIẢI PHÁP	6
1. Đặc điểm dữ liệu	6
2. Thiết kế giải thuật	6
3. Thuật toán KMeans	7
III. CÀI ĐẶT GIẢI PHÁP	8
1. Cài đặt Server	9
1.1. Xử lý file tải lên	9
1.2. Xem trước file	9
1.3. Thực hiện gom nhóm	9
2. Cài đặt giao diện người dùng với ReactJS	11
2.1. Giao diện chọn file	11
2.2. Giao diện chọn trường dữ liệu	12
2.3. Giao diện xem trước dữ liệu đã chọn	13
2.4. Giao diện kết quả gom nhóm	14
PHẦN 3 - KẾT LUẬN	15
1. KẾT QUẢ ĐẠT ĐƯỢC	15
2. HƯỚNG PHÁT TRIỂN	15
TÀI LIỆU THAM KHẢO	16

TÓM LƯỢC

Dự án này nhằm mục đích nghiên cứu để phát triển một ứng dụng web có thể giúp các doanh nghiệp phân tích dữ liệu về sản phẩm của họ một cách tự động. Ứng dụng web này cho phép người dùng tải lên file dữ liệu bằng Excel và tùy chọn các trường dữ liệu cần phân tích cũng như tùy chọn số nhóm sản phẩm muốn phân tách. Đồng thời ứng dụng cũng cho phép người dùng xem trước dữ liệu và tải xuống kết quả phân tích. Ứng dụng sử dụng giải thuật KMeans để phân tích dữ liệu.

Từ khóa: Cluster, Python, NodeJS, Shell Script, Child Process, NumPy, Pandas, sklearn, KMeans.

ABSTRACT

This project is aimed at researching and developing a web application that can help businesses analyze data about their products automatically. This web application allows users to upload data in Excel files and optionally select the data fields to be analyzed as well as the number of product groups to split. The application also allows users to preview data and download analysis results. This application use the KMean algorithm to analyze data.

Keywords: Cluster, Python, NodeJS, Shell Script, Child Process, NumPy, Pandas, sklearn, KMeans.

PHẦN 1 - GIỚI THIỆU

1. ĐẶT VẤN ĐỀ

Trong thời đại công nghệ hiện nay, việc ứng dụng trí tuệ nhân tạo hay các thuật toán máy học để tối ưu nâng suất lao động ngày càng phổ biến. Trong khi đó các doanh nghiệp kinh doanh lâu năm có trong tay một lượng lớn dữ liệu về sản phẩm và khách hàng của họ. Để phân tích lượng dữ liệu lớn như vậy đòi hỏi người thực hiện phải mất rất nhiều thời gian nếu làm theo phương pháp truyền thống. Nắm bắt tình hình đó, đề tài này được nghiên cứu để tạo ra công cụ giúp các doanh nghiệp phân tích dữ liệu của họ một cách tự động nhờ vào các thuật toán máy học.

2. MỤC TIÊU ĐỀ TÀI

Xây dựng ứng dụng giúp các doanh nghiệp tìm hiểu về tiềm năng phát triển của các sản phẩm bằng cách phân tích nguồn dữ liệu có được từ quá trình hoạt động lâu dài trên thị trường, với các chức năng cụ thể như:

- Nhận dữ liệu từ file Excel.
- Tùy chỉnh dữ liệu cần phân tích.
- Tải xuống kết quả phân tích.

3. ĐỐI TƯỢNG VÀ PHẠM VI NGHIÊN CỨU

- Ngôn ngữ lập trình Shell script, Python, Javascript.
- Thư viện máy học scikit-learn, giải thuật clustering.
- NodeJS, ReactJS

4. PHƯƠNG PHÁP NGHIÊN CỨU

- Tìm hiểu, phân tích dữ liệu nguồn
- Thử nghiệm các giải thuật clustering để chọn thuật toán phù hợp.
- Cài đặt giải pháp.
- Kiểm thử và cải tiến.

5. NỘI DUNG NGHIÊN CỨU

- Tìm hiểu dữ liệu, phân tích và thiết kế giải pháp.
- Cài đặt giải pháp bằng Shell script, Python và NodeJS.
- Cài đặt giao diện người dùng với ReactJS.

6. BỐ CỤC BÀI BÁO CÁO

Bài báo cáo gồm 3 phần chính:

- Phần 1: Giới thiệu tổng quan về đề tài nghiên cứu
- Phần 2: Nội dung nghiên cứu, bao gồm mô tả yêu cầu, giải pháp và phương pháp cài đặt.
- Phần 3: Kết luận về kết quả đạt được, hạn chế và hướng phát triển.

PHẦN 2 - NỘI DUNG

I. ĐẶC TẢ YÊU CẦU

1. Mục tiêu

Xây dựng ứng dụng web dùng thuật toán gom cụm (clustering) để phân loại sản phẩm thành từng nhóm theo các thuộc tính nhất định do người dùng chọn.

2. Phạm vi sản phẩm

Dành cho các doanh nghiệp có lượng dữ liệu lớn về sản phẩm và khách hàng của mình, muốn phân tích để lập chiến lược tối ưu lợi nhuận.

4. Các chức năng của sản phẩm

- Tải lên tệp tin Excel với dung lượng lớn.
- Tùy chọn trường dữ liệu cần phân tích.
- Xem trước dữ liệu trước khi phân tích.
- Tùy chọn độ chính xác và số nhóm cần phân loại.

5. Đặc điểm người sử dụng

Người dùng là những người có thẩm quyền truy cập vào dữ liệu của doanh nghiệp, có hiểu biết về đối tượng sẽ phân tích, có khả năng cụ thể hoá kết quả phân tích.

6. Môi trường vận hành

Máy chủ:

- Ubuntu 18.04.3 LTS, python3, shell, xlsx2csv, node

Máy khách:

- Bất kỳ thiết bị nào có thể truy cập Internet và hỗ trợ Javascript.

7. Các ràng buộc về thực thi và thiết kế

- Thực thi: Hoạt động tốt khi có mạng.
- Thiết kế:
 - Ngôn ngữ lập trình: Javascript, Python
 - Nền tảng lập trình: ReactJS, NodeJS, Python.
 - Thiết kế giao diện trực quan, thân thiện, rõ ràng, dễ sử dụng.

II. THIẾT KẾ GIẢI PHÁP

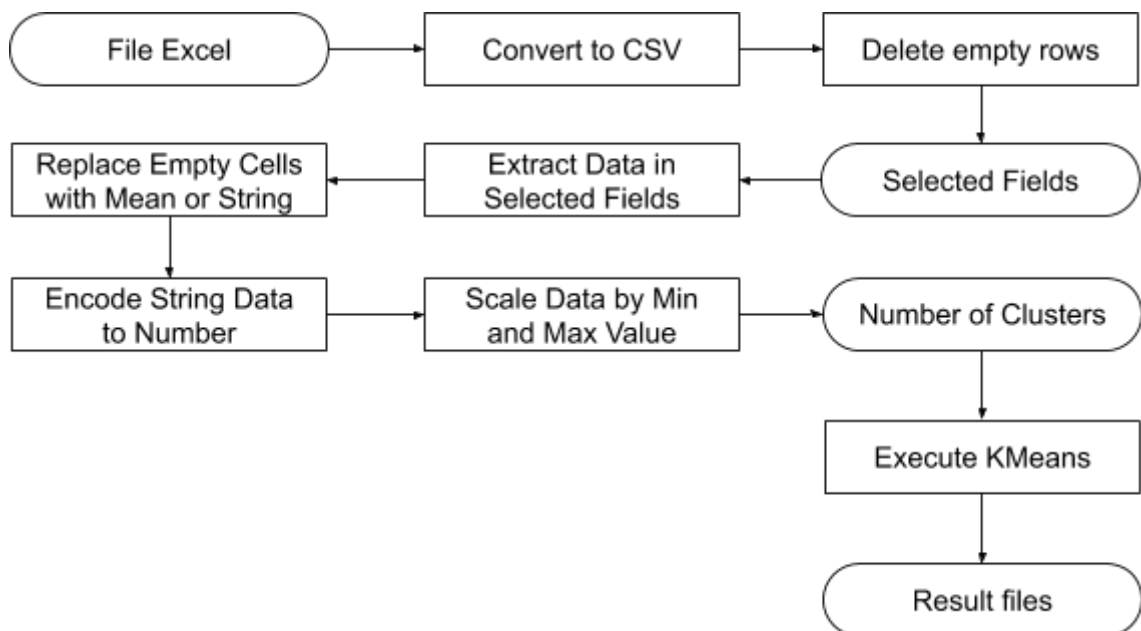
1. Đặc điểm dữ liệu

- Dữ liệu có số lượng lớn, được lưu trữ bằng Excel.
- Dữ liệu có nhiều cột, có cột chứa dữ liệu số, có cột chứa dữ liệu chuỗi.
- Dữ liệu nguồn chưa qua xử lý, có nhiều dòng, ô dữ liệu trống cuối tập tin hoặc xen kẽ.

2. Thiết kế giải thuật

Với đặc điểm dữ liệu phân tích được, giải thuật thiết kế cần phải xử lý và chuẩn hóa dữ liệu trước khi tiến hành gom nhóm để đảm bảo độ chính xác, đồng thời phải xử lý được cả dữ liệu số và chữ.

Sơ đồ giải thuật:



Giải thuật nhận File Excel từ người dùng sau đó chuyển thành file CSV để tối ưu các phần xử lý phía sau. Tiếp đến là xóa các dòng trống không có dữ liệu.

Sau đó hệ thống nhận input là các trường mà người dùng đã chọn để phân tích, dựa vào đó giải thuật trích xuất những dữ liệu được chọn để tiếp tục xử lý. Với dữ liệu được chọn, giải thuật thay thế những ô trống bằng giá trị trung bình (Mean) nếu ô trống nằm trong cột dữ liệu số, hoặc thay bằng một chuỗi ký tự bất kỳ (String) nếu ô trống nằm trong cột dữ liệu chuỗi.

Do giải thuật KMeans chỉ làm việc tốt với dữ liệu số, nên bước tiếp theo, các cột dữ liệu chuỗi sẽ được mã hóa thành số và để giá trị giữa các cột dữ liệu không chênh lệch quá nhiều, ta tiến hành tính toán lại dữ liệu ở tất cả các cột theo tỉ lệ Min Max.

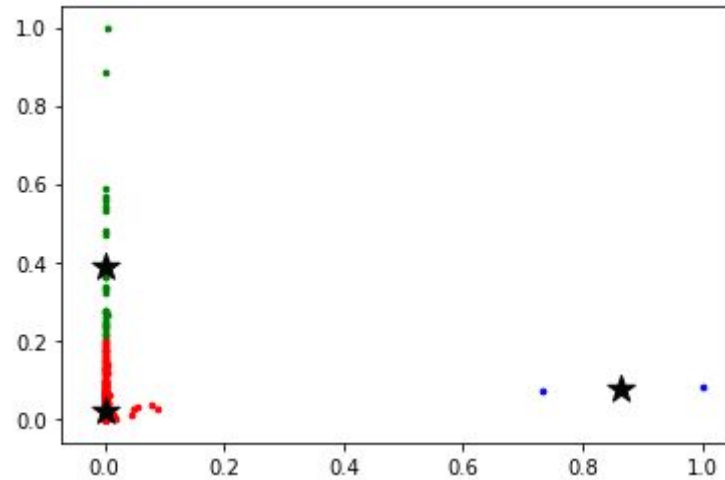
Đến đây, dữ liệu đã sẵn sàng để phân tích. Hệ thống nhận input là số nhóm cần gom từ người dùng và tiến hành chạy thuật toán KMeans.

Sau khi thuật toán hoàn thành, hệ thống xuất ra các file kết quả tương ứng với số nhóm người dùng đã chọn.

3. Thuật toán KMeans

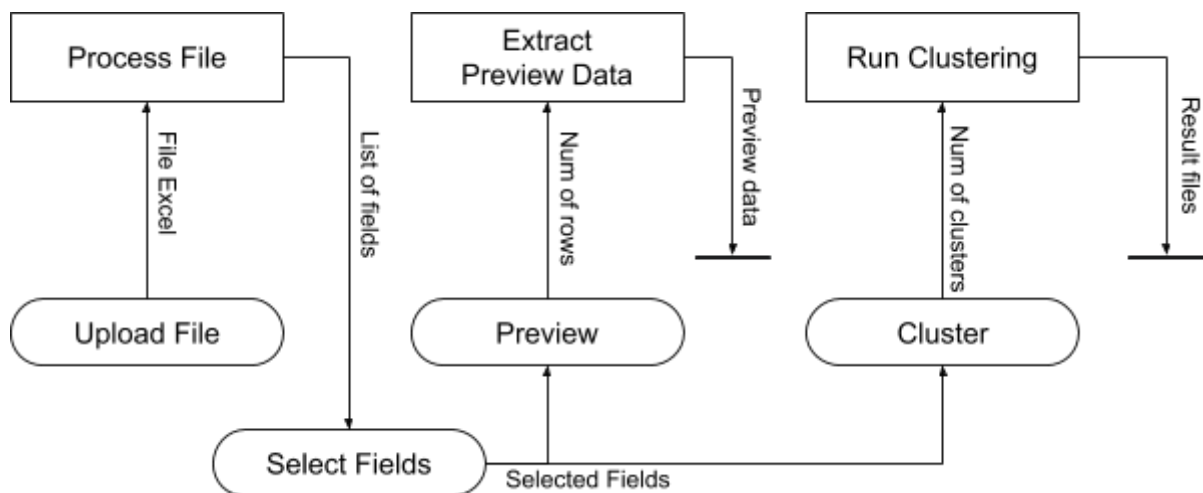
KMeans là một thuật toán gom cụm (cluster) phổ biến trong máy học.

Trong KMeans clustering, chúng ta không biết nhãn (label) của từng điểm dữ liệu. Mục đích là làm thế nào để phân dữ liệu thành các cụm (cluster) khác nhau sao cho dữ liệu trong cùng một cụm có tính chất giống nhau.



Ví dụ như biểu đồ trên với một bộ dữ liệu đơn giản gồm số lượng bán ra và tổng doanh thu của sản phẩm. Gom nhóm với số nhóm cần gom là 3 thì thuật toán KMeans sẽ gom được 3 nhóm với các tính chất khác nhau cho từng nhóm như màu xanh lá là nhóm có số lượng bán ra nhiều nhưng giá thành thấp, nhóm màu đỏ là nhóm có giá thành thấp nhưng bán ra ít và nhóm màu xanh dương là nhóm có giá thành cao và số lượng bán ra cũng ít.

III. CÀI ĐẶT GIẢI PHÁP



Sơ đồ cài đặt hệ thống

Hệ thống sẽ gồm 2 phần chính là Server chịu trách nhiệm xử lý dữ liệu và Giao diện để người dùng tương tác với hệ thống.

1. Cài đặt Server

1.1. Xử lý file tải lên

Tại bước xử lý file, hệ thống sử dụng mã Shell kèm gói mở rộng **xlsx2csv** trên server Ubuntu để xử lý nhanh file tải lên.

File tải lên của người dùng là file excel với lượng dữ liệu lớn, đòi hỏi mất nhiều thời gian để xử lý, do đó để tối ưu tốc độ xử lý, file tải lên được chuyển đổi sang định CSV (dòng 1).

Sau đó tiếp tục loại bỏ những dòng trống, không có dữ liệu (dòng 2).

Cuối cùng là trích xuất dòng đầu tiên của file dữ liệu chứa tiêu đề cột (dòng 3).

1.2. Xem trước file

Để xem trước một phần nội dung file, hệ thống sử dụng lại file CSV thu được từ phần trước, kết hợp với danh sách cột và số dòng do người dùng nhập vào để trích xuất dữ liệu.

Bằng cách sử dụng gói mở rộng **pandas** trong python hệ thống dễ dàng trích xuất được dữ liệu theo lựa chọn của người dùng và trả về dưới định dạng JSON.

1.3. Thực hiện gom nhóm

Để thực hiện gom nhóm, hệ thống sử dụng các gói mở rộng của python để thực hiện tiền xử lý file và gom nhóm. Trong đó có LabelEncoder và MinMaxScaler trong gói sklearn.preprocessing để xử lý dữ liệu trước khi tiến hành gom nhóm với KMeans.

Đầu tiên là đọc vào các dữ liệu cần thiết, bao gồm file dữ liệu nguồn, số nhóm cần gom, số lần lặp tối đa của KMeans (số này càng lớn thì độ chính xác càng cao) và các cột dữ liệu sẽ phân tích.

Tiếp theo, ta tiến hành thay thế những ô dữ liệu trống bằng giá trị trung bình của dữ liệu, sử dụng giá trị trung bình sẽ hạn chế được ảnh hưởng các ô đó đến kết quả của thuật toán. Đối với các cột dữ liệu là chuỗi ký tự thì ta thay các ô trống bằng một giá trị chuỗi bất kỳ.

KMeans là thuật toán gom nhóm hoạt động tốt với dữ liệu số, do đó, đối với các cột dữ liệu dạng chuỗi ta tiến hành mã hóa thành số để có được kết quả tốt nhất.

Sau khi toàn bộ dữ liệu đã được chuyển thành số ta tiến hành điều chỉnh giá trị của chúng theo một tỉ lệ nhất định để đảm bảo giá trị giữa các cột dữ liệu không chênh lệch quá nhiều. Ở đây ta sử dụng MinMaxScaler để điều chỉnh dữ liệu theo giá trị lớn nhất và nhỏ nhất của bộ dữ liệu.

Đến đây toàn bộ dữ liệu đã được chuẩn hóa, sẵn sàng triển khai gom nhóm. Ta sử dụng hàm KMeans trong thư viện sklearn.cluster để tiến hành gom nhóm tự động.

Sau khi gom nhóm thành công ta thu một mảng kết quả là **labels** chứa kết quả gom nhóm. Dựa vào mảng kết quả, ta trích xuất ra file kết quả để người dùng có thể tải về. Đồng thời chương trình cũng tạo ra một báo cáo tóm tắt của các nhóm kết quả để người dùng có thể xem trước khi tải về.

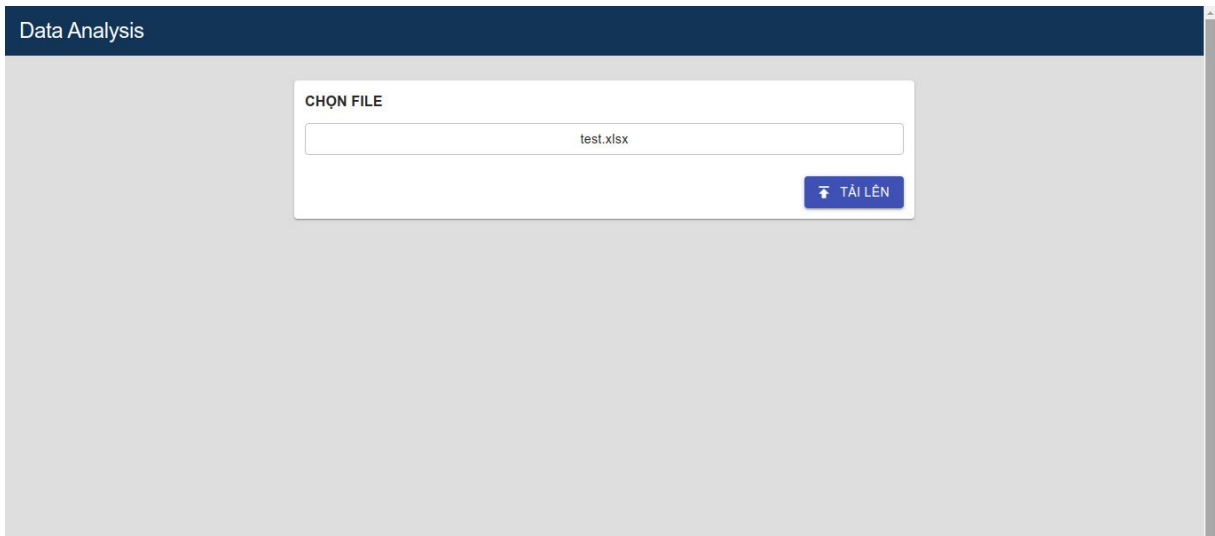
2. Cài đặt giao diện người dùng với ReactJS

2.1. Giao diện chọn file

Gồm một biểu mẫu đơn giản để người dùng chọn file và tải lên. Tuy nhiên người dùng chỉ chọn được file Excel.

Thành phần chính:

- Input [type=file]: để chọn file.
- Button [type=submit]: để tải file lên máy chủ.



The screenshot shows a web application with a dark blue header containing the text "Data Analysis". Below the header, there is a light gray background area. In the center of this area is a white rectangular box titled "CHỌN FILE". Inside this box, there is a file input field that displays "test.xlsx". To the right of the input field is a blue button with a white upload icon and the text "TẢI LÊN".

2.2. Giao diện chọn trường dữ liệu

Dùng để hiển thị tên các cột của dữ liệu mà người dùng tải lên. Đồng thời cho phép người dùng chọn những cột muốn phân tích.

Thành phần chính:

- Unordered List [item type = button]: hiển thị danh sách cột và cho phép chọn cột.
- Input [type = number]: cho phép nhập dòng bắt đầu để xem trước.
- Input [type = number]: cho phép nhập dòng kết thúc để xem trước.
- Button [type = submit]: nút chọn để tiến hành xem trước.
- Input [type = number]: cho phép nhập số nhóm cần gom.
- Input [type = number]: cho phép nhập số lần lặp lặp tối đa của thuật toán.
- Button [type = submit]: nút chọn để tiến hành gom nhóm.

CÁC THUỘC TÍNH CỦA DỮ LIỆU (56)

Month ID ✓

Sales Order Number ✓

SO Doc Type ✓

Day ID ✓

Delivery Doc Number ✓

Billing Number ✓

Model Number ✓

Customer Code ✓

Customer Ship-To Code ✓

Customer name ✓

Salesman Name ✓

Sales Office ✓

Level5 Name ✓

Store Location ✓

Actual Sales Quantity ✓

Actual Sales Amount ✓

Credit Block Quantity ✓

Credit Block Amount ✓

Under Sales Quantity ✓

Under Sales Amount ✓

Under Logistics Quantity ✓

Under Logistics Amount ✓

GIT Amount ✓

GIT Quantity Outbound ✓

GIT Quantity ✓

GIT Amount Outbound ✓

Sales Deal ✓

Sales Deal Code ✓

Actual Sales Quantity (FOC) ✓

GIT Quantity (FOC) ✓

GIT Quantity Outbound (FOC) ✓

Under Logistics Quantity (FOC) ✓

Under Sales Quantity (FOC) ✓

Other Block Base Quantity ✓

Other Block Base Quantity (FOC) ✓

Other Block Amount ✓

Volume ✓

Volume Unit ✓

SO Created Date ✓

DO Created Date ✓

Bill Date ✓

Paymnet term ✓

Sales date ✓

Invoice Number ✓

Dealer code ✓

Sales team ✓

Division ✓

Pending QTY ✓

Pending AMT ✓

Actual + Pending QTY ✓

Actual + Pending AMT ✓

Net AMT_RAC ✓

Multiple outlet ✓

AC team ✓

type ✓

Pending status ✓

Đã chọn 4 thuộc tính

CHỌN FILE

test.xlsx

TẢI LÊN

Nhập số dòng để xem trước

Bắt đầu

Kết thúc

010

XEM TRƯỚC

Tùy chọn gom nhóm

Số nhóm

Số lần lặp

3300

CHỌN LẠI

GOM NHÓM

2.3. Giao diện xem trước dữ liệu đã chọn

Phần này dùng để hiển thị một phần dữ liệu mà người dùng đã chọn để người dùng có thể xem trước khi tiến hành gom nhóm.

Thành phần chính:

- Table: hiển thị dữ liệu
- Button: để quay lại giao diện chọn trường dữ liệu.
- Input [type = text]: để tìm kiếm dữ liệu có trong bảng.
- Nhóm nút chọn phân trang dữ liệu
- Input [type = number]: cho phép nhập dòng bắt đầu để xem trước.
- Input [type = number]: cho phép nhập dòng kết thúc để xem trước.
- Button [type = submit]: nút chọn để tiến hành xem trước.
- Input [type = number]: cho phép nhập số nhóm cần gom.
- Input [type = number]: để nhập số lần lặp tối đa của thuật toán.
- Button [type = submit]: nút chọn để tiến hành gom nhóm

Data Analysis

← QUAY LẠI

Q Search

×

Level5 Name	Actual Sales Quantity	Day ID	Model Number
REF	0	20181020	NR-BA188PSV1
REF	0	20181027	NR-BJ158SSV1
REF	0	20181022	NR-BJ158SSV1
REF	0	20181026	NR-BJ158SSV1
REF	0	20181019	NR-BJ158SSV1
REF	0	20181028	NR-BJ158SSV1
BEAUTY	2	20181030	EH-NE20-K645
BEAUTY	0	20181013	EH-NE20-K645
BEAUTY	2	20181029	EH-NE20-K645
BEAUTY	0	20181021	EH-NE20-K645

10 rows

|<

<

1-10 of 90

>

>|

CHỌN FILE

test.xlsx

TẢI LÊN

Nhập số dòng để xem trước

Bắt đầu

Kết thúc

10100

XEM TRƯỚC

Tùy chọn gom nhóm

Số nhóm

Số lần lặp

3300

CHỌN LẠI

GOM NHÓM

2.4. Giao diện kết quả gom nhóm

Sau khi gom nhóm, kết quả gom nhóm sẽ được hiển thị tóm tắt để người dùng có thể xem trước khi tải xuống kết quả.

Các thành phần chính:

- Danh sách nhóm đã gom được, mỗi nhóm bao gồm nội dung tóm tắt của nhóm đó và nút tải xuống để tải về dữ liệu của nhóm đó.
- Input [type = number]: cho phép nhập dòng bắt đầu để xem trước.
- Input [type = number]: cho phép nhập dòng kết thúc để xem trước.
- Button [type = submit]: nút chọn để tiến hành xem trước.
- Input [type = number]: cho phép nhập số nhóm cần gom.
- Input [type = number]: để nhập số lần lặp tối đa của thuật toán.
- Button [type = submit]: nút chọn để tiến hành gom nhóm

The screenshot displays the 'Data Analysis' web application interface. It features two main data group panels on the left, each with a large number (1 and 2) and a table of statistics. To the right of these panels are three control panels: 'CHỌN FILE', 'Nhập số dòng để xem trước', and 'Tùy chọn gom nhóm'.

Data Group 1:

Level5 Name	Actual Sales Quantity	Day ID	Model Number
count: 140 freq: 89 top: BEAUTY unique: 6	count: 140 min: -1 max: 36000 mean: 257.5928571429	count: 140 min: 20181001 max: 20181031 mean: 20181018.364285715	count: 140 freq: 35 top: EH-ND11-W645 unique: 21

Data Group 2:

Level5 Name	Actual Sales Quantity	Day ID	Model Number
count: 478 freq: 113 top: REF unique: 11	count: 478 min: -2 max: 250 mean: 2.9644351464	count: 478 min: 20181002 max: 20181031 mean: 20181019.288702928	count: 478 freq: 27 top: NA-F100A4GRV unique: 106

CHỌN FILE: A text input field containing 'test.xlsx' and a 'TẢI LÊN' button.

Nhập số dòng để xem trước: Two input fields for 'Bắt đầu' (10) and 'Kết thúc' (100), with an 'XEM TRƯỚC' button.

Tùy chọn gom nhóm: Two input fields for 'Số nhóm' (2) and 'Số lần lặp' (300), with 'CHỌN LẠI' and 'GOM NHÓM' buttons.

PHẦN 3 - KẾT LUẬN

1. KẾT QUẢ ĐẠT ĐƯỢC

Kết quả thu được là một ứng dụng web chạy trên máy chủ NodeJS trên nền Ubuntu với chức năng chính là xử lý gom nhóm dữ liệu từ file Excel với số nhóm và độ chính xác tùy chọn. Đồng thời cũng cho phép người dùng xem trước và tải xuống kết quả phân tích.

Ưu điểm:

- Xử lý được lượng dữ liệu lớn.
- Chạy trên nền web nên có thể hoạt động trên hầu hết thiết bị có Internet.
- Giao diện đơn giản, tiện dụng, hỗ trợ cả trên điện thoại thông minh.

Nhược điểm:

- Chỉ hỗ trợ định dạng Excel.
- Chưa cho phép lọc dữ liệu theo giá trị cột.
- Chưa cung cấp giao diện biểu đồ trực quan.
- Chỉ triển khai được trên máy chủ Linux có hỗ trợ Python, Shell script và NodeJS.

2. HƯỚNG PHÁT TRIỂN

Với kết quả hiện tại, đề tài có thể được nghiên cứu thêm để khắc phục các hạn chế đang gặp phải hoặc cải tiến thêm các chức năng mới như:

- Hỗ trợ thêm các định dạng tập tin khác ngoài Excel.
- Bổ sung các phương pháp lọc dữ liệu nâng cao.
- Thêm các biểu đồ để thể hiện dữ liệu trực quan.
- Hỗ trợ thêm các thuật toán khác ngoài Cluster.
- Hỗ trợ xử lý dữ liệu chuỗi và số

TÀI LIỆU THAM KHẢO

- [1] “Clustering.” *Scikit*, scikit-learn.org/stable/modules/clustering.html.
- [2] Foundation, Node.js. *Node.js*, nodejs.org/en.
- [3] “Node.js v13.1.0 Documentation.” *Child Process* | *Node.js v13.1.0 Documentation*, nodejs.org/api/child_process.html.
- [4] “NumPy.” *Wikipedia*, Wikimedia Foundation, 11 Nov. 2019, en.wikipedia.org/wiki/NumPy.
- [5] “Python Data Analysis Library.” *Pandas*, pandas.pydata.org/.
- [6] “React – A JavaScript Library for Building User Interfaces.” – *A JavaScript Library for Building User Interfaces*, reactjs.org/.
- [7] “xlsx2csv Package in Ubuntu.” *Launchpad*, launchpad.net/ubuntu/source/xlsx2csv.