

TRƯỜNG ĐẠI HỌC TÀI CHÍNH - MARKETING

KHOA CÔNG NGHỆ THÔNG TIN



ĐỒ ÁN MÔN HỌC

KHAI PHÁ DỮ LIỆU

**SỬ DỤNG THUẬT TOÁN K-MEANS
ĐỂ PHÂN TÍCH KHÁCH HÀNG DỰA TRÊN
HÀNH VI MUA HÀNG TRÊN TIKTOKSHOP**

Giảng viên hướng dẫn : Th.S Thái Thị Ngọc Lý

Sinh viên thực hiện 1 : Nguyễn Thị Kim Kiều

Sinh viên thực hiện 2 : Lâm Hồ Thiên Tổng

Mã lớp học phần : 2311112005904

TP.HCM, tháng 4 năm 2023

TRƯỜNG ĐẠI HỌC TÀI CHÍNH - MARKETING

KHOA CÔNG NGHỆ THÔNG TIN



ĐỒ ÁN MÔN HỌC

KHAI PHÁ DỮ LIỆU

**SỬ DỤNG THUẬT TOÁN K-MEANS
ĐỂ PHÂN TÍCH KHÁCH HÀNG DỰA TRÊN
HÀNH VI MUA HÀNG TRÊN TIKTOKSHOP**

Giảng viên hướng dẫn : Th.S Thái Thị Ngọc Lý

Sinh viên thực hiện 1 : 2021010180 - Nguyễn Thị Kim Kiều

Sinh viên thực hiện 2 : 2021010319 - Lâm Hồ Thiên Tổng

Mã lớp học phần : 2311112005904

TP.HCM, tháng 4 năm 2023

LỜI CẢM ƠN

Chúng em xin chân thành cảm ơn cô Thái Thị Ngọc Lý đã giúp đỡ, hỗ trợ tận tình cho chúng em hoàn thành đồ án môn học này.

Với vốn kiến thức cũng như kinh nghiệm còn rất khiêm tốn và là bước đầu làm quen với công việc nghiên cứu mang tính thực nghiệm thì chắc chắn kết quả đạt được của chúng em cũng không tránh khỏi những hạn chế nhất định. Chúng em rất mong muốn được các giảng viên, những bạn sinh viên đi trước hay bất kỳ độc giả nào quan tâm và góp ý để hoàn thiện hơn cho các đồ án cũng như các nghiên cứu tiếp theo của mình.

Xin kính chúc thầy cùng tất cả những người đã hỗ trợ và đóng góp ý kiến cho chúng em cùng những người thân của mình lời chúc sức khỏe, hạnh phúc và thành đạt.

Xin chân thành cảm ơn.

Sinh viên: Lâm Hồ Thiên Tống - Nguyễn Thị Kim Kiều.

ĐÁNH GIÁ VÀ NHẬN XÉT CỦA GIẢNG VIÊN 1

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

- Điểm số:
- Điểm chữ:

Thành phố Hồ Chí Minh, ngày ... tháng ... năm 2023

Giảng viên hướng dẫn

(Ký, ghi rõ họ tên)

Thái Thị Ngọc Lý

ĐÁNH GIÁ VÀ NHẬN XÉT CỦA GIẢNG VIÊN 2

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

- Điểm số:
- Điểm chữ:

Thành phố Hồ Chí Minh, ngày ... tháng ... năm 2023

Giảng viên

(Ký, ghi rõ họ tên)

DANH MỤC TỪ VIẾT TẮT

Từ viết tắt	Diễn giải
DBMS	Hệ quản trị cơ sở dữ liệu
ETL	Quá trình trích xuất, chuyển đổi và tải dữ liệu
OLAP	Xử lý phân tích trực tuyến

DANH MỤC THUẬT NGỮ ANH - VIỆT

Thuật ngữ	Diễn giải
Data Warehouse	Kho dữ liệu
Columnar Storage	Hệ thống lưu trữ dưới dạng cột
In-memory storage	Hệ thống lưu trữ dạng vùng nhớ đệm
Data Mining	Khai phá dữ liệu
Hyperdata	Siêu dữ liệu
Data Mining	Khai phá dữ liệu

DANH MỤC BẢNG BIỂU

Bảng 2. 1: Bảng mô tả tập dữ liệu khảo sát từ 15 khách hàng.....	19
Bảng 3. 1: Bảng mô tả cấu trúc của dữ liệu	39

DANH MỤC HÌNH ẢNH

Hình 2. 1: Star Schema.....	5
Hình 2. 2: Snowflake Schema.....	6
Hình 2. 3: Fact Constellation Schema.....	7
Hình 2. 4: Trước và Sau khi áp dụng thuật toán K-means.....	18
Hình 3. 1: Giao diện phần mềm Weka	26
Hình 3. 2: Giao diện R	31
Hình 3. 3: Python	34
Hình 4. 1: Danh sách dữ liệu khảo sát về 100 khách hàng	42
Hình 4. 2: Quy ước chuyển đổi các thuộc tính từ dạng chữ sang dạng số.....	43
Hình 4. 3: Dữ liệu sau khi tiền xử lý dữ liệu.....	46
Hình 4. 4: Tách thuộc tính “Tần suất mua hàng” và “Thời gian”.....	46
Hình 4. 5: Tách thuộc tính “Tần suất mua hàng”, “Số lượng sản phẩm” và “Giá trị đơn hàng”	47
Hình 4. 6: Tách thuộc tính “Loại sản phẩm quan tâm” và “Giá trị đơn hàng”	47
Hình 4. 7: Các thuộc tính được trích từ tập dữ liệu khảo sát thành các dữ liệu nhỏ.....	48
Hình 4. 8: Mở phần mềm Weka, chọn Explorer	49
Hình 4. 9: Chọn tệp dữ liệu cần phân cụm.....	49
Hình 4. 10: Sau khi mở file thành công	50
Hình 4. 11: Chọn vào tab Cluster.....	50
Hình 4. 12: Chọn vào thuật toán SimpleKMeans	51
Hình 4. 13: Điều chỉnh các giá trị của thuật toán.....	51
Hình 4. 14: Kết quả của thuật toán.....	52
Hình 4. 15: Kết quả của thuật toán.....	53
Hình 4. 16: Hiện thị mô hình hóa dữ liệu trên kết quả	54
Hình 4. 17: Dữ liệu được trực quan hóa.....	55
Hình 4. 18: Chọn “Select Instance” để cho hiện thị bản ghi khi nhấn vào các quan sát.....	56

Hình 4. 19: Thông tin của quan sát	56
Hình 4. 20: Đánh giá mô hình dựa trên chỉ số đánh giá “Sum of Square Errors”	57
Hình 4. 21: Triển khai mô hình thực hiện dự đoán	57
Hình 4. 22: Mở Weka, và chọn KnowledgeFlow	60
Hình 4. 23: Chọn DataSources	60
Hình 4. 24: Đổ DataSource vào màn hình	61
Hình 4. 25: Chọn Configure để nạp dữ liệu	61
Hình 4. 26: Chọn Browse	62
Hình 4. 27: Chọn vào file cần chạy thuật toán	63
Hình 4. 28: File đã được hiển thị	63
Hình 4. 29: Chọn vào TrainingSetMaker tại Evaluation	63
Hình 4. 30: Chọn vào dataSet	64
Hình 4. 31: Truyền dữ liệu đến TrainingSetMaker	64
Hình 4. 32: Chọn thuật toán SimpleKMeans tại Clusterers	65
Hình 4. 33: Chọn TrainingSet	65
Hình 4. 34: Truyền dữ liệu vào thuật toán SimpleKmeans	65
Hình 4. 35: Tạo TextViewer để chứa kết quả thuật toán	66
Hình 4. 36: Chọn Configure để thiết lập thuật toán	66
Hình 4. 37: Chọn giá trị đo và số cụm	67
Hình 4. 38: Chọn vào text	67
Hình 4. 39: Truyền kết quả vào TextViewer	68
Hình 4. 40: Chạy thuật toán	68
Hình 4. 41: Hiển thị thông báo chạy thành công	68
Hình 4. 42: Chọn Show results để hiện kết quả thuật toán	69
Hình 4. 43: Màn hình hiển thị kết quả của thuật toán	69
Hình 4. 44: Gọi hàm trong R	72
Hình 4. 45: Đọc dữ liệu bằng hàm read.csv và lưu vào biến data	72
Hình 4. 46: Viết hàm “wssplot”	73

Hình 4. 47: Gọi hàm “wssplot”	73
Hình 4. 48: Sử dụng “Elbow Method” để xác định số cụm dựa vào biểu đồ được hiện ra.....	74
Hình 4. 49: Chạy thuật toán K-Means với số cụm là 4, lưu vào biến KM	75
Hình 4. 50: Sử dụng hàm "autoplot" của thư viện “ggfortify” để vẽ biểu đồ trực quan	75
Hình 4. 51: Dữ liệu được trực quan hóa với 4 cụm	76
Hình 4. 52: Dữ liệu tâm các cụm được trả về	77
Hình 4. 53: Mở Python.....	77
Hình 4. 54: Import thư viện “Pandas” vào Python	78
Hình 4. 55: Cài đặt thư viện “Pandas” bằng pip trong “cmd”	78
Hình 4. 56: Đọc file bằng hàm read_csv của thư viện “Pandas”	79
Hình 4. 57: Import thư viện “sklearn”, “numpy”, “matplotlib” vào Python	80
Hình 4. 58: Khởi tạo đối tượng kmeans.....	80
Hình 4. 59: Thực hiện phân cụm dựa trên các thuộc tính	80
Hình 4. 60: In tọa độ tâm cụm.....	81
Hình 4. 61: Gán màu tương ứng từng cụm	81
Hình 4. 62: Tạo hình ảnh và thêm một trục 3D	81
Hình 4. 63: Truyền dữ liệu vào hình ảnh và thiết lập các trục tương ứng các thuộc tính	82
Hình 4. 64: Lệnh hiển thị hình ảnh	83
Hình 4. 65: Đồ thị 3D trực quan hóa 1.....	83
Hình 4. 66: Đồ thị 3D trực quan hóa 2.....	84

MỤC LỤC

LỜI CẢM ƠN	I
ĐÁNH GIÁ VÀ NHẬN XÉT CỦA GIẢNG VIÊN 1	II
ĐÁNH GIÁ VÀ NHẬN XÉT CỦA GIẢNG VIÊN 2	III
DANH MỤC TỪ VIẾT TẮT	IV
DANH MỤC THUẬT NGỮ ANH - VIỆT	V
DANH MỤC BẢNG BIỂU	VI
DANH MỤC HÌNH ẢNH	VII
MỤC LỤC.....	X
CHƯƠNG 1. TỔNG QUAN.....	1
1.1 TỔNG QUAN VỀ TIKTOKSHOP	1
1.2 TỔNG QUAN VỀ ĐỀ TÀI.....	1
1.3 PHẠM VI ĐỀ TÀI	3
CHƯƠNG 2. CƠ SỞ LÝ THUYẾT	4
2.1 KHO DỮ LIỆU	4
2.1.1 Khái niệm kho dữ liệu.....	4
2.1.2 Data Warehouse Schema.....	4
2.1.3 Mô hình dữ liệu đa chiều	8
2.2 KHÁI NIỆM KHAI PHÁ DỮ LIỆU	9
2.2.1 Khái niệm.....	9
2.2.2 Qui trình khai phá dữ liệu	9
2.2.3 Các phương pháp khai phá dữ liệu	10

2.2.4 Các ứng dụng của khai phá dữ liệu.....	14
2.3 PHƯƠNG PHÁP KHAI PHÁ DỮ LIỆU ĐƯỢC SỬ DỤNG TRONG ĐỀ TÀI	18
CHƯƠNG 3. PHẦN MỀM KHAI PHÁ DỮ LIỆU MÃ NGUỒN MỞ	26
3.1 WEKA.....	26
3.1.1 Giới thiệu	26
3.1.2 Chức năng	27
3.1.3 Ưu điểm	28
3.1.4 Nhược điểm.....	29
3.2 R	30
3.2.1 Giới thiệu	30
3.2.2 Chức năng	32
3.2.3 Ưu điểm	33
3.2.4 Nhược điểm.....	33
3.3 PYTHON	34
3.3.1 Giới thiệu	34
3.3.2 Chức năng	35
3.3.3 Ưu điểm	36
3.3.4 Nhược điểm.....	37
CHƯƠNG 4. KHAI PHÁ DỮ LIỆU	38
4.1 XÁC ĐỊNH VẤN ĐỀ	38
4.2 HIỂU DỮ LIỆU	39

4.3 CHUẨN BỊ DỮ LIỆU	41
4.4 LẬP MÔ HÌNH VÀ CHẠY THUẬT TOÁN K-MEANS TRÊN WEKA....	48
4.5 ĐÁNH GIÁ MÔ HÌNH.....	57
4.6 TRIỂN KHAI MÔ HÌNH	57
4.7 CHẠY THUẬT TOÁN K-MEANS TRÊN WEKA KNOWLEDGEFLOW	60
4.8 CHẠY THUẬT TOÁN K-MEANS TRÊN R.....	71
4.9 CHẠY THUẬT TOÁN K-MEANS TRÊN PYTHON	77
CHƯƠNG 5. KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN	85
5.1 KẾT LUẬN.....	85
5.1.1 Những kết quả đạt được.....	85
5.1.2 Hạn chế	85
5.2 HƯỚNG PHÁT TRIỂN.....	86
5.2.1 Hướng khắc phục những hạn chế	86
5.2.2 Hướng mở rộng của đề tài	86
TÀI LIỆU THAM KHẢO.....	87

CHƯƠNG 1. TỔNG QUAN

1.1 Tổng quan về TiktokShop

TiktokShop là một nền tảng thương mại điện tử trực tuyến, cho phép người dùng mua và bán các sản phẩm trên ứng dụng TikTok. Nền tảng này được ra mắt vào năm 2020 và được phát triển bởi ByteDance - công ty mẹ của TikTok. TiktokShop cung cấp cho người dùng trên TikTok một trải nghiệm mua sắm tiện lợi, nhanh chóng và trực quan, thông qua các video quảng cáo sản phẩm được đăng trên ứng dụng TikTok.

TiktokShop cũng có các tính năng giúp người bán quản lý đơn hàng và giao hàng cho khách hàng, bao gồm tính năng đặt hàng, thanh toán trực tuyến, quản lý đơn hàng và vận chuyển hàng hóa. Nền tảng này cũng cung cấp cho người bán các công cụ quảng cáo và phân tích dữ liệu để giúp họ quản lý và tối ưu hóa chiến dịch quảng cáo trên TikTok.

Hiện tại, TiktokShop đã chính thức ra mắt ở Việt Nam từ cuối năm 2020 và đang trở thành một trong những nền tảng thương mại điện tử được sử dụng phổ biến tại Việt Nam, đặc biệt trong giới trẻ. TiktokShop cho phép người dùng mua sắm và thanh toán trực tuyến thông qua ứng dụng TikTok, giúp tăng tính tiện lợi và nhanh chóng cho người dùng.

Việc phát triển nền tảng thương mại điện tử trên TikTok cũng được đánh giá là một trong những xu hướng mới của thị trường thương mại điện tử tại Việt Nam, với tiềm năng phát triển lớn. Do đó, việc nghiên cứu và ứng dụng các thuật toán phân tích dữ liệu để tối ưu hóa hoạt động kinh doanh trên TiktokShop là rất cần thiết.

1.2 Tổng quan về đề tài

Phân tích dữ liệu khách hàng là một quá trình quan trọng trong hoạt động kinh doanh của một doanh nghiệp. Việc phân tích dữ liệu khách hàng giúp các doanh nghiệp

Sử dụng thuật toán K-Means để phân tích khách hàng dựa trên hành vi mua hàng trên TiktokShop

hiểu rõ hơn về hành vi mua hàng, nhu cầu và sở thích của khách hàng, từ đó có thể đưa ra các chiến lược kinh doanh và marketing hiệu quả để tăng doanh số và lợi nhuận.

Cụ thể, phân tích dữ liệu khách hàng giúp:

- Hiểu rõ hơn về khách hàng: Dữ liệu khách hàng cung cấp thông tin về độ tuổi, giới tính, địa chỉ, sở thích và hành vi mua hàng của khách hàng. Từ đó, doanh nghiệp có thể xác định đối tượng khách hàng tiềm năng, đưa ra các sản phẩm và dịch vụ phù hợp với nhu cầu của họ.
- Tối ưu hóa chiến lược marketing: Phân tích dữ liệu khách hàng giúp doanh nghiệp hiểu rõ hơn về những kênh marketing hiệu quả để tiếp cận khách hàng tiềm năng. Từ đó, doanh nghiệp có thể đưa ra các chiến lược marketing phù hợp với đối tượng khách hàng, tối ưu hóa chi phí và tăng tỷ lệ chuyển đổi.
- Đưa ra các chiến lược kinh doanh phù hợp: Phân tích dữ liệu khách hàng giúp doanh nghiệp hiểu rõ hơn về hành vi mua hàng của khách hàng, từ đó có thể đưa ra các chiến lược kinh doanh phù hợp như đưa ra các chương trình khuyến mãi, giảm giá để thu hút khách hàng.

Đề tài "***Sử dụng thuật toán Kmeans để phân tích khách hàng dựa trên hành vi mua hàng trên TiktokShop***" là một đề tài nghiên cứu trong lĩnh vực phân tích dữ liệu và khai thác dữ liệu khách hàng. Mục đích của đề tài là sử dụng thuật toán Kmeans để phân tích dữ liệu khách hàng trên TiktokShop, định tính hóa các thông tin về hành vi mua hàng, và tìm ra những nhóm khách hàng có các hành vi mua hàng tương tự nhau. Kết quả của đề tài sẽ giúp nhà bán hàng trên TiktokShop hiểu rõ hơn về hành vi mua hàng của khách hàng và đưa ra chiến lược kinh doanh phù hợp.

Sử dụng thuật toán K-Means để phân tích khách hàng dựa trên hành vi mua hàng trên TiktokShop

1.3 Phạm vi đề tài

Phạm vi đề tài "Sử dụng thuật toán Kmeans để phân tích khách hàng dựa trên hành vi mua hàng trên TiktokShop" là tập trung vào việc áp dụng thuật toán Kmeans để phân tích dữ liệu hành vi mua hàng của khách hàng trên nền tảng TiktokShop.

Cụ thể, phạm vi của đề tài sẽ bao gồm:

Xác định được vấn đề: Hiểu được các yêu cầu và mục tiêu của án. Từ đó sẽ thu thập dữ liệu, để thực hiện phân tích dữ liệu, đề tài sẽ tiến hành thu thập dữ liệu hành vi mua hàng của khách hàng trên nền tảng TiktokShop.

Hiểu được dữ liệu: Hiểu được các dữ liệu về hành vi mua hàng của khách hàng trên TiktokShop đã thu thập được, và mô tả cấu trúc của dữ liệu đó.

Chuẩn bị dữ liệu: Dữ liệu thu thập được sẽ được tiền xử lý, loại bỏ các dữ liệu không cần thiết, chuẩn hóa dữ liệu và xử lý các giá trị khuyết.

Áp dụng thuật toán Kmeans: Sau khi tiền xử lý dữ liệu, đề tài sẽ áp dụng thuật toán Kmeans để phân tích hành vi mua hàng của khách hàng trên TiktokShop.

Đánh giá kết quả phân tích: Cuối cùng, đề tài sẽ đánh giá kết quả phân tích để đưa ra những nhận xét và giải pháp thích hợp để cải thiện hoạt động kinh doanh trên TiktokShop.

Triển khai mô hình: Thực hiện các kết quả dự đoán và giải thích các kết quả.

Tuy nhiên, đề tài sẽ không bao gồm việc triển khai các giải pháp để cải thiện hoạt động kinh doanh trên TiktokShop, mà chỉ tập trung vào việc phân tích dữ liệu khách hàng để đưa ra những dự đoán.

CHƯƠNG 2. CƠ SỞ LÝ THUYẾT

2.1 Kho dữ liệu

2.1.1 Khái niệm kho dữ liệu

Kho dữ liệu (Data Warehouse) là một hệ thống lưu trữ dữ liệu được thiết kế để hỗ trợ việc phân tích dữ liệu và ra quyết định trong tổ chức. Kho dữ liệu được thiết kế để hỗ trợ việc tập hợp dữ liệu từ nhiều nguồn khác nhau, tiêu chuẩn hóa dữ liệu, lưu trữ dữ liệu lịch sử và cung cấp các công cụ phân tích dữ liệu để phục vụ cho các nhu cầu kinh doanh, quản lý, phân tích, đánh giá và dự báo.

Một kho dữ liệu thường được xây dựng dựa trên hệ thống quản trị cơ sở dữ liệu (DBMS) và các công nghệ lưu trữ dữ liệu như hệ thống lưu trữ dạng cột (Columnar Storage) và hệ thống lưu trữ dạng vùng nhớ đệm (in-memory storage). Kho dữ liệu cũng có thể được xây dựng bằng cách tập hợp các dữ liệu từ các nguồn khác nhau thông qua quá trình trích xuất, chuyển đổi và tải dữ liệu (ETL).

Mục đích chính của kho dữ liệu là giúp tổ chức quản lý dữ liệu một cách hiệu quả hơn, tạo điều kiện thuận lợi cho việc phân tích dữ liệu và ra quyết định kinh doanh thông qua việc tạo ra các báo cáo, biểu đồ, dashboard và các ứng dụng phân tích dữ liệu khác.

2.1.2 Data Warehouse Schema

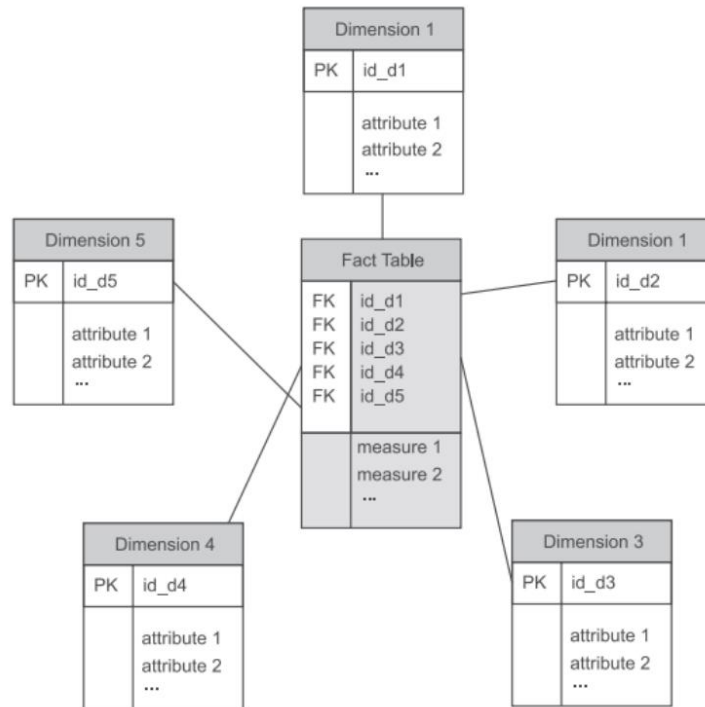
Data Warehouse Schema là một cấu trúc dữ liệu được thiết kế để lưu trữ dữ liệu trong kho dữ liệu (Data Warehouse). Schema được sử dụng để mô tả cách thức tổ chức các bảng và các quan hệ giữa chúng trong kho dữ liệu.

Có ba loại schema chính được sử dụng trong kho dữ liệu: Star Schema, Snowflake Schema, Fact Constellation Schema (Galaxy Schema).

Mỗi kiểu schema có ưu điểm và nhược điểm riêng, tùy thuộc vào yêu cầu kinh doanh và dữ liệu được lưu trữ trong kho dữ liệu.

Sử dụng thuật toán K-Means để phân tích khách hàng dựa trên hành vi mua hàng trên TiktokShop

2.1.2.1 Star Schema



Hình 2. 1: Star Schema

Trong Star Schema, các bảng dữ liệu được tổ chức sao cho mỗi bảng chứa thông tin về một sự kiện cụ thể, ví dụ như đơn hàng hoặc hóa đơn, và được kết nối với nhau thông qua một bảng trung gian gọi là bảng kết nối (fact table).

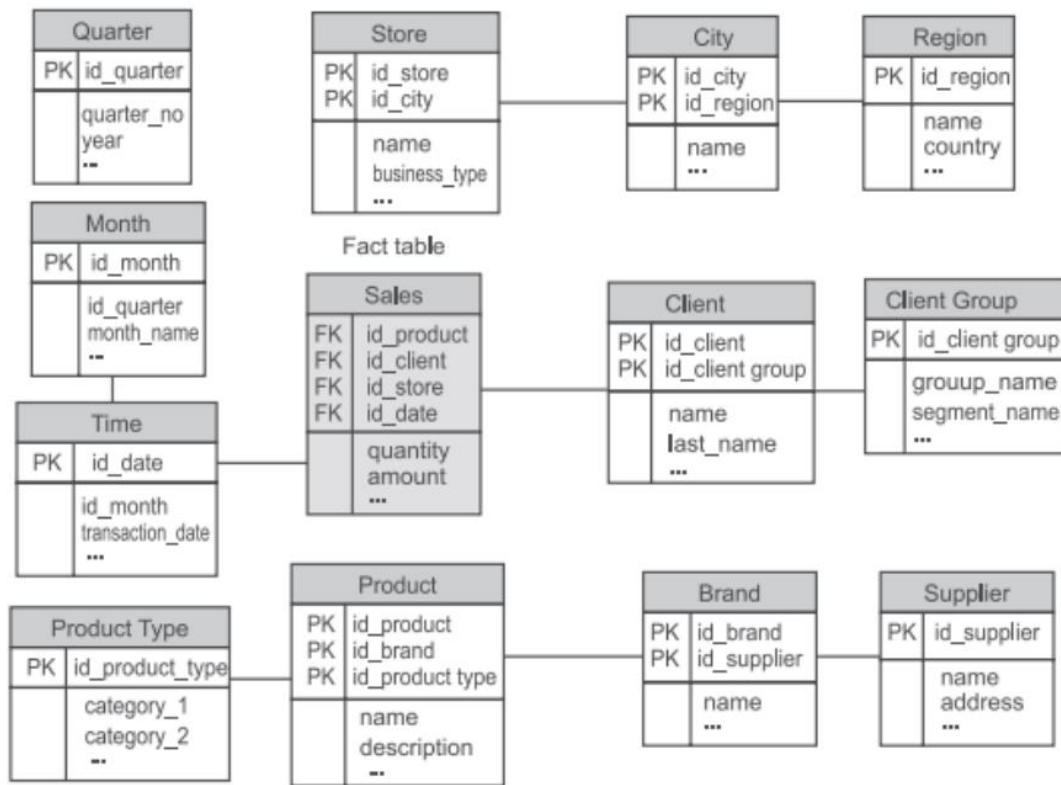
Bảng kết nối chứa các thuộc tính chung của các bảng dữ liệu, nhưng không chứa thông tin chi tiết về các sự kiện cụ thể. Thay vào đó, nó chứa các khóa ngoại của các bảng chi tiết để kết nối chúng lại với nhau.

Với Star Schema, các bảng dữ liệu được thiết kế sao cho chúng có cấu trúc đơn giản và dễ dàng truy vấn. Điều này làm cho việc truy xuất dữ liệu và phân tích dữ liệu trở nên dễ dàng hơn. Nó cũng cho phép tối ưu hóa hiệu suất truy vấn và giúp giảm thiểu thời gian truy vấn.

Tuy nhiên, Star Schema cũng có nhược điểm, đó là nó có thể dẫn đến việc lặp lại dữ liệu và tạo ra các bảng trùng lặp, dẫn đến tăng kích thước của kho dữ liệu.

Sử dụng thuật toán K-Means để phân tích khách hàng dựa trên hành vi mua hàng trên TiktokShop

2.1.2.2 Snowflake Schema



Hình 2. 2: Snowflake Schema

Snowflake Schema là một kiểu cấu trúc dữ liệu được sử dụng trong kho dữ liệu (data warehouse). Snowflake Schema giống như Star Schema, tuy nhiên khác ở chỗ các bảng chi tiết được chia nhỏ thành các bảng con, và các bảng con này được kết nối với bảng cha thông qua các bảng trung gian.

Trong Snowflake Schema, bảng kết nối (fact table) được kết nối với các bảng chi tiết thông qua các bảng trung gian. Các bảng trung gian này chứa các thuộc tính chi tiết của các bảng chi tiết và các khóa ngoại để kết nối chúng lại với bảng kết nối.

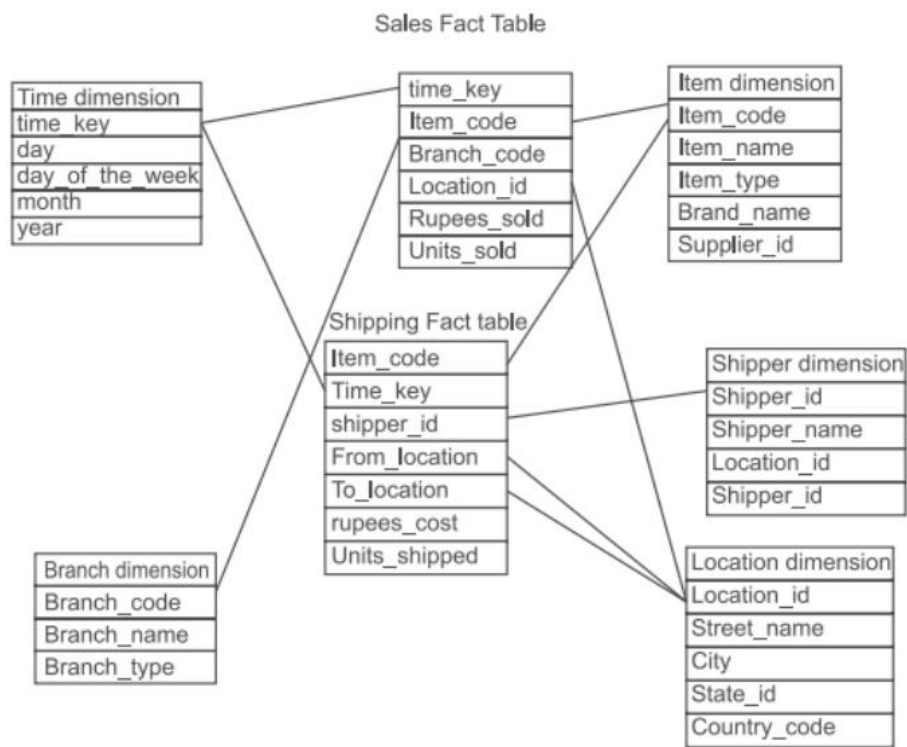
Các bảng chi tiết trong Snowflake Schema được phân chia thành các bảng con, mỗi bảng con chứa một tập hợp các thuộc tính liên quan đến một loại thông tin cụ thể.

Sử dụng thuật toán K-Means để phân tích khách hàng dựa trên hành vi mua hàng trên TiktokShop

Snowflake Schema cho phép tối ưu hóa việc truy xuất dữ liệu và giảm thiểu thời gian truy vấn. Nó cũng giúp giảm kích thước của kho dữ liệu bằng cách chia nhỏ các bảng chi tiết thành các bảng con.

Tuy nhiên, Snowflake Schema cũng có nhược điểm, đó là nó có thể dẫn đến phức tạp hóa thiết kế và truy xuất dữ liệu. Nó cũng có thể tạo ra các bảng trung gian trùng lặp, làm tăng kích thước của kho dữ liệu.

2.1.2.3 Fact Constellation Schema (Galaxy Schema)



Hình 2. 3: Fact Constellation Schema

Fact Constellation Schema là một mô hình cấu trúc dữ liệu trong kho dữ liệu (data warehouse) cho phép lưu trữ thông tin từ nhiều bảng kết nối (fact table) và các bảng chi tiết (dimension table) liên quan đến chúng.

Sử dụng thuật toán K-Means để phân tích khách hàng dựa trên hành vi mua hàng trên TiktokShop

Trong Fact Constellation Schema, các bảng kết nối và bảng chi tiết được tổ chức thành nhiều nhóm liên quan đến nhau, gọi là hình sao (star schema), với mỗi hình sao tương ứng với một tập hợp các bảng kết nối và các bảng chi tiết có liên quan.

Một hình sao trong Fact Constellation Schema bao gồm một bảng kết nối trung tâm và các bảng chi tiết được kết nối trực tiếp với nó. Ngoài ra, mỗi bảng chi tiết còn có thể kết nối với các bảng chi tiết khác.

Fact Constellation Schema cho phép lưu trữ dữ liệu theo các mô hình phức tạp và đa dạng, và giúp tối ưu hóa việc truy vấn dữ liệu trong kho dữ liệu. Nó cũng cho phép dễ dàng mở rộng kho dữ liệu khi cần thiết.

Tuy nhiên, Fact Constellation Schema cũng có nhược điểm, đó là thiết kế và triển khai có thể phức tạp hơn so với các mô hình cấu trúc dữ liệu khác. Nó cũng có thể dẫn đến các vấn đề về hiệu suất nếu không được xây dựng và quản lý đúng cách.

2.1.3 Mô hình dữ liệu đa chiều

Mô hình dữ liệu đa chiều (multidimensional data model) là một mô hình được sử dụng để lưu trữ dữ liệu trong các hệ thống quản lý cơ sở dữ liệu (DBMS). Khác với *mô hình dữ liệu dạng bảng* (relational data model) có dữ liệu được tổ chức dưới dạng bảng với các hàng và cột, thì *mô hình dữ liệu đa chiều* cho phép dữ liệu được tổ chức dưới dạng hình khối (cube) hoặc siêu dữ liệu (hyperdata). Trong đó dữ liệu được phân tích thành nhiều chiều, mỗi chiều biểu diễn một khía cạnh của dữ liệu. Mô hình này cho phép dữ liệu được truy xuất và phân tích dễ dàng hơn, bởi vì dữ liệu được phân tách ra thành các đối tượng con và các thông tin chi tiết có thể được lấy ra một cách độc lập.

Mô hình dữ liệu đa chiều phù hợp cho các hệ thống dữ liệu lớn và đòi hỏi phân tích dữ liệu phức tạp, như data warehouse, business intelligence hay data mining. Một trong những ứng dụng phổ biến của mô hình dữ liệu đa chiều là OLAP (Online Analytical Processing), trong đó dữ liệu được phân tích từ nhiều góc độ khác nhau để đưa ra quyết định hoặc dự đoán trong thời gian thực.

Sử dụng thuật toán K-Means để phân tích khách hàng dựa trên hành vi mua hàng trên TiktokShop

2.2 Khái niệm khai phá dữ liệu

2.2.1 Khái niệm

Khai phá dữ liệu (data mining) là quá trình khám phá và tìm ra các mối quan hệ, khuynh hướng, kiến thức ẩn và thông tin hữu ích từ trong cơ sở dữ liệu lớn và phức tạp.

Khai phá dữ liệu sử dụng các kỹ thuật, phương pháp và công cụ để phân tích và xử lý các dữ liệu lớn và phức tạp nhằm tìm ra các mô hình, quy luật, thông tin quan trọng, cấu trúc ẩn, hay các dạng kiến thức hữu ích khác giúp hỗ trợ quyết định, dự báo, hay hiểu rõ hơn về các vấn đề trong thực tế.

Khai phá dữ liệu giúp cho người dùng có thể trích xuất thông tin hữu ích, từ đó tạo ra các giải pháp, quyết định và phát triển các ứng dụng và sản phẩm mới. Khai phá dữ liệu được ứng dụng rộng rãi trong nhiều lĩnh vực như kinh doanh, y tế, khoa học dữ liệu, tài chính, bảo mật mạng, v.v.

2.2.2 Quy trình khai phá dữ liệu

Quy trình khai phá dữ liệu bao gồm 6 giai đoạn chính:

Giai đoạn 1: Xác định vấn đề: Giai đoạn này tập trung vào việc xác định mục tiêu và mục đích khai phá dữ liệu, đưa ra các câu hỏi cần trả lời và các giải pháp cần đưa ra để giải quyết vấn đề. Đây là giai đoạn quan trọng nhất trong quy trình khai phá dữ liệu.

Giai đoạn 2: Hiểu dữ liệu: Ở giai đoạn này, người thực hiện khai phá dữ liệu cần hiểu rõ về dữ liệu đang có, các đặc tính của dữ liệu, các mối quan hệ giữa các thuộc tính và các kết quả cần đạt được. Đây là bước để có thể tạo ra các giả định cho bước tiếp theo.

Giai đoạn 3: Chuẩn bị dữ liệu: Giai đoạn này tập trung vào việc chuẩn bị dữ liệu để có thể sử dụng cho việc khai phá. Các bước trong giai đoạn này bao gồm: thu thập dữ liệu, xử lý dữ liệu, lọc dữ liệu, chuẩn hóa dữ liệu, chọn mẫu dữ liệu và chia dữ liệu thành các tập train và test.

Sử dụng thuật toán K-Means để phân tích khách hàng dựa trên hành vi mua hàng trên TiktokShop

Giai đoạn 4: Mô hình hóa: Ở giai đoạn này, người thực hiện khai phá dữ liệu sẽ áp dụng các phương pháp khai phá dữ liệu để tạo ra mô hình. Các phương pháp khai phá dữ liệu có thể bao gồm: phân tích nhân tố, cây quyết định, hồi quy tuyến tính, clustering, mạng nơ-ron, ...

Giai đoạn 5: Đánh giá: Giai đoạn này tập trung vào việc đánh giá mô hình đã được tạo ra ở giai đoạn trước. Các phương pháp đánh giá mô hình có thể bao gồm: cross-validation, hold-out, bootstrapping, ...

Giai đoạn 6: Triển khai: Giai đoạn cuối cùng trong quy trình khai phá dữ liệu là triển khai mô hình đã được tạo ra ở giai đoạn mô hình hóa. Việc triển khai này có thể bao gồm: triển khai trực tiếp, phân phối trên nhiều máy tính, đưa vào sản xuất hoặc chạy trên các hệ thống thời gian thực.

2.2.3 Các phương pháp khai phá dữ liệu

2.2.3.1 Phương pháp phân lớp

Phương pháp phân lớp trong khai phá dữ liệu là một phương pháp dùng để xây dựng mô hình dự đoán một biến mục tiêu dựa trên các biến đầu vào. Nó dựa trên việc học từ dữ liệu đã biết để tạo ra một mô hình dự đoán cho các dữ liệu mới.

Mục đích của phân lớp là xác định phân loại cho các dữ liệu đầu vào dựa trên các thông tin sẵn có, chẳng hạn như thuộc tính hoặc các biến đầu vào. Các phương pháp phân lớp có thể được sử dụng trong nhiều lĩnh vực, bao gồm phân loại hình ảnh, phát hiện gian lận tín dụng, dự đoán độ chính xác của bệnh lý và nhiều ứng dụng khác.

Quy trình gồm 2 bước:

- + Bước học (giai đoạn huấn luyện): xây dựng bộ phận phân loại (classifier) bằng việc phân tích, học tập huấn luyện.

Sử dụng thuật toán K-Means để phân tích khách hàng dựa trên hành vi mua hàng trên TiktokShop

- + Bước phân loại (Classification): phân loại dữ liệu, đối tượng mới nếu độ chính xác của bộ phân loại được đánh giá là có thể chấp nhận được (acceptable)

Các thuật phân loại dữ liệu:

- + Phân loại dữ liệu với cây quyết định (decision tree)
- + Phân loại dữ liệu với mạng Bayesian
- + Phân loại dữ liệu với mạng neural
- + Phân loại dữ liệu với K phần tử gần nhất (K – nearest neighbor)
- + Phân loại dữ liệu với suy diễn dựa trên tình huống (cased – based reasoning)
- + Phân loại dữ liệu dựa trên tiến hóa gen (genetic algorithms)
- + Phân loại dữ liệu với lý thuyết tập thô (rough sets)
- + Phân loại dữ liệu với lý thuyết tập mờ (fuzzy sets)

2.2.3.2 Phương pháp gom cụm

Phương pháp gom cụm (clustering) trong khai phá dữ liệu là một kỹ thuật rất phổ biến được sử dụng để phân nhóm các đối tượng dữ liệu vào những nhóm tương đồng nhau dựa trên các đặc trưng của chúng. Phương pháp này tập trung vào việc phân tích và khám phá các mẫu và cấu trúc bên trong dữ liệu mà không cần có các giả định trước về các lớp hay phân phối của dữ liệu.

Trong phương pháp gom cụm, các điểm dữ liệu được gán vào các nhóm sao cho các điểm trong cùng một nhóm có sự tương đồng lớn hơn với nhau so với các điểm trong các nhóm khác. Sự tương đồng giữa các điểm dữ liệu thường được định nghĩa bằng cách sử dụng khoảng cách Euclid hoặc các phương pháp đo tương đồng khác.

Sử dụng thuật toán K-Means để phân tích khách hàng dựa trên hành vi mua hàng trên TiktokShop

Phương pháp gom cụm được sử dụng rộng rãi trong nhiều lĩnh vực, ví dụ như trong kinh doanh để phân nhóm khách hàng dựa trên hành vi mua hàng, trong y học để phân loại bệnh nhân dựa trên các chỉ số sức khỏe, trong địa lý để phân tích mô hình phân bố địa lý của các đối tượng, và nhiều lĩnh vực khác.

Một số phương pháp phân cụm phổ biến:

- + Phân hoạch (Partitioning): Phương pháp này chia dữ liệu thành các cụm (cluster) độc lập dựa trên khoảng cách giữa các điểm dữ liệu. Phân hoạch phổ biến nhất là phương pháp K-Means.
- + Phân cấp (Hierarchical): Phương pháp này xây dựng các cụm bằng cách liên tiếp gom các điểm dữ liệu lại với nhau. Phân cấp có thể được chia thành hai loại là phân cấp trên (agglomerative) và phân cấp dưới (divisive).
- + Dựa trên mật độ (Density-based): Phương pháp này phát hiện ra các cụm dữ liệu dựa trên mật độ của chúng. Phân cụm dựa trên mật độ phổ biến nhất là DBSCAN.
- + Dựa trên lưới (Grid-based): Phương pháp này chia không gian dữ liệu thành các ô lưới rồi xác định các cụm bằng cách gom các ô lưới lại với nhau.
- + Dựa trên mô hình (Model-based): Phương pháp này dựa trên một mô hình xác định các cụm dữ liệu. Phân cụm dựa trên mô hình phổ biến nhất là Gaussian Mixture Models (GMM).
- + Có ràng buộc (Constraint-based): Phương pháp này giải quyết vấn đề phân cụm với các ràng buộc cần tuân thủ. Ví dụ: phân cụm dữ liệu với các điểm dữ liệu thuộc các cụm khác nhau.

Sử dụng thuật toán K-Means để phân tích khách hàng dựa trên hành vi mua hàng trên TiktokShop

Các phương pháp này có ưu điểm và hạn chế riêng, tùy thuộc vào bộ dữ liệu và mục đích sử dụng, người sử dụng có thể chọn phương pháp phù hợp để phân cụm dữ liệu.

2.2.3.3 Phương pháp luật kết hợp

Phương pháp luật kết hợp (Association Rule Mining) là một trong những phương pháp khai phá dữ liệu phổ biến nhất. Phương pháp này tìm kiếm các mối quan hệ tần suất giữa các mục trong một tập dữ liệu lớn. Nó cho phép khám phá ra các quy tắc kết hợp giữa các mục, nhằm giúp cho các tổ chức đưa ra các quyết định trong kinh doanh và tiếp thị.

Phương pháp luật kết hợp hoạt động bằng cách tìm kiếm các luật kết hợp trong tập dữ liệu, tức là các mẫu thường xuyên xuất hiện cùng nhau. Ví dụ, nếu chúng ta có một tập dữ liệu về mua sắm trực tuyến, phương pháp luật kết hợp có thể giúp chúng ta tìm ra các sản phẩm thường được mua cùng nhau, như giày thể thao và áo thể thao.

Các luật kết hợp thường được viết là $X \rightarrow Y$, tức là:

- + Bất cứ khi nào X xuất hiện Y cũng có xu hướng xuất hiện.
- + X và Y có thể là các mục đơn lẻ hoặc tập hợp các mục.
- + X được gọi là tiền đề của quy tắc và Y là hậu quả của nó.

Một số khái niệm cơ bản trong luật kết hợp

Tập hợp món hàng được biểu diễn là $I = \{i_1, i_2, \dots, i_n\}$ với n là số lượng món hàng.

Tập hợp các giao dịch được kí hiệu là $T = \{t_1, t_2, \dots, t_n\}$ với n là số lượng giao dịch; trong mỗi giao dịch sẽ có m món hàng được ký hiệu là $t_i = \{i_1, i_2, \dots, i_m\}$, $m < n$; mỗi giao dịch được định danh là TID.

Sử dụng thuật toán K-Means để phân tích khách hàng dựa trên hành vi mua hàng trên TiktokShop

Support (độ hỗ trợ): là tần suất xuất hiện của X trong số giao dịch. Kí hiệu: Support (X) hoặc $P(X)$

Confidence (độ tin cậy): là số phần trăm giao dịch luôn chứa Y trong khi chứa X hoặc được định nghĩa là tỷ số giữa độ hỗ trợ của X và Y với độ hỗ trợ của X (giống như xác suất có điều kiện của Y khi X đã xảy ra. Kí hiệu: Confidence of $(X \rightarrow Y)$ hoặc $P(Y|X)$ hoặc $\text{Support}(XY) / \text{Support}(X)$

Lift (độ nâng): là tỷ lệ giữa xác suất có điều kiện của Y khi X được cho với xác suất không điều kiện của Y trong tập dữ liệu hoặc nghĩa là độ tin cậy của $X \rightarrow Y$ chia cho xác suất của Y

2.2.4 Các ứng dụng của khai phá dữ liệu

Việc khai phá dữ liệu có rất nhiều ứng dụng trong nhiều lĩnh vực khác nhau. Dưới đây là một số ứng dụng chính của việc khai phá dữ liệu:

❖ Khoa học dữ liệu:

- Phân tích dữ liệu y học: Khai thác dữ liệu từ các nguồn y tế khác nhau như tài liệu y học, bảng điểm, hồ sơ bệnh án, báo cáo kiểm tra... để xây dựng các mô hình phân loại, dự đoán và phát hiện các mối liên hệ giữa các yếu tố để hỗ trợ việc chẩn đoán và điều trị bệnh tật. Ví dụ như việc áp dụng các mô hình học máy để dự đoán nguy cơ mắc ung thư cổ tử cung dựa trên thông tin của bệnh nhân, giúp phát hiện sớm và đưa ra những phương pháp phòng ngừa hiệu quả hơn.
- Phân tích dữ liệu tài chính: Khai thác dữ liệu từ các nguồn tài chính khác nhau như giao dịch ngân hàng, thẻ tín dụng, lịch sử giá cổ phiếu... để phát hiện các mô hình, xu hướng và tình hình tài chính của doanh nghiệp hoặc cá nhân. Ví dụ như việc sử dụng phân tích dữ liệu để đưa ra dự báo xu hướng giá cổ phiếu, giúp nhà đầu tư đưa ra các quyết định đầu tư hợp lý.

Sử dụng thuật toán K-Means để phân tích khách hàng dựa trên hành vi mua hàng trên TiktokShop

- Phân tích dữ liệu thương mại điện tử: Khai thác dữ liệu từ các nguồn thương mại điện tử như đơn hàng, đánh giá sản phẩm, hành vi mua hàng... để phát hiện xu hướng mua sắm, tìm kiếm, yêu cầu của khách hàng và đưa ra các phương án kinh doanh phù hợp. Ví dụ như việc sử dụng phân tích dữ liệu để tìm kiếm những sản phẩm phổ biến, khuyến mãi tốt nhất để quảng bá trên các kênh quảng cáo.

❖ An ninh mạng:

- Phát hiện tấn công mạng: Khai phá dữ liệu có thể giúp phát hiện các hoạt động tấn công mạng bằng cách xử lý các dữ liệu mạng, bao gồm các giao thức, luồng dữ liệu và thông tin nhận dạng của hệ thống. Các phương pháp khai phá dữ liệu như phân tích hành vi mạng, phát hiện xâm nhập và phát hiện giả mạo địa chỉ IP đều được sử dụng để phát hiện tấn công mạng.
- Giám sát mạng: Khai phá dữ liệu có thể được sử dụng để giám sát mạng và phát hiện các hành vi đáng ngờ như phát tán virus, tấn công từ chối dịch vụ (DDoS), tấn công đánh cắp thông tin cá nhân, vv.
- Phân tích log: Khai phá dữ liệu cũng có thể được sử dụng để phân tích các tập tin log của hệ thống, nhằm giúp định vị các sự kiện an ninh quan trọng và giúp ngăn chặn các cuộc tấn công mạng.
- Phát hiện các tài khoản đăng nhập bị xâm nhập: Các phương pháp khai phá dữ liệu có thể được sử dụng để phát hiện các tài khoản đăng nhập bị xâm nhập bằng cách phân tích các hoạt động đăng nhập, lịch sử đăng nhập và các thông tin liên quan đến người dùng.
- Phát triển chính sách an ninh: Khai phá dữ liệu có thể giúp tạo ra các chính sách an ninh mạng hiệu quả bằng cách phân tích các thông tin về các cuộc tấn công mạng, các mối đe dọa tiềm tàng và các xu hướng an ninh.

❖ Kinh doanh và tiếp thị:

Sử dụng thuật toán K-Means để phân tích khách hàng dựa trên hành vi mua hàng trên TiktokShop

- Phân tích khách hàng: Khai phá dữ liệu khách hàng giúp doanh nghiệp hiểu rõ hơn về đặc tính, hành vi và nhu cầu của khách hàng, từ đó tối ưu hóa chiến lược tiếp thị và kinh doanh. Ví dụ: Sử dụng phương pháp phân cụm để chia khách hàng thành các nhóm dựa trên các đặc tính như tuổi, giới tính, thu nhập, sở thích, v.v. từ đó tạo ra các chiến lược tiếp thị hiệu quả hơn đối với từng nhóm khách hàng.
- Dự báo xu hướng thị trường: Khai phá dữ liệu thị trường giúp doanh nghiệp dự báo được xu hướng của thị trường, nhận diện các cơ hội mới và đưa ra quyết định kinh doanh chính xác hơn. Ví dụ: Phân tích dữ liệu từ các trang mạng xã hội, diễn đàn để đánh giá các bài viết, ý kiến của khách hàng về sản phẩm, dịch vụ để đưa ra những cải tiến mới phù hợp với nhu cầu thị trường.
- Đo lường hiệu quả chiến dịch tiếp thị: Khai phá dữ liệu tiếp thị giúp doanh nghiệp đo lường hiệu quả của các chiến dịch tiếp thị, từ đó tối ưu hóa chiến lược tiếp thị và kinh doanh. Ví dụ: Sử dụng phương pháp phân tích đường cong ROC để đánh giá hiệu quả của chiến dịch quảng cáo trực tuyến, từ đó điều chỉnh chiến lược quảng cáo để đạt được hiệu quả cao hơn.
- Dự báo doanh số: Khai phá dữ liệu doanh số giúp doanh nghiệp dự báo được doanh số bán hàng, từ đó có thể đưa ra các chiến lược kinh doanh phù hợp. Ví dụ: Sử dụng phương pháp phân tích chuỗi thời gian để dự báo doanh số bán hàng của một sản phẩm trong tương lai.

❖ Trí tuệ nhân tạo

Thật thiếu sót nếu nhắc đến ứng dụng của khai phá dữ liệu mà không nói đến trí tuệ nhân tạo, trí tuệ nhân tạo (Artificial Intelligence - AI) và khai phá dữ liệu (Data Mining) là hai lĩnh vực gần gũi nhau và thường được áp dụng kết hợp với nhau trong nhiều ứng dụng.

Sử dụng thuật toán K-Means để phân tích khách hàng dựa trên hành vi mua hàng trên TiktokShop

Khai phá dữ liệu cung cấp những dữ liệu quan trọng và kiến thức cần thiết cho việc xây dựng các hệ thống trí tuệ nhân tạo. Chẳng hạn như khi xây dựng một mô hình học máy, chúng ta cần dữ liệu để huấn luyện và kiểm tra mô hình. Dữ liệu được lựa chọn, tiền xử lý và trích xuất thông tin từ quá trình khai phá dữ liệu có thể giúp cải thiện độ chính xác của mô hình học máy.

Ngoài ra, khai phá dữ liệu và trí tuệ nhân tạo cũng có ứng dụng trong việc phát hiện dữ liệu gian lận, xác định các nhóm khách hàng tiềm năng, tối ưu hóa quy trình sản xuất, tối ưu hóa chiến lược marketing và nhiều lĩnh vực khác. Ví dụ, công ty Amazon sử dụng khai phá dữ liệu để phát hiện xu hướng mua sắm của khách hàng, đồng thời sử dụng trí tuệ nhân tạo để tạo ra các gợi ý sản phẩm và chiến lược tiếp thị định hướng tới từng khách hàng cụ thể.

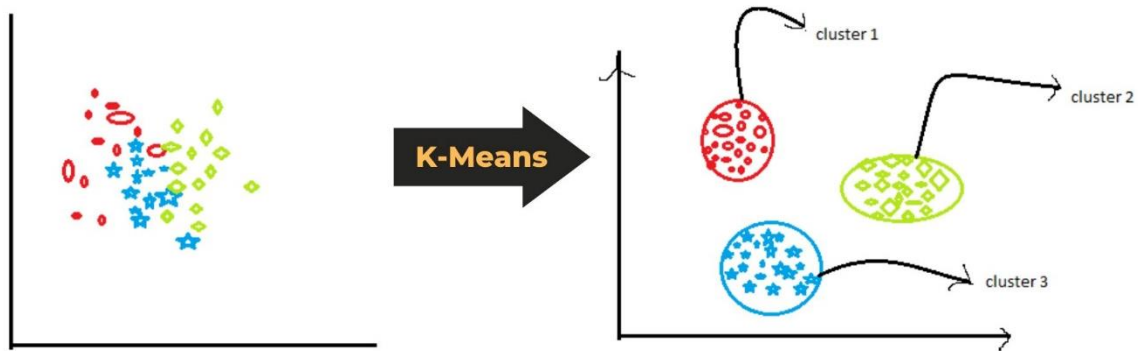
Tóm lại, khai phá dữ liệu và trí tuệ nhân tạo là hai lĩnh vực có quan hệ mật thiết và thường được kết hợp với nhau trong nhiều ứng dụng. Việc áp dụng kết hợp hai lĩnh vực này đem lại nhiều lợi ích trong việc tối ưu hóa các quy trình và cải thiện hiệu quả kinh doanh.

Như vậy, khai phá dữ liệu có nhiều ứng dụng hữu ích trong nhiều lĩnh vực khác nhau, từ trí tuệ nhân tạo, khoa học dữ liệu, an ninh mạng, đến kinh doanh và tiếp thị, ngoài ra còn rất nhiều lĩnh vực khác, ... Các phương pháp khai phá dữ liệu giúp cho các nhà quản trị và nhà nghiên cứu có thể xử lý và phân tích các tập dữ liệu lớn để tìm ra các mối quan hệ ẩn giấu và đưa ra các dự báo hữu ích. Các ví dụ cụ thể như phát hiện gian lận tín dụng, tăng doanh số bán hàng, cải thiện trải nghiệm người dùng, hay tối ưu hóa chiến lược marketing, ... đều chứng minh được tính hiệu quả và tiềm năng của khai phá dữ liệu trong thực tế.

Sử dụng thuật toán K-Means để phân tích khách hàng dựa trên hành vi mua hàng trên TiktokShop

2.3 Phương pháp khai phá dữ liệu được sử dụng trong đề tài

Trình bày thuật toán



Hình 2. 4: Trước và Sau khi áp dụng thuật toán K-means

Thuật toán K-Means là một trong những thuật toán gom cụm phổ biến trong khai phá dữ liệu. Thuật toán này được sử dụng để phân cụm dữ liệu, trong đó các điểm dữ liệu được phân vào các nhóm dựa trên sự tương đồng về đặc trưng giữa chúng.

Trong đề tài, thuật toán K-Means được sử dụng để phân tích và phân loại khách hàng dựa trên hành vi mua hàng của họ. Bằng cách áp dụng thuật toán này trên dữ liệu hành vi mua hàng (họ và tên, giới tính, khu vực khách hàng, tần suất mua hàng mỗi tháng, số lượng sản phẩm trung bình trong mỗi đơn hàng, giá trị trung bình mỗi đơn hàng, thời gian đặt hàng), sau đó các nhóm khách hàng có các đặc trưng tương tự nhau sẽ được gom lại thành các cụm.

Các bước thực hiện của thuật toán:

Bước 1: Xác định số lượng cụm: Trước khi triển khai thuật toán, ta cần xác định số lượng cụm phù hợp để phân chia khách hàng dựa trên hành vi mua hàng. Số lượng này có thể được xác định bằng cách sử dụng các phương pháp như phương pháp Elbow hoặc phương pháp Silhouette.

Sử dụng thuật toán K-Means để phân tích khách hàng dựa trên hành vi mua hàng trên TiktokShop

Bước 2: Khởi tạo các trung tâm cụm: Sau khi xác định được số lượng cụm, ta cần khởi tạo các trung tâm ban đầu cho từng cụm. Các trung tâm này có thể được chọn ngẫu nhiên từ dữ liệu hoặc dựa trên các giá trị trung bình của các đặc trưng của dữ liệu.

Bước 3: Gán các điểm dữ liệu vào các cụm: Tiếp theo, ta sử dụng các giá trị của các đặc trưng của từng khách hàng để tính toán khoảng cách giữa khách hàng đó và các trung tâm cụm. Sau đó, ta sẽ gán khách hàng đó vào cụm có trung tâm gần nhất.

Bước 4: Cập nhật các trung tâm cụm: Sau khi gán các khách hàng vào các cụm, ta sẽ tính toán lại giá trị trung bình của các đặc trưng của các khách hàng trong cùng một cụm để cập nhật lại trung tâm của cụm.

Bước 5: Lặp lại quá trình: Tiếp tục lặp lại quá trình gán các điểm dữ liệu vào các cụm và cập nhật các trung tâm cụm cho đến khi không có sự thay đổi nào nữa

Minh họa ví dụ bằng từng bước thực hiện

Một tập dữ liệu khảo sát từ 15 khách hàng, gồm có tần suất mua hàng mỗi tháng, giá trị mỗi đơn hàng, được thể hiện như sau:

Bảng 2. 1: Bảng mô tả tập dữ liệu khảo sát từ 15 khách hàng

Tên Khách Hàng	Tần suất mua hàng mỗi tháng	Giá trị mỗi đơn hàng
Lâm Hồ Thiên Tổng	5	500
Vũ Vương Vinh	5	800
Nguyễn Kiều Nhã Linh	10	300
Đinh Như Ý	6	200
Lương Công Tiến	2	500
Vũ Tường Nguyên	7	800
Doãn Thị Đài Trang	2	300
Nguyễn Phùng Vân Anh	15	900
Nguyễn Thị Kim Kiều	2	200

Sử dụng thuật toán K-Means để phân tích khách hàng dựa trên hành vi mua hàng trên TiktokShop

Dương Thị Bích Muội	3	100
Nguyễn Thùy Trang	1	600
Nguyễn Ngọc Ánh	3	300
Nguyễn Lâm Gia Thịnh	2	300
Hồ Lâm Duy Khang	3	800
Hồ Lâm Gia Mỹ	4	700

❖ **Bước 1: Xác định số cụm**

$k = 2$.

❖ **Bước 2: Khởi tạo trung tâm cụm**

Chọn ngẫu nhiên k tâm cụm

- Tâm cụm 1 (C1) chọn khách hàng Vũ Vương Vinh (5,800).
- Tâm cụm 2 (C2) chọn khách hàng Nguyễn Lâm Gia Thịnh (2,300).

❖ **Bước 3:**

Công thức tính khoảng cách Euclidean dist:

$$\text{Euclidean dist}((x,y),(a,b)) = \sqrt{(x - a)^2 + (y - b)^2}$$

Tên Khách Hàng	Tâm C1 (5,800)	Tâm C2 (2,300)	Cụm
Lâm Hồ Thiên Tổng (5,500)	300	200.0224987	2
Vũ Vương Vinh (5,800)	0	500.0089999	1
Nguyễn Kiều Nhã Linh (10,300)	500.0249994	8	2

**Sử dụng thuật toán K-Means để phân tích khách hàng dựa trên hành vi mua hàng
trên TiktokShop**

Đinh Như Ý (6,200)	600.0008333	100.079968	2
Lương Công Tiến (2,500)	300.0149996	200	2
Vũ Tường Nguyên (7,800)	2	500.0249994	1
Doãn Thị Đài Trang (2,300)	500.0089999	0	2
Nguyễn Phùng Vân Anh (15,900)	100.4987562	600.1408168	1
Nguyễn Thị Kim Kiều (2,200)	600.0075	100	2
Dương Thị Bích Muội (3,100)	700.0028571	200.0025	2
Nguyễn Thùy Trang (1,600)	200.039996	300.0016667	1
Nguyễn Ngọc Ánh (3,300)	500.004	1	2
Nguyễn Lâm Gia Thịnh (2,300)	500.0089999	0	2
Hồ Lâm Duy Khang (3,800)	2	500.001	1

**Sử dụng thuật toán K-Means để phân tích khách hàng dựa trên hành vi mua hàng
trên TiktokShop**

Hồ Lâm Gia Mỹ (4,700)	100.0049999	400.005	1
--------------------------	-------------	---------	---

Cụm quan sát gồm các quan sát:

- C1 = {Vũ Vương Vinh, Vũ Tường Nguyên, Nguyễn Phùng Vân Anh, Nguyễn Thùy Trang, Hồ Lâm Duy Khang, Hồ Lâm Gia Mỹ}
- C2 = {Lâm Hồ Thiên Tổng, Nguyễn Kiều Nhã Linh, Đinh Như Ý, Lương Công Tiến, Doãn Thị Đài Trang, Nguyễn Thị Kim Kiều, Dương Bích Muội, Nguyễn Ngọc Ánh, Nguyễn Lâm Gia Thịnh}

❖ **Bước 4: Tính lại tâm cụm bằng trung bình cộng các quan sát**

$$\text{Tâm C1} = \left(\frac{5+7+15+1+3+4}{6}, \frac{800+800+900+600+800+700}{6} \right) \\ = (5,83, 766,67)$$

$$\text{Tâm C2} = \left(\frac{5+10+6+2+2+2+3+3+2}{9}, \frac{500+300+200+500+300+200+100+300+300}{9} \right) \\ = (3,89, 300)$$

❖ **Bước 5:**

Lập lại vì:

	Lần khởi tạo	
Tâm cụm 1	(5,800)	(5,83, 766,67)
Tâm cụm 2	(2,300)	(3,89, 300)

❖ **Quay lại Bước 2:**

Chọn lại tâm cụm

- Tâm cụm 1 (C1) = (5,83, 766,67)

**Sử dụng thuật toán K-Means để phân tích khách hàng dựa trên hành vi mua hàng
trên TiktokShop**

- Tâm cụm 2 (C2) = (3,89, 300)

❖ **Bước 3:**

Tên Khách Hàng	Tâm C1 (5,83, 766,67)	Tâm C2 (3,89, 300)	Cụm
Lâm Hồ Thiên Tổng (5,500)	266.6679687	200.0030864	2
Vũ Vương Vinh (5,800)	33.34374837	500.0012346	1
Nguyễn Kiều Nhã Linh (10,300)	466.6852675	6.111111111	2
Đinh Như Ý (6,200)	566.6666912	100.0222815	2
Lương Công Tiến (2,500)	266.6942173	200.0089196	2
Vũ Tường Nguyên (7,800)	33.35374375	500.0096789	1
Doãn Thị Đài Trang (2,300)	466.6824104	1.888888889	2
Nguyễn Phùng Vân Anh (15,900)	133.648066	600.1028718	1
Nguyễn Thị Kim Kiều (2,200)	566.6796322	100.0178379	2

Sử dụng thuật toán K-Means để phân tích khách hàng dựa trên hành vi mua hàng trên TiktokShop

Dương Thị Bích Muội (3,100)	666.6726875	200.0019753	2
Nguyễn Thùy Trang (1,600)	166.7367353	300.0139091	1
Nguyễn Ngọc Ánh (3,300)	466.6752678	0.888888889	2
Nguyễn Lâm Gia Thịnh (2,300)	466.6824104	1.888888889	2
Hồ Lâm Duy Khang (3,800)	33.45353328	500.0007901	1
Hồ Lâm Gia Mỹ (4,700)	66.69187024	400.0000154	1

Cụm quan sát gồm các quan sát:

$C1 = \{ \text{Vũ Vương Vinh, Vũ Tường Nguyên, Nguyễn Phùng Vân Anh, Nguyễn Thùy Trang, Hồ Lâm Duy Khang, Hồ Lâm Gia Mỹ} \}$

$C2 = \{ \text{Lâm Hồ Thiên Tổng, Nguyễn Kiều Nhã Linh, Đinh Như Ý, Lương Công Tiến, Doãn Thị Đài Trang, Nguyễn Thị Kim Kiều, Dương Bích Muội, Nguyễn Ngọc Ánh, Nguyễn Lâm Gia Thịnh} \}$

❖ Bước 4: Tính lại tâm cụm bằng trung bình cộng các quan sát

$$\begin{aligned} \text{Tâm } C1 &= \left(\frac{5+7+15+1+3+4}{6}, \frac{800+800+900+600+800+700}{6} \right) \\ &= (5,83, 766,67) \end{aligned}$$

**Sử dụng thuật toán K-Means để phân tích khách hàng dựa trên hành vi mua hàng
trên TiktokShop**

$$\text{Tâm C2} = \left(\frac{5 + 10 + 6 + 2 + 2 + 2 + 3 + 3 + 2}{9}, \frac{500 + 300 + 200 + 500 + 300 + 200 + 100 + 300 + 300}{9} \right) \\ = (3,89, 300)$$

❖ **Bước 5:**

Không lặp lại vì:

	Lần khởi tạo	Lặp lần 1	
Tâm cụm 1	(5,800)	(5,83, 766,67)	(5,83, 766,67)
Tâm cụm 2	(2,300)	(3,89, 300)	(3,89, 300)

Vậy:

$C1 = \{\text{Vũ Vương Vinh, Vũ Tường Nguyên, Nguyễn Phùng Vân Anh, Nguyễn Thùy Trang, Hồ Lâm Duy Khang, Hồ Lâm Gia Mỹ}\}$

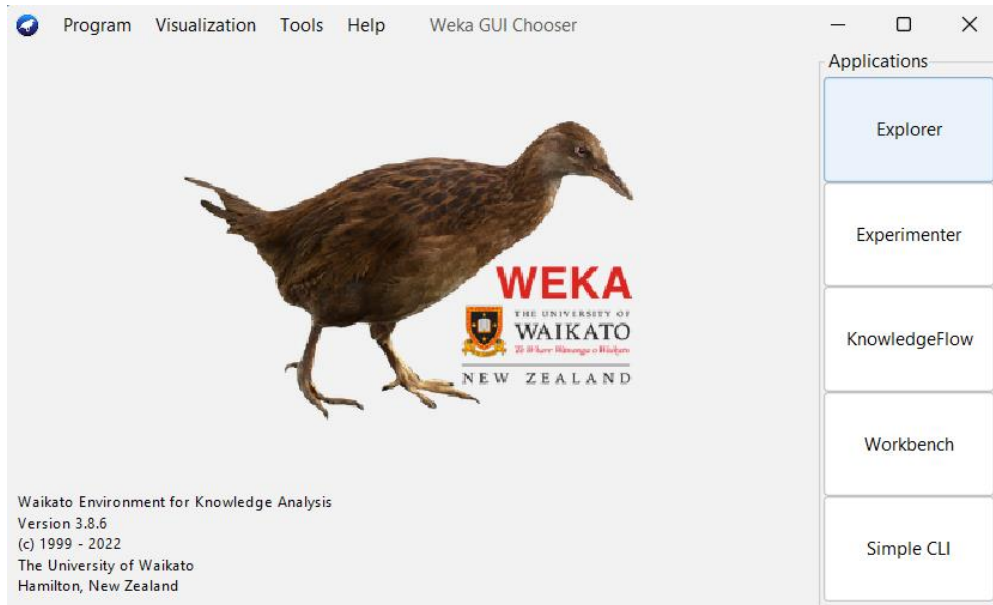
$C2 = \{\text{Lâm Hồ Thiên Tổng, Nguyễn Kiều Nhã Linh, Đinh Như Ý, Lương Công Tiến, Doãn Thị Đài Trang, Nguyễn Thị Kim Kiều, Dương Bích Muội, Nguyễn Ngọc Ánh, Nguyễn Lâm Gia Thịnh}\}$

CHƯƠNG 3. PHẦN MỀM KHAI PHÁ DỮ LIỆU MÃ NGUỒN MỞ

3.1 WEKA

3.1.1 Giới thiệu

WEKA là một phần mềm mã nguồn mở để thực hiện các tác vụ khai phá dữ liệu và học máy trên dữ liệu số. Tên gọi "WEKA" là viết tắt của "Waikato Environment for Knowledge Analysis", tức là môi trường phân tích tri thức của đại học Waikato ở New Zealand, nơi phần mềm này được phát triển.



Hình 3. 1: Giao diện phần mềm Weka

WEKA được phát triển bằng ngôn ngữ Java và cung cấp một bộ công cụ để phân tích dữ liệu và xây dựng mô hình học máy. WEKA cung cấp các công cụ tiền xử lý dữ liệu để làm sạch và chuyển đổi dữ liệu trước khi xử lý, các phương pháp phân tích dữ liệu như phân tích thành phần chính (PCA), phân tích độ tương tự (cluster), và phân tích giải thích (attribute selection), và các phương pháp học máy khác nhau bao gồm cây

Sử dụng thuật toán K-Means để phân tích khách hàng dựa trên hành vi mua hàng trên TiktokShop

quyết định, hồi quy tuyến tính, mạng neural, SVM và các phương pháp học tập có giám sát và không giám sát khác.

WEKA cũng cung cấp một giao diện đồ họa để dễ dàng sử dụng, cho phép người dùng tạo các luồng làm việc và thực hiện các tác vụ khai phá dữ liệu và học máy. WEKA cũng được tích hợp trong các công cụ phân tích dữ liệu và học máy khác như KNIME và RapidMiner.

Lịch sử phát triển:

- Vào những năm 1993, WEKA được phát triển bởi đại học Waikato ở New Zealand vào những năm 1993 bởi một nhóm các nhà khoa học máy tính và nghiên cứu viên. Ban đầu, WEKA được phát triển để giúp cho việc phân loại các loài chim dựa trên các thuộc tính sinh học. Sau đó, nó đã được mở rộng để sử dụng trong các tác vụ khai phá dữ liệu và học máy khác.
- Vào năm 1997, Phiên bản đầu tiên của WEKA được phát hành và nó đã trở thành một trong những phần mềm khai phá dữ liệu và học máy phổ biến nhất trên thế giới.
- Đến năm 2005, WEKA đã được tải xuống hơn 500.000 lần từ trang web chính thức của nó. Phần mềm Weka đã xuất sắc nhận được giải thưởng danh giá SIGKDD Data Mining and Knowledge Discovery Service Award.
- Đến nay, WEKA vẫn được phát triển và cập nhật liên tục, và nó đã trở thành một trong những công cụ phân tích dữ liệu và học máy được ưa chuộng nhất trên thế giới.

3.1.2 Chức năng

WEKA là một phần mềm khai phá dữ liệu và học máy mạnh mẽ, nó cung cấp các tính năng và công cụ để thực hiện các tác vụ khai phá dữ liệu và học máy trên các tập dữ liệu số.

Sử dụng thuật toán K-Means để phân tích khách hàng dựa trên hành vi mua hàng trên TiktokShop

Các chức năng chính của WEKA bao gồm:

- Tiền xử lý dữ liệu: WEKA cung cấp các công cụ để làm sạch và chuyển đổi dữ liệu trước khi xử lý, bao gồm xử lý dữ liệu bị thiếu, chuẩn hóa và biến đổi dữ liệu, lọc và lấy mẫu dữ liệu, và thêm hoặc xóa các thuộc tính.
- Phân tích dữ liệu: WEKA cung cấp các phương pháp phân tích dữ liệu như phân tích thành phần chính (PCA), phân tích độ tương tự (cluster), và phân tích giải thích (attribute selection).
- Học máy: WEKA cung cấp các phương pháp học máy khác nhau bao gồm cây quyết định, hồi quy tuyến tính, mạng neural, SVM và các phương pháp học tập có giám sát và không giám sát khác.
- Đánh giá mô hình: WEKA cung cấp các công cụ để đánh giá mô hình học máy, bao gồm các phương pháp đánh giá như độ chính xác, độ nhạy và độ đặc hiệu, và các kỹ thuật đánh giá mô hình như cross-validation và bootstrapping.
- Giao diện đồ họa: WEKA cung cấp giao diện đồ họa để dễ dàng sử dụng, cho phép người dùng tạo các luồng làm việc và thực hiện các tác vụ khai phá dữ liệu và học máy.
- Tích hợp với các công cụ khác: WEKA được tích hợp trong các công cụ phân tích dữ liệu và học máy khác như KNIME và RapidMiner.

Tóm lại, WEKA cung cấp một bộ công cụ mạnh mẽ để thực hiện các tác vụ khai phá dữ liệu và học máy trên các tập dữ liệu số, và nó được sử dụng rộng rãi trong nghiên cứu và ứng dụng thực tiễn.

3.1.3 Ưu điểm

WEKA là một phần mềm khai phá dữ liệu và học máy phổ biến, có nhiều ưu điểm nổi bật, bao gồm:

Sử dụng thuật toán K-Means để phân tích khách hàng dựa trên hành vi mua hàng trên TiktokShop

- Miễn phí và mã nguồn mở: WEKA là phần mềm miễn phí và mã nguồn mở, cho phép các nhà phát triển tùy chỉnh và mở rộng các tính năng của nó theo nhu cầu.
- Đa nền tảng: WEKA được hỗ trợ trên nhiều nền tảng, bao gồm Windows, Linux và macOS, vì vậy nó có thể được sử dụng trên hầu hết các hệ thống.
- Dễ sử dụng: WEKA cung cấp giao diện đồ họa dễ sử dụng và đơn giản, giúp người dùng thực hiện các tác vụ khai phá dữ liệu và học máy một cách dễ dàng.
- Các tính năng phong phú: WEKA cung cấp nhiều tính năng và công cụ để tiền xử lý dữ liệu, phân tích dữ liệu và học máy, bao gồm các phương pháp học máy phổ biến nhất.
- Tích hợp với các công cụ khác: WEKA có thể tích hợp với các công cụ phân tích dữ liệu và học máy khác, bao gồm KNIME và RapidMiner.
- Weka còn hỗ trợ cung cấp các quyền truy cập vào hệ thống cơ sở dữ liệu SQL bằng cách sử dụng Java Database Connectivity và nó có thể xử lý và kết quả sẽ được trả về bởi lệnh truy vấn cơ sở dữ liệu.
- Tài liệu phong phú: WEKA có nhiều tài liệu và hướng dẫn sử dụng, bao gồm các tài liệu trực tuyến, sách và bài giảng, giúp người dùng hiểu rõ các tính năng và thực hiện các tác vụ khai phá dữ liệu và học máy một cách hiệu quả.

Tóm lại, WEKA là một phần mềm khai phá dữ liệu và học máy đa năng, dễ sử dụng và miễn phí, với nhiều tính năng và tích hợp với các công cụ phân tích dữ liệu và học máy khác.

3.1.4 Nhược điểm

Mặc dù WEKA là một phần mềm khai phá dữ liệu và học máy mạnh mẽ, nhưng còn một số nhược điểm cần được cải thiện, bao gồm:

Sử dụng thuật toán K-Means để phân tích khách hàng dựa trên hành vi mua hàng trên TiktokShop

- Khả năng xử lý dữ liệu lớn: WEKA khó khăn trong việc xử lý các tập dữ liệu lớn và phức tạp. Điều này có thể khiến việc sử dụng WEKA trở nên chậm chạp và không hiệu quả.
- Khả năng tương thích với các công nghệ mới: WEKA được viết bằng Java và sử dụng các thư viện Java để thực hiện các tác vụ khai phá dữ liệu và học máy. Tuy nhiên, các công nghệ mới như Python và R đang trở nên phổ biến trong lĩnh vực này, do đó WEKA cần được cập nhật để có thể tương thích với các công nghệ mới này.
- Hỗ trợ tài liệu và người dùng: Mặc dù WEKA có một cộng đồng người dùng lớn, nhưng việc tìm kiếm thông tin và hỗ trợ có thể khó khăn đối với người dùng mới. WEKA cần cải thiện hỗ trợ tài liệu và hỗ trợ người dùng để giúp người dùng tận dụng tối đa các tính năng của phần mềm.

Tóm lại, những nhược điểm cần cải thiện của WEKA bao gồm khả năng xử lý dữ liệu lớn, khả năng tương thích với các công nghệ mới và hỗ trợ tài liệu và người dùng. Các cải tiến trong những lĩnh vực này có thể giúp WEKA trở nên tốt hơn và hữu ích hơn đối với người dùng.

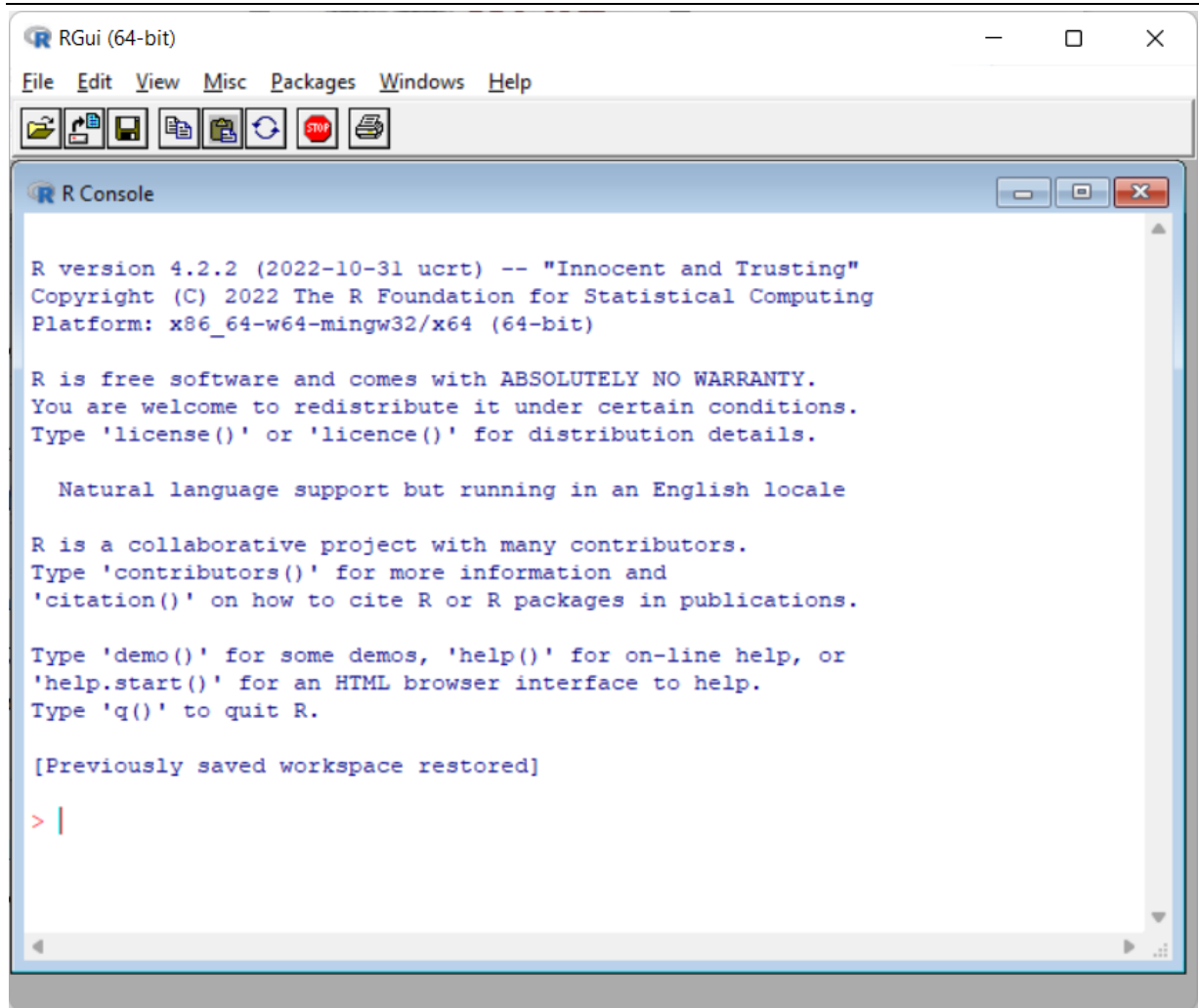
3.2 R

3.2.1 Giới thiệu

R là một ngôn ngữ lập trình và môi trường tính toán thống kê miễn phí và mã nguồn mở. Nó được sử dụng rộng rãi trong các lĩnh vực liên quan đến dữ liệu như khoa học dữ liệu, thống kê, tài chính, kinh tế, y tế và nhiều lĩnh vực khác. R cung cấp nhiều công cụ và thư viện cho phân tích dữ liệu, khai thác tri thức và mô hình hóa thống kê.

R được phát triển bởi Ross Ihaka và Robert Gentleman từ năm 1993 tại Đại học Auckland, New Zealand. Ngôn ngữ R được phát triển dựa trên ngôn ngữ S, một ngôn ngữ lập trình được sử dụng rộng rãi trong cộng đồng thống kê.

Sử dụng thuật toán K-Means để phân tích khách hàng dựa trên hành vi mua hàng trên TiktokShop



Hình 3. 2: Giao diện R

R có thể chạy trên nhiều nền tảng phần cứng như Intel, PowerPC, Alpha, Sparc và hỗ trợ nhiều hệ điều hành khác nhau như Unix, Linux, Windows và Mac OS X. R được phân phối theo giấy phép GNU GPL, cho phép người dùng sử dụng, sao chép, phân phối và sửa đổi mã nguồn của nó.

R có cú pháp đơn giản và dễ học, với nhiều cú pháp tương tự với các ngôn ngữ lập trình khác như C++, Python. R cung cấp nhiều gói và thư viện được phát triển bởi cộng đồng người dùng, giúp cho việc phân tích dữ liệu trở nên dễ dàng và nhanh chóng.

Sử dụng thuật toán K-Means để phân tích khách hàng dựa trên hành vi mua hàng trên TiktokShop

3.2.2 Chức năng

Các chức năng của R bao gồm:

- Phân tích dữ liệu: R cung cấp rất nhiều thư viện và gói phần mềm cho phân tích dữ liệu như dplyr, tidyr, reshape2, plyr ... giúp cho việc xử lý và phân tích dữ liệu trở nên dễ dàng hơn. Ví dụ: sử dụng thư viện dplyr để lọc và tóm tắt dữ liệu.
- Xử lý dữ liệu: R cung cấp các công cụ mạnh mẽ cho việc xử lý dữ liệu, bao gồm các gói phần mềm như stringr, lubridate, tidyr ... Ví dụ: sử dụng gói phần mềm stringr để xử lý dữ liệu chuỗi.
- Trực quan hóa dữ liệu: R có rất nhiều gói phần mềm để trực quan hóa dữ liệu như ggplot2, lattice, plotly, ggvis ... Giúp cho việc trình bày dữ liệu trở nên dễ dàng và hiệu quả hơn. Ví dụ: sử dụng ggplot2 để tạo biểu đồ.
- Phân tích thống kê: R cung cấp rất nhiều gói phần mềm cho phân tích thống kê như stats, car, MASS, moments ... Ví dụ: sử dụng gói phần mềm stats để tính toán thống kê mô tả.
- Machine learning: R là một trong những ngôn ngữ được sử dụng nhiều trong Machine Learning. R cung cấp các gói phần mềm cho Machine Learning như caret, mlr, randomForest ... Ví dụ: sử dụng gói phần mềm caret để huấn luyện mô hình dự báo.
- Dự báo: R là một trong những công cụ mạnh mẽ cho việc dự báo với các gói phần mềm như forecast, prophet ... Ví dụ: sử dụng gói phần mềm forecast để dự báo xu hướng giá cổ phiếu.
- Lập trình: Cho phép người dùng tạo ra các chương trình, hàm và gói phần mềm của riêng mình để sử dụng trong quá trình phân tích dữ liệu, xử lý dữ

Sử dụng thuật toán K-Means để phân tích khách hàng dựa trên hành vi mua hàng trên TiktokShop

liệu và trực quan hóa dữ liệu. Việc lập trình trên R cũng giúp người dùng tối ưu hóa quá trình xử lý dữ liệu và tăng tốc độ tính toán.

3.2.3 Ưu điểm

Các ưu điểm của ngôn ngữ R là:

- Miễn phí và mã nguồn mở: R là một phần mềm miễn phí, điều này rất hữu ích cho các nhà nghiên cứu, sinh viên và các nhà phân tích dữ liệu vì họ không phải chi trả chi phí cho việc sử dụng phần mềm. Vì là mã nguồn mở nên bất kỳ ai cũng có thể sử dụng và cải tiến nó, và có thể được dùng mọi lúc mọi nơi cho bất cứ việc gì, kể cả bán các sản phẩm từ R theo điều kiện của giấy phép.
- Cộng đồng lớn: R có một cộng đồng lớn và đầy đủ các tài liệu hướng dẫn, các gói phần mềm, các bài viết về các phương pháp phân tích dữ liệu. Điều này giúp cho người sử dụng R có thể dễ dàng tìm kiếm các thông tin, giải pháp cho các vấn đề mà mình đang gặp phải.
- Đa nền tảng: R có thể chạy trên nhiều nền tảng khác nhau, bao gồm cả Windows, Linux, MacOS và các hệ thống UNIX.
- Công cụ trực quan hóa dữ liệu: R cung cấp nhiều công cụ trực quan hóa dữ liệu cho phép người dùng có thể tạo ra các biểu đồ, đồ thị, bản đồ và các hình ảnh khác để mô tả dữ liệu một cách rõ ràng và trực quan.

3.2.4 Nhược điểm

Mặc dù có nhiều ưu điểm như đã nêu ở trên, tuy nhiên ngôn ngữ R cũng có một số nhược điểm như sau:

- Thiếu sự thống nhất trong các gói phần mềm: Vì R là một ngôn ngữ mã nguồn mở nên có rất nhiều gói phần mềm được phát triển bởi các cá nhân hoặc nhóm phát triển khác nhau, điều này dẫn đến sự thiếu sự thống nhất trong cách thức sử dụng và cú pháp.

Sử dụng thuật toán K-Means để phân tích khách hàng dựa trên hành vi mua hàng trên TiktokShop

- Tốc độ chạy chậm hơn so với một số ngôn ngữ khác như C++, Python, vì R là ngôn ngữ thông dịch và sử dụng đa số các thư viện phức tạp.
- Khó khăn trong việc xử lý dữ liệu lớn: R có thể gặp khó khăn trong việc xử lý các tập dữ liệu lớn vì bộ nhớ có hạn. Tuy nhiên, hiện nay đã có nhiều gói phần mềm hỗ trợ xử lý dữ liệu lớn như dplyr, data.table, ...
- Khó sử dụng cho một số lĩnh vực chuyên biệt: R phù hợp với phân tích thống kê, khoa học dữ liệu, nhưng không phải cho mọi lĩnh vực, ví dụ như phát triển ứng dụng web.
- Cú pháp phức tạp: Với những người mới bắt đầu học R, cú pháp của ngôn ngữ này có thể làm khó khăn trong quá trình học tập và phát triển ứng dụng.

Tuy nhiên, những nhược điểm trên cũng có thể được khắc phục bằng cách sử dụng các công cụ hỗ trợ và cải thiện kỹ năng lập trình của người dùng.

3.3 Python

3.3.1 Giới thiệu

Python là một ngôn ngữ lập trình thông dịch, có cú pháp đơn giản, dễ học và đọc. Nó được phát triển bởi Guido van Rossum vào năm 1989, được công bố vào năm 1991 và hiện nay đã trở thành một trong những ngôn ngữ lập trình phổ biến nhất trên thế giới.



Hình 3. 3: Python

Sử dụng thuật toán K-Means để phân tích khách hàng dựa trên hành vi mua hàng trên TiktokShop

Với cú pháp dễ đọc và hiểu, Python được sử dụng rộng rãi trong các lĩnh vực như khoa học dữ liệu, trí tuệ nhân tạo, web development, game development, và nhiều lĩnh vực khác.

Một số đặc điểm nổi bật của Python bao gồm:

- Cú pháp đơn giản, dễ đọc và dễ học.
- Hỗ trợ nhiều kiểu dữ liệu.
- Có rất nhiều thư viện mở rộng và phong phú để làm việc với nhiều tác vụ khác nhau.
- Được hỗ trợ bởi nhiều nền tảng và hệ điều hành, bao gồm Windows, macOS và Linux.
- Có cộng đồng phát triển mạnh mẽ và hỗ trợ đầy đủ tài liệu và ví dụ.

Python cũng là một trong những ngôn ngữ lập trình được sử dụng nhiều trong lĩnh vực khoa học dữ liệu và trí tuệ nhân tạo. Nhờ vào các thư viện như NumPy, Pandas, Scikit-learn và TensorFlow, Python đã trở thành công cụ không thể thiếu cho các nhà khoa học dữ liệu và nhà phát triển trí tuệ nhân tạo.

3.3.2 Chức năng

- Là một ngôn ngữ lập trình đa năng: Python có thể được sử dụng cho các mục đích khác nhau, bao gồm phát triển web, khoa học dữ liệu, máy học, trí tuệ nhân tạo, đồ họa máy tính, game, viễn thông, mã nguồn mở, và nhiều lĩnh vực khác.
- Cú pháp đơn giản: Python có cú pháp đơn giản và dễ đọc, dễ học, dễ sử dụng.
- Xử lý dữ liệu mạnh mẽ: Python có nhiều thư viện và công cụ hỗ trợ mạnh mẽ cho việc xử lý và phân tích dữ liệu.

Sử dụng thuật toán K-Means để phân tích khách hàng dựa trên hành vi mua hàng trên TiktokShop

- Hỗ trợ đa nền tảng: Python có thể chạy trên nhiều hệ điều hành khác nhau, bao gồm Windows, Linux, MacOS, và các hệ điều hành khác.
- Tính mở rộng cao: Python có thể được mở rộng thông qua việc sử dụng các thư viện và công cụ bên thứ ba.
- Cộng đồng sử dụng lớn: Python có cộng đồng sử dụng lớn, đông đảo, có nhiều tài liệu, hỗ trợ và giải đáp các thắc mắc của người dùng.

3.3.3 Ưu điểm

- Dễ học và sử dụng: Python có cú pháp đơn giản và trực quan, giúp người mới bắt đầu dễ dàng học và sử dụng. Nó cũng có nhiều tài liệu và hỗ trợ đầy đủ từ cộng đồng lập trình viên trên toàn thế giới.
- Đa nền tảng: Python có thể chạy trên nhiều nền tảng, bao gồm Windows, Linux và Mac OS. Điều này giúp người dùng có thể phát triển và triển khai ứng dụng trên nhiều nền tảng khác nhau.
- Đa năng: Python có thể được sử dụng trong nhiều lĩnh vực khác nhau, từ phân tích dữ liệu, trí tuệ nhân tạo, đến lập trình web, game, thiết kế đồ họa, và nhiều lĩnh vực khác.
- Thư viện phong phú: Python có một số lượng lớn các thư viện và module hỗ trợ cho nhiều ứng dụng khác nhau, từ khoa học dữ liệu đến lập trình web. Các thư viện phổ biến như NumPy, Pandas, Scikit-learn, TensorFlow,... cung cấp cho người dùng một lượng lớn các công cụ và chức năng để xử lý dữ liệu và phát triển ứng dụng.
- Tính mở rộng: Python là một ngôn ngữ lập trình có tính mở rộng cao, cho phép người dùng phát triển các module và thư viện mới để sử dụng cho ứng dụng của mình. Nó cũng có thể tích hợp với các ngôn ngữ khác như C và C++.

Sử dụng thuật toán K-Means để phân tích khách hàng dựa trên hành vi mua hàng trên TiktokShop

- Dễ dàng tương tác và thực thi: Python cung cấp các công cụ tương tác như trình thông dịch Python, Jupyter Notebook,... để người dùng có thể thực thi code và tương tác với kết quả trực tiếp.

3.3.4 Nhược điểm

- Tốc độ chậm hơn so với các ngôn ngữ lập trình khác: Python là một ngôn ngữ lập trình thông dịch, điều này làm cho tốc độ của Python chậm hơn so với các ngôn ngữ biên dịch như C++, Java.
- Cấu trúc dữ liệu không hiệu quả: Một số cấu trúc dữ liệu của Python như danh sách (list) và từ điển (dictionary) không hiệu quả về mặt tốc độ so với các ngôn ngữ lập trình khác.
- Bảo mật thấp: Python không được thiết kế để có tính bảo mật cao, nên việc sử dụng trong ứng dụng có tính bảo mật cao không được khuyến khích.
- Python có thể gặp vấn đề quản lý bộ nhớ khi dùng cho các ứng dụng nặng.

CHƯƠNG 4. KHAI PHÁ DỮ LIỆU

4.1 Xác định vấn đề

Một cửa hàng muốn bán được nhiều sản phẩm hơn cho khách thì đầu tiên họ phải hiểu rõ hơn về khách hàng của mình, từ đó mới có thể đưa ra các chiến lược kinh doanh và marketing phù hợp cho từng nhóm khách. Việc hiểu khách hàng còn giúp doanh nghiệp cải thiện chất lượng sản phẩm và dịch vụ của mình để đáp ứng nhu cầu của khách hàng và tăng sự hài lòng của khách hàng.

Do đó họ cần giải quyết những câu hỏi sau:

- Các khách hàng đang có hành vi mua hàng tương tự nhau trên Tiktokshop thường có những đặc điểm và nhu cầu gì?
- Làm thế nào để phát hiện ra những xu hướng mua hàng của khách hàng trên Tiktokshop để có thể đưa ra các chiến lược tiếp thị và bán hàng hiệu quả hơn?
- Khách hàng thường chi bao nhiêu tiền cho mỗi lần mua hàng?
- Các sản phẩm nào đang được khách hàng quan tâm và mua nhiều nhất trên Tiktokshop?

Để giải quyết các câu hỏi trên thì ta cần tiến hành thu thập dữ liệu về hành vi mua hàng của khách hàng trên TiktokShop:

Dữ liệu thu thập gồm có các thuộc tính: Họ và tên, Giới tính, Khu vực, Tần suất mua hàng, Số lượng sản phẩm mỗi đơn hàng, Loại sản phẩm quan tâm, Giá trị đơn hàng, Thời gian đặt hàng.

Bằng cách tập trung vào hành vi mua hàng của khách từ những dữ liệu trên, ta có thể đưa ra các quyết định về các chiến lược dựa trên từng nhóm khách có điểm chung, qua đó cải thiện tình hình kinh doanh của mình, đồng thời nâng cao sự cạnh tranh trong thị trường.

Sử dụng thuật toán K-Means để phân tích khách hàng dựa trên hành vi mua hàng trên TiktokShop

4.2 Hiểu dữ liệu

Tập dữ liệu thu thập được, gồm có các thuộc tính sau: Họ và tên, Giới tính, Khu vực, Tần suất mua hàng, Số lượng sản phẩm mỗi đơn hàng, Loại sản phẩm quan tâm, Giá trị đơn hàng, Thời gian đặt hàng.

Với tập dữ liệu trên, hoàn toàn có thể giải quyết vấn đề kinh doanh, đặc biệt là trong việc tối ưu hóa chiến lược kinh doanh và marketing đối với từng nhóm khách hàng khác nhau:

- + Thời gian mua hàng và tần số mua hàng: hai thuộc tính này thể hiện thói quen mua sắm của khách hàng.
- + Tần số mua hàng và giá trị đơn hàng: hai thuộc tính này thể hiện mức độ mua sắm của khách hàng và giá trị đơn hàng của họ.
- + Loại sản phẩm quan tâm và giá trị đơn hàng: hai thuộc tính này thể hiện sở thích và mức độ chi tiêu của khách hàng.
- + Số lượng sản phẩm mỗi đơn và giá trị đơn hàng: hai thuộc tính này thể hiện mức độ mua sắm của khách hàng và giá trị đơn hàng của họ.

Mô tả cấu trúc của dữ liệu:

Dữ liệu gồm có 8 thuộc tính.

Số lượng instance trong tập dữ liệu là: 100.

Bảng 3. 1: Bảng mô tả cấu trúc của dữ liệu

Tên thuộc tính	Kiểu dữ liệu	Ý nghĩa của thuộc tính	Ghi chú thêm
Họ và tên	Character	Họ và tên khách hàng.	
Giới tính	Character	Giới tính khách hàng.	

Sử dụng thuật toán K-Means để phân tích khách hàng dựa trên hành vi mua hàng trên TiktokShop

Khu vực	Character	Khu vực khách hàng đang ở.	Chỉ bao gồm 3 miền: Miền Bắc, Miền Trung, Miền Nam.
Tần suất mua hàng	Integer	Cho biết trong 1 tháng, khách hàng mua trung bình bao nhiêu lần.	Chỉ tính số lần mua trên TiktokShop. Không bao gồm các sàn thương mại điện tử khác.
Số lượng sản phẩm mỗi đơn	Integer	Số lượng sản phẩm trung bình trong mỗi đơn hàng.	
Loại sản phẩm quan tâm	Character	Những loại sản phẩm mà khách hàng quan tâm đến	Gồm có 7 loại sản phẩm: <ul style="list-style-type: none"> - Thời trang - Nhà cửa & Đời sống - Gia dụng - Mỹ phẩm - Sức khỏe - Trang sức và phụ kiện - Chăm sóc nhà cửa
Giá trị đơn hàng	Integer	Mức giá của một đơn hàng.	Mức giá này là mức giá trung bình các đơn hàng của khách hàng. Đơn giá: Nghìn đồng.
Thời gian đặt hàng	Character	Khoảng thời gian khách hàng thường đặt hàng.	<ul style="list-style-type: none"> - Sáng - Trưa - Chiều - Tối

Sử dụng thuật toán K-Means để phân tích khách hàng dựa trên hành vi mua hàng trên TiktokShop

4.3 Chuẩn bị dữ liệu

- ❖ **Thu thập dữ liệu:** Tạo khảo sát về hành vi mua hàng trên TiktokShop của 100 khách hàng khác nhau.

Tập dữ liệu khảo sát thu về được như sau:

	A	B	C	D	E	F	G	H
1	Họ tên	Giới tính	Khu vực	Tần suất	Số lượng	Loại sản phẩm quan tâm	Giá trị đơn hàng	Thời gian
2	Lâm Hồ Thiên Tống	Nam	Miền Nam	5	3	Thời trang	500000	Tối
3	Vũ Vương Vinh	Nam	Miền Nam	10	5	Thời trang	1000000	Tối
4	Nguyễn Kiều Nhã Linh	Nữ	Miền Nam	10	2	Thời trang	300000	Tối
5	Đinh Như Ý	Nữ	Miền Nam	2	1	Thời trang	100000	Tối
6	Lương Công Tiến	Nam	Miền Nam	6	5	Thời trang	900000	Tối
7	Vũ Tường Nguyên	Nam	Miền Nam	1	1	Thời trang	400000	Tối
8	Nguyễn Sĩ Nguyên	Nam	Miền Nam	1	1	Thời trang	600000	Sáng
9	Doãn Thị Cẩm Hương	Nữ	Miền Nam	7	2	Thời trang	500000	Chiều
10	Vương Triều Lê	Nam	Miền Nam	1	1	Thời trang	700000	Sáng
11	Lê Thị Kim Giang	Nữ	Miền Bắc	1	1	Nhà cửa & Đời sống	600000	Tối
12	Trần Văn Trường	Nam	Miền Bắc	1	1	Nhà cửa & Đời sống	200000	Sáng
13	To Khánh Huyền	Nữ	Miền Bắc	15	1	Nhà cửa & Đời sống	200000	Tối
14	Bùi Hồng Diệp	Nữ	Miền Bắc	2	2	Gia dụng	500000	Sáng
15	Võ Thành Thế	Nam	Miền Bắc	3	1	Gia dụng	800000	Tối
16	Nguyễn Kim Thủy	Nữ	Miền Bắc	1	1	Gia dụng	700000	Trưa
17	Nguyễn Ngọc Hoàng Duyên	Nam	Miền Bắc	1	1	Gia dụng	500000	Tối
18	Doãn Thị Đài Trang	Nữ	Miền Bắc	5	1	Nhà cửa & Đời sống	200000	Tối
19	Nguyễn Phùng Văn Anh	Nữ	Miền Bắc	1	3	Nhà cửa & Đời sống	100000	Tối
20	Nguyễn Thị Kim Kiều	Nữ	Miền Bắc	8	1	Nhà cửa & Đời sống	600000	Tối
21	Dương Thị Bích Muội	Nữ	Miền Trung	3	4	Thời trang	500000	Trưa
22	Nguyễn Thùy Trang	Nữ	Miền Nam	10	1	Thời trang	100000	Tối
23	Nguyễn Ngọc Ánh	Nữ	Miền Nam	1	1	Thời trang	200000	Tối
24	Nguyễn Lâm Gia Thịnh	Nam	Miền Bắc	3	1	Thời trang	100000	Tối
25	Hồ Lâm Duy Khang	Nam	Miền Nam	10	4	Thời trang	300000	Tối
26	Nguyễn Thị Phúc	Nữ	Miền Nam	8	1	Thời trang	900000	Tối
27	Hồ Lâm Gia Mỹ	Nữ	Miền Bắc	1	1	Thời trang	400000	Tối
28	Trần Kim Tuyết	Nữ	Miền Nam	3	5	Thời trang	1000000	Tối
29	Lý Huỳnh My	Nữ	Miền Bắc	4	1	Gia dụng	600000	Chiều
30	Đào Nga My	Nữ	Miền Bắc	1	4	Gia dụng	500000	Chiều
31	Nguyễn Thanh Nhân	Nam	Miền Bắc	10	5	Gia dụng	1000000	Tối
32	Trần Ngọc Như Hoa	Nữ	Miền Bắc	1	1	Gia dụng	200000	Tối
33	Đinh Thị Xuân Hoa	Nữ	Miền Bắc	4	1	Gia dụng	200000	Tối
34	Trần Thị Như Ý	Nữ	Miền Bắc	4	1	Nhà cửa & Đời sống	200000	Tối
35	Nguyễn Cẩm Vy	Nữ	Miền Bắc	8	2	Nhà cửa & Đời sống	400000	Tối
36	Thới Việt Trà	Nữ	Miền Bắc	6	1	Nhà cửa & Đời sống	200000	Tối
37	Lê Thị Anh Thư	Nữ	Miền Bắc	4	3	Thời trang	800000	Tối
38	Nguyễn Duy Khánh	Nam	Miền Bắc	1	1	Thời trang	300000	Tối
39	Trần Ý Vy	Nữ	Miền Nam	1	4	Thời trang	500000	Chiều
40	Trần Thị Uyên Vy	Nữ	Miền Trung	8	2	Thời trang	900000	Trưa
41	Huỳnh Thị Mơ Muội	Nữ	Miền Trung	8	4	Thời trang	300000	Tối
42	Vũ Đức Lộc	Nam	Miền Nam	7	4	Thời trang	400000	Tối
43	Đỗ Kiều Duy	Nữ	Miền Nam	8	2	Thời trang	300000	Tối
44	Tô Văn Duy Đang	Nam	Miền Nam	4	2	Thời trang	400000	Tối
45	Nguyễn Trí Hào	Nam	Miền Bắc	4	1	Thời trang	600000	Sáng
46	Lê Hoài Thương	Nam	Miền Nam	9	1	Mỹ phẩm	400000	Sáng
47	Đoàn Thị Kim Anh	Nữ	Miền Nam	10	3	Mỹ phẩm	400000	Trưa
48	Nguyễn Nhật Linh	Nữ	Miền Nam	3	1	Thời trang	300000	Trưa
49	Tô Khánh Văn	Nam	Miền Nam	4	1	Thời trang	100000	Chiều
50	Doãn Chí Bình	Nam	Miền Nam	7	2	Thời trang	400000	Trưa

Sử dụng thuật toán K-Means để phân tích khách hàng dựa trên hành vi mua hàng trên TiktokShop

51	Trần Hữu Lộc	Nam	Miền Nam	8	1	Thời trang	300000	Trưa
52	Lâm Nhất Đức Duy	Nam	Miền Nam	1	1	Thời trang	200000	Sáng
53	Trần Bích Ngọc	Nữ	Miền Nam	8	4	Mỹ phẩm	800000	Chiều
54	Nguyễn Vũ Luân	Nam	Miền Nam	4	1	Mỹ phẩm	200000	Tối
55	Nguyễn Hoàng Tuấn	Nam	Miền Nam	4	1	Thời trang	100000	Tối
56	Lâm Vĩnh Long	Nam	Miền Nam	3	4	Thời trang	300000	Tối
57	Dương Minh Nhí	Nam	Miền Nam	2	4	Mỹ phẩm	400000	Chiều
58	Hồ Phương Anh	Nữ	Miền Trung	10	5	Sức khỏe	800000	Trưa
59	Nguyễn Thị Ngọc Yến	Nữ	Miền Trung	1	1	Sức khỏe	600000	Sáng
60	Trần Kim Như	Nữ	Miền Nam	2	1	Thời trang	700000	Tối
61	Nguyễn Thảo Nguyên	Nữ	Miền Nam	6	4	Mỹ phẩm	100000	Tối
62	Lâm Bích Du	Nữ	Miền Nam	8	3	Mỹ phẩm	800000	Tối
63	Nguyễn Thúy Vi	Nữ	Miền Nam	4	3	Mỹ phẩm	300000	Chiều
64	Hoàng Văn Hoài	Nam	Miền Trung	3	3	Sức khỏe	500000	Trưa
65	Hoàng Triệu Vy	Nữ	Miền Nam	6	2	Trang sức và phụ kiện	300000	Sáng
66	Lâm Mẫn Vy	Nữ	Miền Nam	3	2	Trang sức và phụ kiện	400000	Sáng
67	Hồ Khánh Vân	Nữ	Miền Trung	3	1	Sức khỏe	200000	Chiều
68	Trần Kim Chi	Nữ	Miền Trung	4	4	Sức khỏe	500000	Trưa
69	Trần Thanh Tâm	Nữ	Miền Trung	8	2	Sức khỏe	200000	Tối
70	Trần Gia Khang	Nam	Miền Trung	3	3	Sức khỏe	400000	Tối
71	Đặng Văn Bi	Nam	Miền Trung	5	5	Sức khỏe	500000	Chiều
72	Ngô Khánh Thái	Nam	Miền Trung	3	4	Sức khỏe	500000	Tối
73	Trần Hoàn Hào	Nam	Miền Nam	1	1	Trang sức và phụ kiện	100000	Sáng
74	Trương Đình	Nam	Miền Nam	3	1	Trang sức và phụ kiện	200000	Chiều
75	Lâm Nhất	Nam	Miền Trung	5	3	Sức khỏe	300000	Chiều
76	Nguyễn Ngọc Bích	Nữ	Miền Nam	1	2	Trang sức và phụ kiện	200000	Trưa
77	Trần Bích Ngọc	Nữ	Miền Nam	2	2	Trang sức và phụ kiện	300000	Tối
78	Nguyễn Trí Hữu	Nam	Miền Trung	2	4	Sức khỏe	400000	Chiều
79	Nguyễn Ngọc Hòa	Nam	Miền Bắc	4	4	Nhà cửa & Đời sống	400000	Tối
80	Lâm Nhất Thiên	Nam	Miền Bắc	4	2	Gia dụng	300000	Trưa
81	Trần Công Bằng	Nam	Miền Trung	3	4	Sức khỏe	300000	Tối
82	Lương Chí Hải	Nam	Miền Trung	2	3	Sức khỏe	400000	Sáng
83	Tạ Công Bằng	Nam	Miền Trung	2	2	Sức khỏe	400000	Trưa
84	Lâm Chí Khanh	Nam	Miền Trung	8	2	Sức khỏe	200000	Chiều
85	Đặng Văn Lâm	Nam	Miền Trung	4	5	Sức khỏe	700000	Tối
86	Đặng Thủy Trâm	Nữ	Miền Trung	10	5	Sức khỏe	1000000	Tối
87	Trần Minh Khánh	Nam	Miền Trung	8	4	Sức khỏe	800000	Sáng
88	Đặng Văn Hòa	Nam	Miền Trung	7	2	Sức khỏe	300000	Trưa
89	Đinh Minh Khang	Nam	Miền Trung	4	2	Sức khỏe	200000	Tối
90	Vũ Khánh Nguyễn	Nam	Miền Bắc	4	4	Gia dụng	400000	Chiều
91	Doãn Ninh Bình	Nam	Miền Bắc	3	3	Chăm sóc nhà cửa	400000	Tối
92	Đoàn Ngọc Hải	Nam	Miền Bắc	5	2	Chăm sóc nhà cửa	400000	Trưa
93	Trần Thị Mỹ Ái	Nữ	Miền Trung	4	5	Chăm sóc nhà cửa	700000	Chiều
94	Hoàng Tiểu Bào	Nam	Miền Trung	4	3	Chăm sóc nhà cửa	400000	Tối
95	Trần Quốc Khánh	Nam	Miền Trung	2	2	Chăm sóc nhà cửa	300000	Sáng
96	Trần Quốc Tuấn	Nam	Miền Bắc	3	3	Chăm sóc nhà cửa	300000	Sáng
97	Đinh Minh Hoàng	Nam	Miền Bắc	7	4	Chăm sóc nhà cửa	400000	Chiều
98	Đoàn Khánh Quốc	Nam	Miền Bắc	4	2	Thời trang	300000	Trưa
99	Đoàn Ngọc Khánh Linh	Nữ	Miền Bắc	3	3	Gia dụng	500000	Trưa
100	Nguyễn Khánh Hà	Nữ	Miền Trung	4	3	Sức khỏe	400000	Tối
101	Lưu Bảo Hà	Nữ	Miền Bắc	9	2	Thời trang	300000	Tối

Hình 4. 1: Danh sách dữ liệu khảo sát về 100 khách hàng

Sử dụng thuật toán K-Means để phân tích khách hàng dựa trên hành vi mua hàng trên TiktokShop

❖ Giai đoạn Tiền xử lý dữ liệu:

- Làm sạch dữ liệu:

Dữ liệu mẫu đã được xử lý và kiểm tra kỹ lưỡng trước khi thu thập, do đó dữ liệu đã khá sạch và có thể sử dụng trực tiếp cho mục đích khai phá dữ liệu mà không cần phải làm sạch thêm.

- Chọn lọc dữ liệu:

Loại bỏ thuộc tính “Họ và Tên khách hàng”. Do mục đích của dự án là tìm ra nhóm khách hàng dựa trên hành vi mua hàng và không quan tâm đến thông tin cá nhân của từng khách hàng.

- Chuyển đổi dữ liệu:

- + Chuyển đổi các thuộc tính “Giới Tính”, “Khu Vực”, “Loại sản phẩm quan tâm” và “Thời Gian” của tập dữ liệu, từ dạng chữ sang số để phù hợp với thuật toán khai phá dữ liệu. Quy ước như hình sau:

Giới tính		
Nam = 1		Nữ = 2
Khu vực		
Miền Bắc = 1	Miền Trung = 2	Miền Nam = 3
Thời gian		
Sáng = 1		Trưa = 2
Chiều = 3		Tối = 4
Loại sản phẩm quan tâm		
Thời trang = 1		Nhà cửa & Đời sống = 2
Gia dụng = 3		Mỹ phẩm = 4
Sức khỏe = 5		Chăm sóc nhà cửa = 6
Trang sức và phụ kiện = 7		

Hình 4. 2: Quy ước chuyển đổi các thuộc tính từ dạng chữ sang dạng số

- + Ngoài ra đối với thuộc tính “Giá trị đơn hàng” (đơn vị đang là “nghìn đồng”), ta có thể cân nhắc xem xét chuyển chúng sang đơn vị là “trăm nghìn đồng”. Việc chuyển đơn vị này, giúp giảm đáng kể độ lớn của dữ liệu.

Sử dụng thuật toán K-Means để phân tích khách hàng dựa trên hành vi mua hàng trên TiktokShop

- Tạo mẫu dữ liệu:

Tách tập dữ liệu thành các cặp thuộc tính sẽ giúp tìm ra các đặc trưng cơ bản của dữ liệu và làm giảm độ phức tạp của bài toán. Tuy nhiên, ta sẽ giữ nguyên các dữ liệu của khách, chỉ là tách cách thuộc tính ra theo từng nhu cầu khác nhau:

- + Đối với nhu cầu cần phân tích thói mua sắm của khách hàng ta sẽ sử dụng cặp thuộc tính “Thời gian mua hàng” và “Tần số mua hàng”.
- + Đối với nhu cầu cần phân tích về mức độ mua sắm và giá trị đơn hàng của khách, ta sẽ sử dụng bộ ba thuộc tính “Tần số mua hàng”, “Số lượng sản phẩm mỗi đơn”, và “Giá trị đơn hàng”.
- + Đối với nhu cầu cần phân tích về sở thích nào được chi tiêu nhiều nhất thì sẽ dùng cặp thuộc tính “Loại sản phẩm quan tâm” và “Giá trị đơn hàng”.

❖ Kết quả của bước Tiền xử lý dữ liệu

Đây là dữ liệu sau giai đoạn tiền xử lý dữ liệu:

- Thuộc tính “Họ và Tên khách hàng” đã được loại bỏ.
- Các thuộc tính như “Giới Tính”, “Khu vực”, “Loại sản phẩm quan tâm”, “Thời Gian” đã được chuyển đổi từ chữ sang số đúng theo quy ước ở trên.
- Thuộc tính “Giá trị mỗi đơn” đã được chuyển từ đơn vị “Nghìn đồng” sang “Trăm Nghìn” để giảm độ lớn của dữ liệu.

Sử dụng thuật toán K-Means để phân tích khách hàng dựa trên hành vi mua hàng trên TiktokShop

	A	B	C	D	E	F	G
1	Giới tính ▾	Khu vực ▾	Tần suất mua hàng (tháng) ▾	Số lượng sản phẩm mỗi đơn ▾	Loại sản phẩm quan tâm ▾	Giá trị mỗi đơn (Trăm Nghìn) ▾	Thời gian ▾
2	1	3	5	3	1	5	4
3	1	3	10	5	1	10	4
4	2	3	10	2	1	3	4
5	2	3	2	1	1	1	4
6	1	3	6	5	1	9	4
7	1	3	1	1	1	4	4
8	1	3	1	1	1	6	1
9	2	3	7	2	1	5	3
10	1	3	1	1	1	7	1
11	2	1	1	1	2	6	4
12	1	1	1	1	2	2	1
13	2	1	15	1	2	2	4
14	2	1	2	2	3	5	1
15	1	1	3	1	3	8	4
16	2	1	1	1	3	7	2
17	1	1	1	1	3	5	4
18	A	B	C	D	E	F	G
18	2	1	5	1	2	2	4
19	2	1	1	3	2	1	4
20	2	1	8	1	2	6	4
21	2	2	3	4	1	5	2
22	2	3	10	1	1	1	4
23	2	3	1	1	1	2	4
24	1	1	3	1	1	1	4
25	1	3	10	4	1	3	4
26	2	3	8	1	1	9	4
27	2	1	1	1	1	4	4
28	2	3	3	5	1	10	4
29	2	1	4	1	3	6	3
30	2	1	1	4	3	5	3
31	1	1	10	5	3	10	4
32	2	1	1	1	3	2	4
33	2	1	4	1	3	2	4
34	2	1	4	1	2	2	4
35	2	1	8	2	2	4	4
36	2	1	6	1	2	2	4
37	2	1	4	3	1	8	4
38	1	1	1	1	1	3	4
39	2	3	1	4	1	5	3
40	2	2	8	2	1	9	2
41	2	2	8	4	1	3	4
42	1	3	7	4	1	4	4
43	A	B	C	D	E	F	G
43	2	3	8	2	1	3	4
44	1	3	4	2	1	4	4
45	1	1	4	1	1	6	1
46	1	3	9	1	4	4	1
47	2	3	10	3	4	4	2
48	2	3	3	1	1	3	2
49	1	3	4	1	1	1	3
50	1	3	7	2	1	4	2
51	1	3	8	1	1	3	2
52	1	3	1	1	1	2	1
53	2	3	8	4	4	8	3
54	1	3	4	1	4	2	4
55	1	3	4	1	1	1	4
56	1	3	3	4	1	3	4
57	1	3	2	4	4	4	3
58	2	2	10	5	5	8	2
59	2	2	1	1	5	6	1
60	2	3	2	1	1	7	4
61	2	3	6	4	4	1	4
62	2	3	8	3	4	8	4
63	2	3	4	3	4	3	3
64	1	2	3	3	5	5	2
65	2	3	6	2	7	3	1
66	2	3	3	2	7	4	1
67	2	2	3	1	5	2	3

Sử dụng thuật toán K-Means để phân tích khách hàng dựa trên hành vi mua hàng trên TiktokShop

	A	B	C	D	E	F	G
68	2	2	4	4	5	5	2
69	2	2	8	2	5	2	4
70	1	2	3	3	5	4	4
71	1	2	5	5	5	5	3
72	1	2	3	4	5	5	4
73	1	3	1	1	7	1	1
74	1	3	3	1	7	2	3
75	1	2	5	3	5	3	3
76	2	3	1	2	7	2	2
77	2	3	2	2	7	3	4
78	1	2	2	4	5	4	3
79	1	1	4	4	2	4	4
80	1	1	4	2	3	3	2
81	1	2	3	4	5	3	4
82	1	2	2	3	5	4	1
83	1	2	2	2	5	4	2
84	1	2	8	2	5	2	3
85	1	2	4	5	5	7	4
86	2	2	10	5	5	10	4
87	1	2	8	4	5	8	1
88	1	2	7	2	5	3	2
89	1	2	4	2	5	2	4
90	1	1	4	4	3	4	3

	A	B	C	D	E	F	G
91	1	1	3	3	6	4	4
92	1	1	5	2	6	4	2
93	2	2	4	5	6	7	3
94	1	2	4	3	6	4	4
95	1	2	2	2	6	3	1
96	1	1	3	3	6	3	1
97	1	1	7	4	6	4	3
98	1	1	4	2	1	3	2
99	2	1	3	3	3	5	2
100	2	2	4	3	5	4	4
101	2	1	9	2	1	3	4

Hình 4. 3: Dữ liệu sau khi tiền xử lý dữ liệu

Dưới đây là dữ liệu được tách ra thành các cặp thuộc tính. để dễ dàng xem xét và phân tích (Chỉ chụp một phần dữ liệu, tuy nhiên được trích ra hoàn toàn từ dữ liệu gốc):

- “Tần số mua hàng” và “Thời gian”.

	A	B
1	Tần suất mua hàng	Thời gian
2	5	4
3	10	4
4	10	4
5	2	4
6	6	4
7	1	4
8	1	1
9	7	3
10	1	1
11	1	4
12	1	1
13	15	4
14	2	1
15	3	4
16	1	2
17	1	4
18	5	4
19	1	4
20	8	4

Hình 4. 4: Tách thuộc tính “Tần suất mua hàng” và “Thời gian”

Sử dụng thuật toán K-Means để phân tích khách hàng dựa trên hành vi mua hàng trên TiktokShop

- “Tần số mua hàng”, “Số lượng sản phẩm mỗi đơn”, và “Giá trị đơn hàng”.

	A	B	C
1	Tần suất mua hàng	Số lượng sản phẩm	Giá trị đơn hàng
2	5	3	5
3	10	5	10
4	10	2	3
5	2	1	1
6	6	5	9
7	1	1	4
8	1	1	6
9	7	2	5
10	1	1	7
11	1	1	6
12	1	1	2
13	15	1	2
14	2	2	5
15	3	1	8
16	1	1	7
17	1	1	5
18	5	1	2
19	1	3	1
20	0	1	6

Hình 4. 5: Tách thuộc tính “Tần suất mua hàng”, “Số lượng sản phẩm” và “Giá trị đơn hàng”

- “Loại sản phẩm quan tâm” và “Giá trị đơn hàng”.

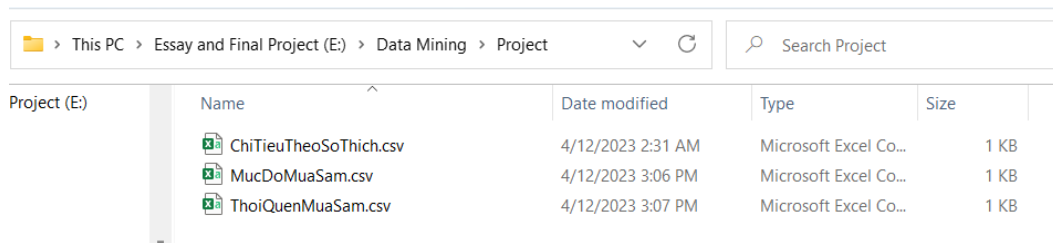
	A	B
1	Loại sản phẩm quan tâm	Giá trị đơn hàng
2	1	5
3	1	10
4	1	3
5	1	1
6	1	9
7	1	4
8	1	6
9	1	5
10	1	7
11	2	6
12	2	2
13	2	2
14	3	5
15	3	8
16	3	7
17	3	5
18	2	2
19	2	1
20	2	6
21	1	5

Hình 4. 6: Tách thuộc tính “Loại sản phẩm quan tâm” và “Giá trị đơn hàng”

Sử dụng thuật toán K-Means để phân tích khách hàng dựa trên hành vi mua hàng trên TiktokShop

Cả 3 dữ liệu này đều được lưu lại dưới dạng file csv. Tuy nhiên, có một số thay đổi là tên thuộc tính được bỏ dấu Tiếng Việt, để phù hợp với file csv. Và được đặt tên như sau:

- + Tên của file dữ liệu gồm có hai thuộc tính “Tần suất mua hàng” và “Thời gian” là: “ThoiQuenMuaSam.csv” để chỉ thói quen mua sắm của khách hàng.
- + Tên của file dữ liệu gồm có ba thuộc tính “Tần suất mua hàng”, “Số lượng sản phẩm” và “Giá trị đơn hàng” là: “MucDoMuaSam.csv” để chỉ mức độ mua sắm của khách hàng.
- + Tên của file dữ liệu gồm có hai thuộc tính “Loại sản phẩm quan tâm” và “Giá trị đơn hàng” là: “ChiTieuTheoSoThich.csv” để chỉ việc mua sắm của khách hàng dựa trên sở thích, loại sản phẩm quan tâm và độ chi tiêu.



Project (E:)	Name	Date modified	Type	Size
	ChiTieuTheoSoThich.csv	4/12/2023 2:31 AM	Microsoft Excel Co...	1 KB
	MucDoMuaSam.csv	4/12/2023 3:06 PM	Microsoft Excel Co...	1 KB
	ThoiQuenMuaSam.csv	4/12/2023 3:07 PM	Microsoft Excel Co...	1 KB

Hình 4. 7: Các thuộc tính được trích từ tập dữ liệu khảo sát thành các dữ liệu nhỏ

4.4 Lập mô hình và chạy thuật toán K-means trên Weka

Chọn phương pháp khai phá: Chọn phương pháp phân cụm (Clustering). Vì:

- Tập dữ liệu này là dữ liệu không được gán nhãn. Vì vậy, phương pháp Clustering có thể được sử dụng để phân nhóm các khách hàng dựa trên các đặc trưng mua hàng của họ một cách tự động.
- Clustering là phương pháp phổ biến trong khai phá dữ liệu, cho phép phân nhóm các mẫu dữ liệu tương tự với nhau vào các cụm riêng biệt. Điều này có thể giúp tìm ra các đặc trưng phổ biến của mỗi nhóm khách hàng, từ đó giúp

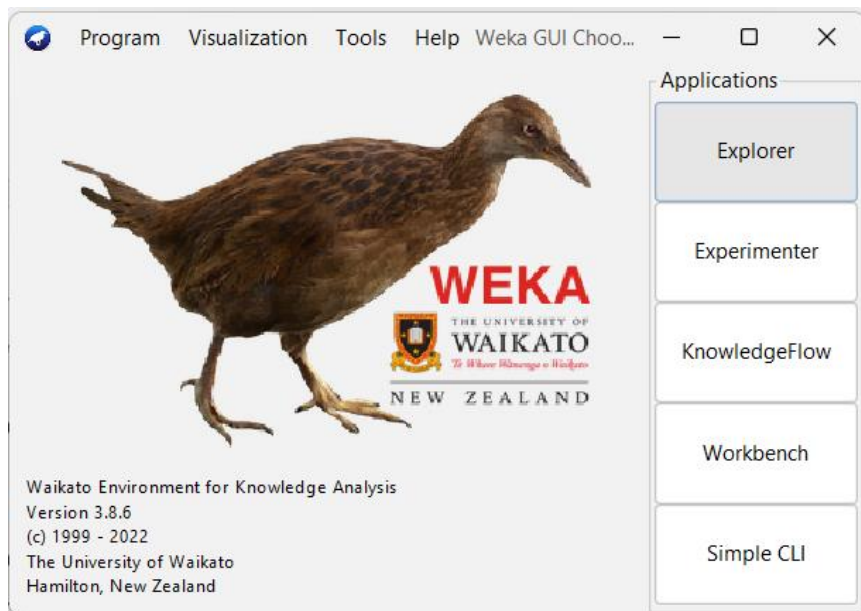
Sử dụng thuật toán K-Means để phân tích khách hàng dựa trên hành vi mua hàng trên TiktokShop

chúng ta hiểu hơn về hành vi mua hàng của từng nhóm khách hàng và thiết kế chiến lược kinh doanh phù hợp.

Chọn thuật toán: Sử dụng thuật toán Kmeans.

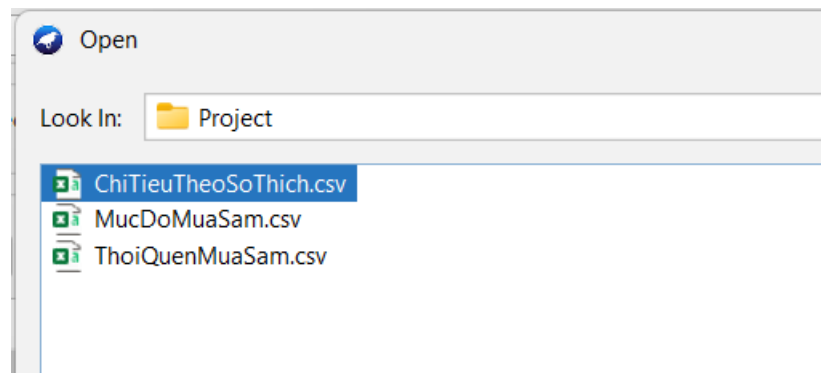
Các bước chạy thuật toán trên Weka:

❖ **Bước 1:** Mở phần mềm Weka và chọn Explorer



Hình 4. 8: Mở phần mềm Weka, chọn Explorer

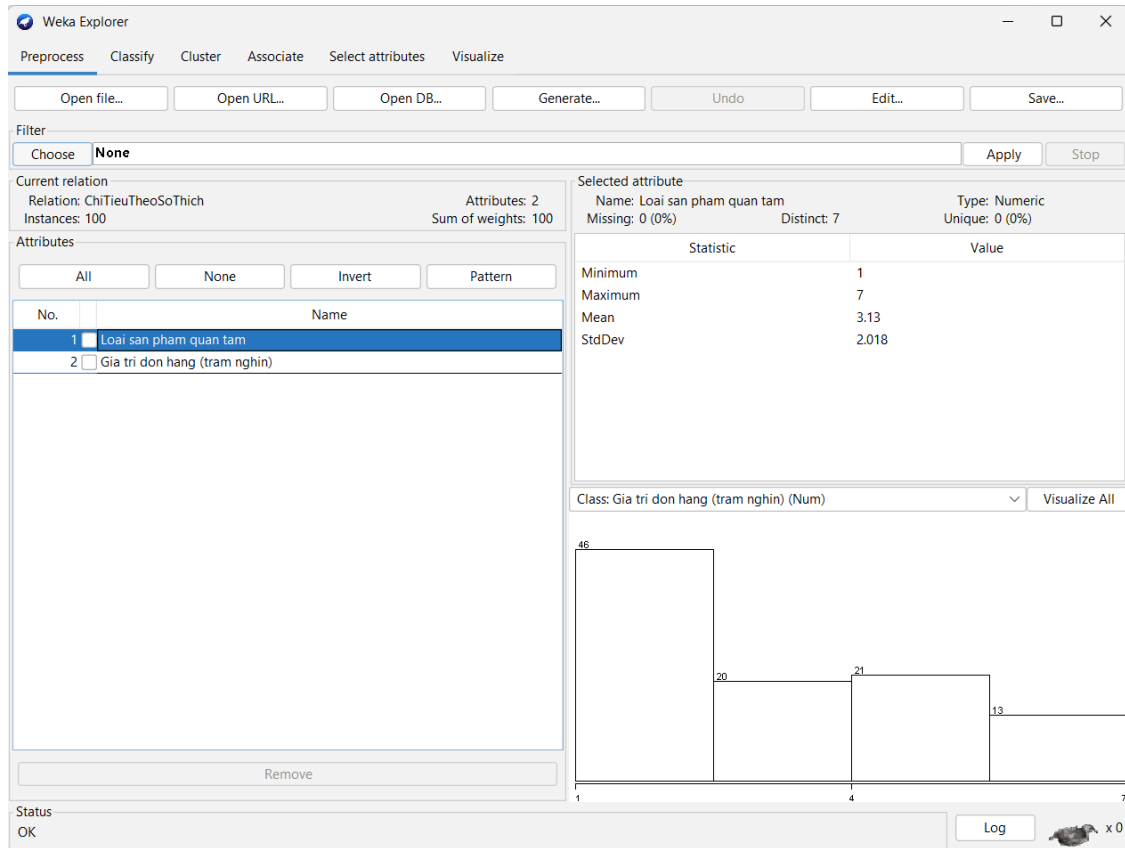
❖ **Bước 2:** Tại tab Preprocess -> Openfile và tìm chọn tệp ChiTieuTheoSoThich.csv (File này tạo ở bước tiền xử lý dữ liệu nằm trong giai đoạn “Chuẩn bị dữ liệu”):



Hình 4. 9: Chọn tệp dữ liệu cần phân cụm

Sử dụng thuật toán K-Means để phân tích khách hàng dựa trên hành vi mua hàng trên TiktokShop

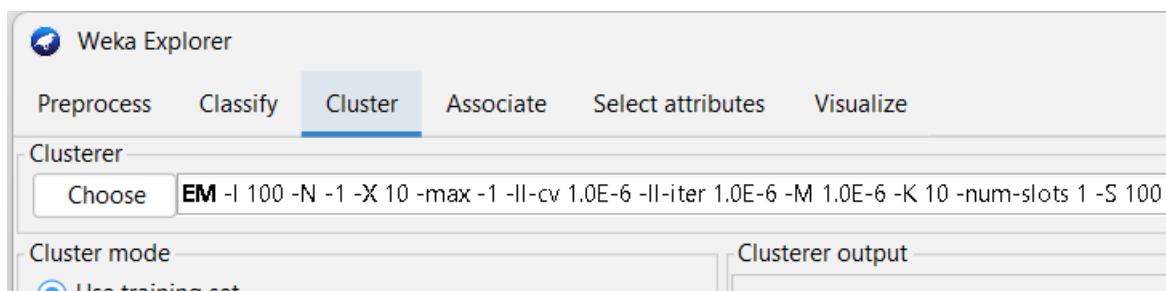
Sau khi mở file, chúng ta sẽ thấy 2 thuộc tính như hình dưới:



Hình 4. 10: Sau khi mở file thành công

Do dữ liệu đã được “sạch”, và xử lý đồng bộ với nhau ở bước tiền xử lý dữ liệu. Và cũng không còn thuộc tính thừa. Nên ta sẽ sang bước tiếp theo.

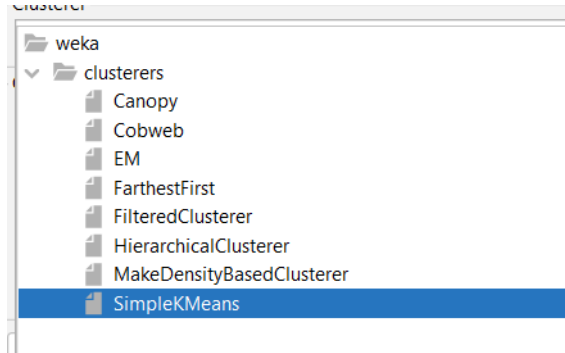
❖ **Bước 3:** Chuyển sang tab Cluster - > Choose. Và chọn vào SimpleKMeans, để sử dụng thuật toán Kmeans.



Hình 4. 11: Chọn vào tab Cluster

Sử dụng thuật toán K-Means để phân tích khách hàng dựa trên hành vi mua hàng trên TiktokShop

Bấm chọn vào “SimpleKMeans”:



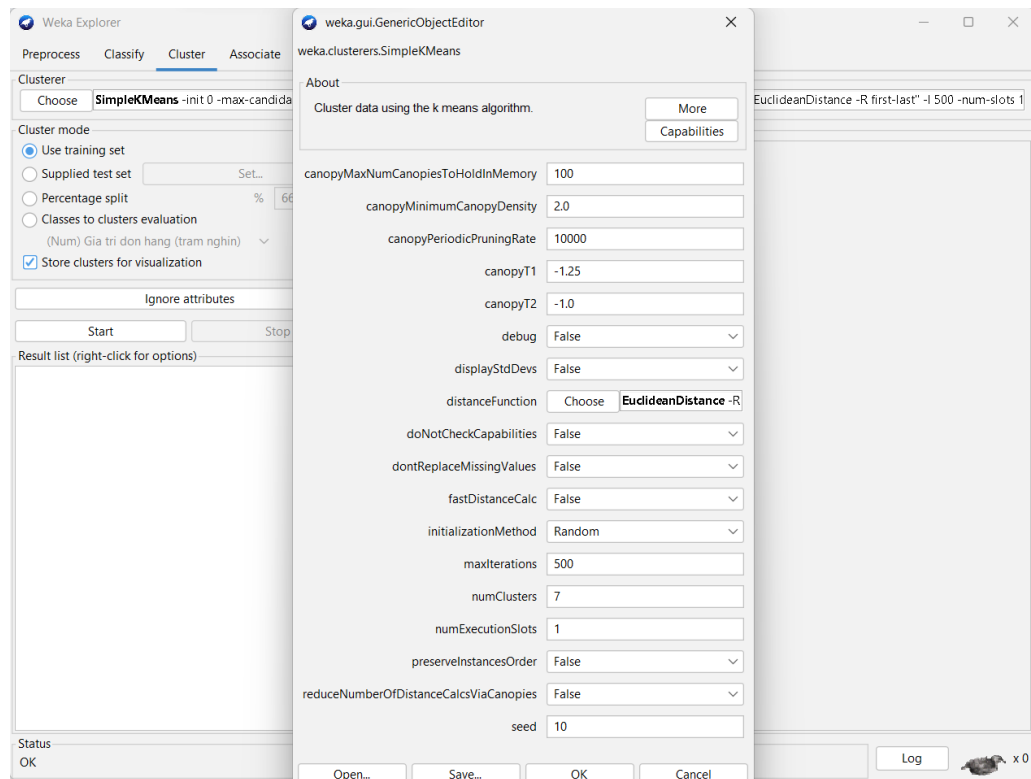
Hình 4. 12: Chọn vào thuật toán SimpleKMeans

❖ **Bước 4:** Điều chỉnh các giá trị như sau:

Tại distanceFunction, sẽ sử dụng độ đo khoảng cách: EuclideanDistance

Tại numClusters, sẽ chọn số cụm là: 7.

Sau khi chỉnh xong sẽ bấm OK.



Hình 4. 13: Điều chỉnh các giá trị của thuật toán

Sử dụng thuật toán K-Means để phân tích khách hàng dựa trên hành vi mua hàng trên TiktokShop

❖ **Bước 5:** Bấm vào Start, và kết quả sẽ hiện ra như sau:

```
kMeans
=====

Number of iterations: 7
Within cluster sum of squared errors: 2.1068865396888268

Initial starting points (random):

Cluster 0: 3,8
Cluster 1: 5,5
Cluster 2: 2,6
Cluster 3: 1,2
Cluster 4: 1,3
Cluster 5: 5,4
Cluster 6: 5,3

Missing values globally replaced with mean/mode

Final cluster centroids:

Attribute          Full Data      Cluster#
                   (100.0)      (8.0)      1      2      3      4      5      6
=====
Loai san pham quan tam      3.13      1.5      4.875      1.9412      1.75      1.1176      4.7273      6.5
Gia tri don hang (tram nghin) 4.27      9.125      7.75      5.5882      1.6875      3.4118      3.4091      3.0833

Time taken to build model (full training data) : 0.02 seconds
```

Hình 4. 14: Kết quả của thuật toán

Giải thích kết quả:

- "Number of iterations": 7.

Cho biết số lần cập nhật lại vị trí của các trung tâm cụm là 7.

- "Within cluster sum of squared errors": 2.1068865396888268.

Tổng bình phương khoảng cách từ các điểm dữ liệu đến trung tâm cụm gần nhất là: 2.1068865396888268. Giá trị này càng thấp thì các điểm trong cùng một cụm sẽ càng gần nhau.

- "Initial starting points" là vị trí ban đầu của các trung tâm cụm được chọn ngẫu nhiên. 7 Cluster được chọn với hai thuộc tính “Loại sản phẩm quan tâm” và “Giá trị đơn hàng”:

- + Cluster 0: 3,8

- + Cluster 1: 5,5

Sử dụng thuật toán K-Means để phân tích khách hàng dựa trên hành vi mua hàng trên TiktokShop

- + Cluster 2: 2,6
 - + Cluster 3: 1,2
 - + Cluster 4: 1,3
 - + Cluster 5: 5,4
 - + Cluster 6: 5,3
- "Final cluster centroids": thông tin về các tâm cụm cuối cùng của các cụm được tạo ra dựa trên tập dữ liệu đã cho. Nó hiển thị giá trị trung bình của các thuộc tính cho mỗi cụm, cùng với số lượng các mẫu trong mỗi cụm.

Ta có thể thấy bảng trên gồm có 7 Cluster được đánh số từ 0 đến 6, và 2 thuộc tính là “Loại sản phẩm quan tâm” và “Giá trị đơn hàng”. Ví dụ: Tại Cluster 5, ta có 22 mẫu dữ liệu trong cụm này, và giá trị trung bình của “Loại sản phẩm quan tâm” là 4.7273 và giá trị trung bình của “Giá trị đơn hàng” là 3.4091.

```
=== Model and evaluation on training set ===  
  
Clustered Instances  
  
0      8 ( 8%)  
1      8 ( 8%)  
2     17 ( 17%)  
3     16 ( 16%)  
4     17 ( 17%)  
5     22 ( 22%)  
6     12 ( 12%)
```

Hình 4. 15: Kết quả của thuật toán

Giải thích kết quả:

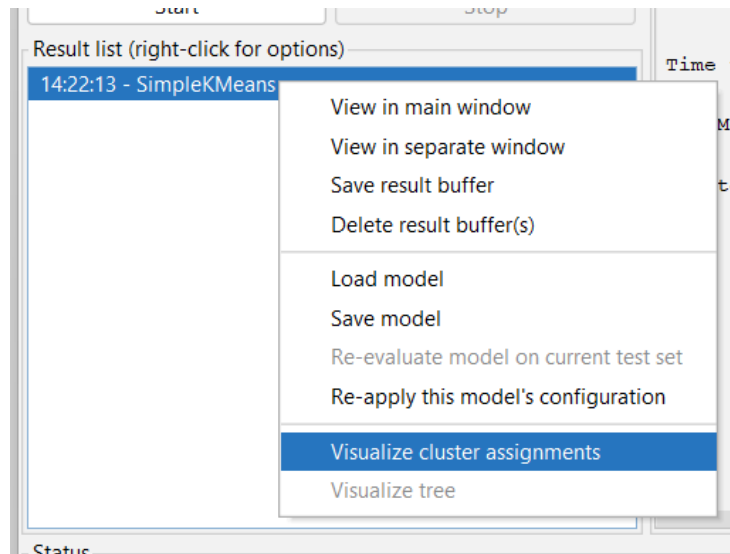
- Clustered Instances: Số lượng các mẫu dữ liệu đã được phân vào từng cụm tương ứng.

Sử dụng thuật toán K-Means để phân tích khách hàng dựa trên hành vi mua hàng trên TiktokShop

- Số trong ngoặc đơn (%): tỉ lệ phần trăm số lượng mẫu dữ liệu trong từng cụm so với tổng số mẫu dữ liệu. Ví dụ: cụm 5 có 22 mẫu dữ liệu, chiếm 22% trong tổng số mẫu dữ liệu.

Thông tin này cung cấp cái nhìn tổng quan về phân bố dữ liệu sau khi đã được phân cụm.

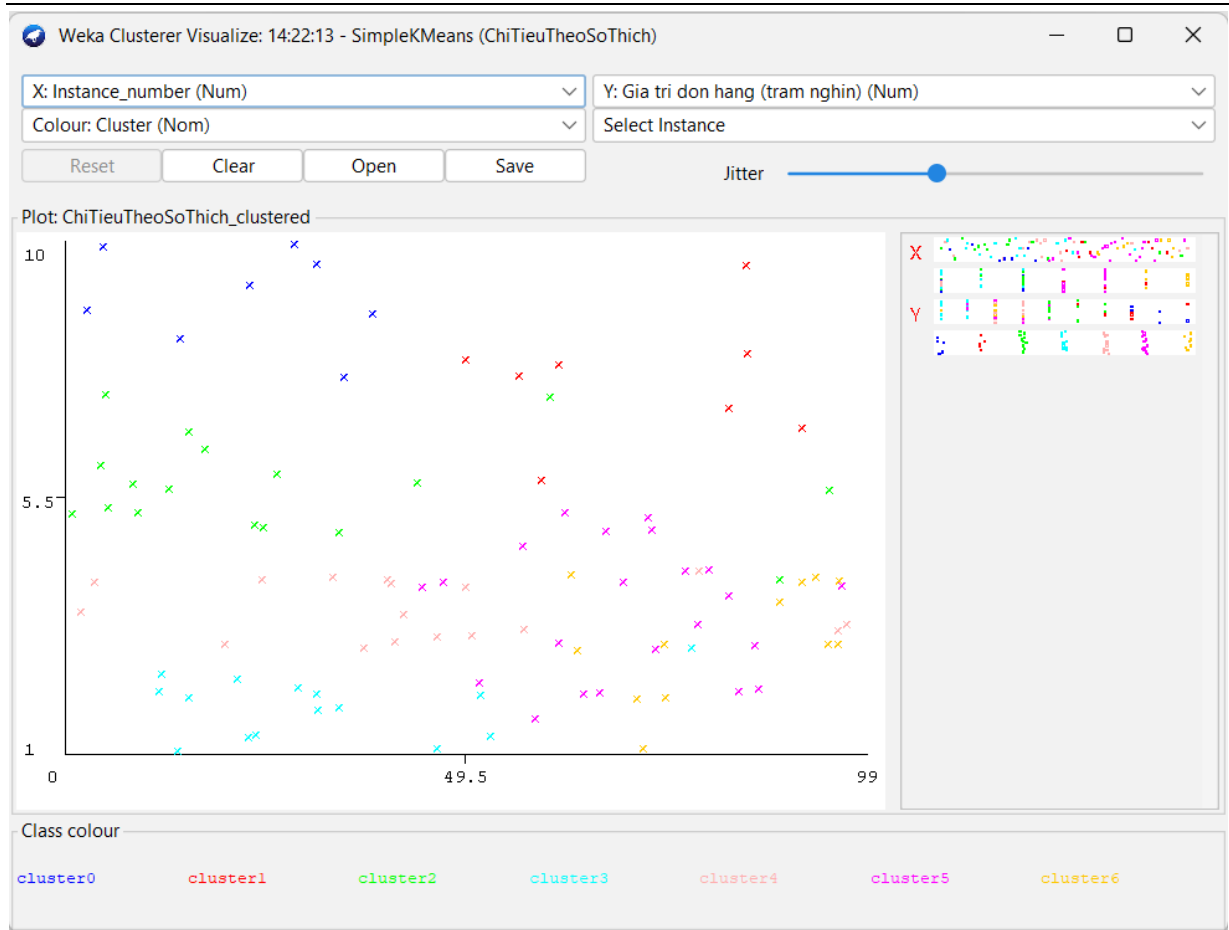
- ❖ **Bước 6:** Để trực quan hóa dữ liệu, ta sẽ nhìn vào màn hình “Result List” ở góc trái màn hình. Và nhấp chuột phải vào kết quả đã chạy và chọn “Visualize cluster assignments”:



Hình 4. 16: Hiển thị mô hình hóa dữ liệu trên kết quả

Lúc này màn hình trực quan hóa dữ liệu sẽ được hiện lên:

Sử dụng thuật toán K-Means để phân tích khách hàng dựa trên hành vi mua hàng trên TiktokShop



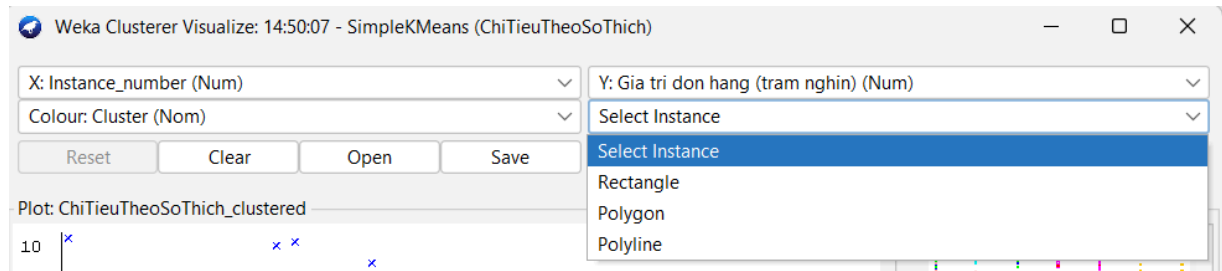
Hình 4. 17: Dữ liệu được trực quan hóa

Giải thích kết quả:

- Class Colour: Là màu sắc của mỗi cụm. Ở đây dữ liệu có 7 cụm tương ứng với các màu sắc từ Cluster 0 cho đến Cluster 6.
- X: Instance_number: Ở đây trục x được chọn là thuộc tính instance_number. Đây là một thuộc tính ảo được tạo ra bởi Weka trong quá trình trực quan hóa kết quả. Thuộc tính này chỉ đơn giản là số thứ tự của các mẫu dữ liệu trong tập dữ liệu của bạn
- Y: Gia tri don hang (tram nghin): Ở đây trục y được chọn là thuộc tính “Giá trị đơn hàng” của tập dữ liệu.

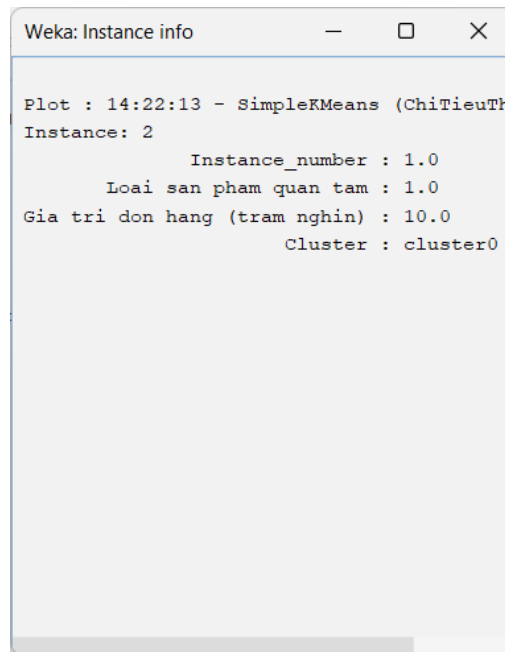
Sử dụng thuật toán K-Means để phân tích khách hàng dựa trên hành vi mua hàng trên TiktokShop

Để có thể xem thông tin của điểm nào đó thì chúng ta sẽ chọn vào “Select Instance”. Sau đó “Click trái” vào điểm đó để xem thông tin:



Hình 4. 18: Chọn “Select Instance” để cho hiển thị bản ghi khi nhấn vào các quan sát

Đây là thông tin của quan sát:



Hình 4. 19: Thông tin của quan sát

Giải thích kết quả:

Kết quả này được tính cho Instance số 2 trong tập dữ liệu, và Instance này có các giá trị thuộc tính sau: "Loai san pham quan tam" là 1.0, "Gia tri don hang (tram nghin)" là 10.0. Sau khi chạy thuật toán SimpleKMeans, Instance này đã được phân vào Cluster "cluster0".

Sử dụng thuật toán K-Means để phân tích khách hàng dựa trên hành vi mua hàng trên TiktokShop

4.5 Đánh giá mô hình

```
kMeans
=====

Number of iterations: 7
Within cluster sum of squared errors: 2.1068865396888268
```

Hình 4. 20: Đánh giá mô hình dựa trên chỉ số đánh giá “Sum of Square Errors”

Trong trường hợp này, Chỉ số đánh giá "Within cluster sum of squared errors" của thuật toán Kmeans là 2.1068865396888268. Điều này cho thấy các điểm dữ liệu trong cùng một cluster có xu hướng gần nhau hơn so với các điểm dữ liệu ở các cluster khác. Hay nói cách khác thuật toán đã tạo ra các cụm có tính tương đồng cao với nhau.

4.6 Triển khai mô hình

```
Final cluster centroids:

Attribute          Full Data          Cluster#
                   (100.0)         0          1          2          3          4          5          6
=====
Loai san pham quan tam    3.13          1.5          4.875          1.9412          1.75          1.1176          4.7273          6.5
Gia tri don hang (tram nghin)  4.27          9.125          7.75          5.5882          1.6875          3.4118          3.4091          3.0833

Time taken to build model (full training data) : 0 seconds

=== Model and evaluation on training set ===

Clustered Instances
0          8 ( 8%)
1          8 ( 8%)
2         17 ( 17%)
3         16 ( 16%)
4         17 ( 17%)
5         22 ( 22%)
6         12 ( 12%)
```

Hình 4. 21: Triển khai mô hình thực hiện dự đoán

Thực hiện dự đoán:

Dựa vào kết quả của thuật toán Kmeans, ta có thể thấy các mẫu dữ liệu khách hàng đã được phân vào các cụm khách hàng có điểm chung với nhau, qua đây chúng ta có thể đưa ra những dự đoán như:

Sử dụng thuật toán K-Means để phân tích khách hàng dựa trên hành vi mua hàng trên TiktokShop

- Tại Cluster 0:
 - + Có 8 mẫu khách hàng có chung hành vi mua hàng với nhau, chiếm tổng số 8% tổng số mẫu.
 - + Trung bình “Loại sản phẩm quan tâm” của họ là 1.5, ứng với sản phẩm “Nhà cửa & Đời Sống”.
 - + “Giá trị đơn hàng (trăm nghìn)” của nhóm này là 9.125, nghĩa là trung bình mọi người trong nhóm này chi tiền trung bình cho mỗi đơn hàng là 912.500đ.
- Tại Cluster 1:
 - + Có 8 mẫu khách hàng có chung hành vi mua hàng với nhau, chiếm tổng số 8% tổng số mẫu.
 - + Trung bình “Loại sản phẩm quan tâm” của họ là 4.875, ứng với sản phẩm “Sức Khỏe”.
 - + “Giá trị đơn hàng (trăm nghìn)” của nhóm này là 7.75, nghĩa là trung bình mọi người trong nhóm này chi tiền trung bình cho mỗi đơn hàng là 775.000đ.
- Tại Cluster 2:
 - + Có 17 mẫu khách hàng có chung hành vi mua hàng với nhau, chiếm tổng số 17% tổng số mẫu.
 - + Trung bình “Loại sản phẩm quan tâm” của họ là 1.9412, ứng với sản phẩm “Nhà cửa & Đời Sống”.
 - + “Giá trị đơn hàng (trăm nghìn)” của nhóm này là 5.5882, nghĩa là trung bình mọi người trong nhóm này chi tiền trung bình cho mỗi đơn hàng là 558.820đ.
- Tại Cluster 3:
 - + Có 16 mẫu khách hàng có chung hành vi mua hàng với nhau, chiếm tổng số 16% tổng số mẫu.

Sử dụng thuật toán K-Means để phân tích khách hàng dựa trên hành vi mua hàng trên TiktokShop

- + Trung bình “Loại sản phẩm quan tâm” của họ là 1.75, ứng với sản phẩm “Nhà cửa & Đời Sống”.
- + “Giá trị đơn hàng (trăm nghìn)” của nhóm này là 1.6875, nghĩa là trung bình mọi người trong nhóm này chi tiền trung bình cho mỗi đơn hàng là 168.750đ.
- Tại Cluster 4:
 - + Có 17 mẫu khách hàng có chung hành vi mua hàng với nhau, chiếm tổng số 17% tổng số mẫu.
 - + Trung bình “Loại sản phẩm quan tâm” của họ là 1.1176, ứng với sản phẩm “Thời trang”.
 - + “Giá trị đơn hàng (trăm nghìn)” của nhóm này là 3.4118, nghĩa là trung bình mọi người trong nhóm này chi tiền trung bình cho mỗi đơn hàng là 341.180đ.
- Tại Cluster 5:
 - + Có 22 mẫu khách hàng có chung hành vi mua hàng với nhau, chiếm tổng số 22% tổng số mẫu.
 - + Trung bình “Loại sản phẩm quan tâm” của họ là 4.7273, ứng với sản phẩm “Sức Khỏe”.
 - + “Giá trị đơn hàng (trăm nghìn)” của nhóm này là 3.4091, nghĩa là trung bình mọi người trong nhóm này chi tiền trung bình cho mỗi đơn hàng là 340.910đ.
- Tại Cluster 6:
 - + Có 12 mẫu khách hàng có chung hành vi mua hàng với nhau, chiếm tổng số 12% tổng số mẫu.
 - + Trung bình “Loại sản phẩm quan tâm” của họ là 6.5, ứng với sản phẩm “Trang sức và phụ kiện”.

Sử dụng thuật toán K-Means để phân tích khách hàng dựa trên hành vi mua hàng trên TiktokShop

- + “Giá trị đơn hàng (trăm nghìn)” của nhóm này là 3.0833, nghĩa là trung bình mọi người trong nhóm này chi tiền trung bình cho mỗi đơn hàng là 308.330đ.

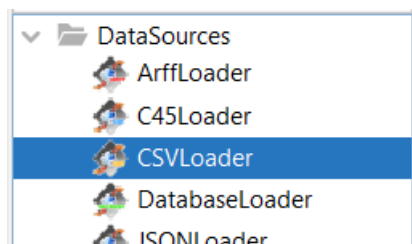
4.7 Chạy thuật toán K-Means trên Weka KnowledgeFlow

- ❖ **Bước 1:** Đầu tiên chúng ta mở Weka lên, và chọn vào KnowledgeFlow.



Hình 4. 22: Mở Weka, và chọn KnowledgeFlow

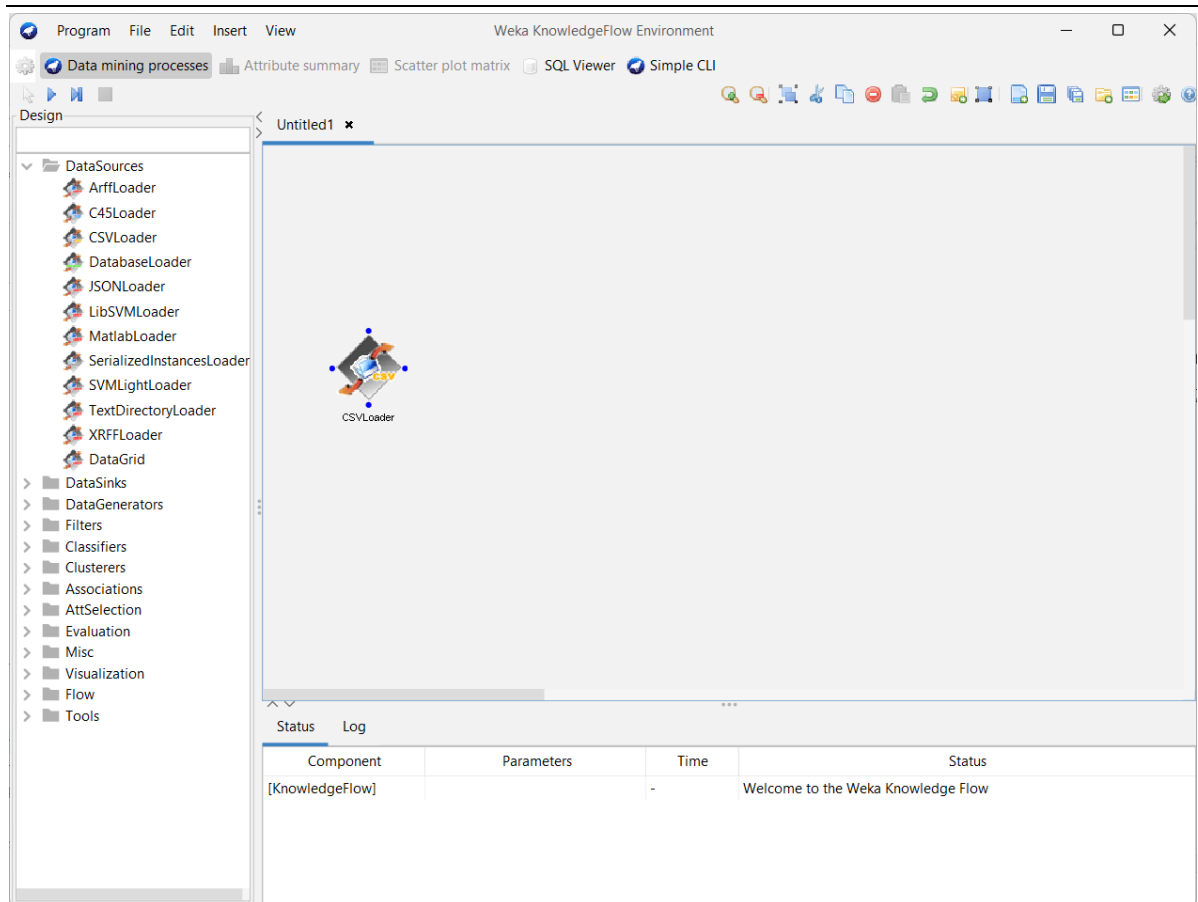
- ❖ **Bước 2:** Để tải file dữ liệu lên ta cần vào “DataSources”, và chọn định dạng file tương ứng với mô hình. Ở đây, do cần tải file csv lên, nên sẽ chọn vào “CSVLoader”:



Hình 4. 23: Chọn DataSources

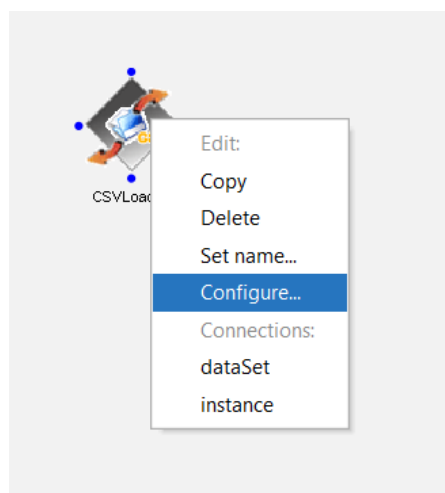
Sau khi chọn được, ta sẽ bấm vào màn hình kế bên để nó hiện lên:

Sử dụng thuật toán K-Means để phân tích khách hàng dựa trên hành vi mua hàng trên TiktokShop



Hình 4. 24: Đặt DataSource vào màn hình

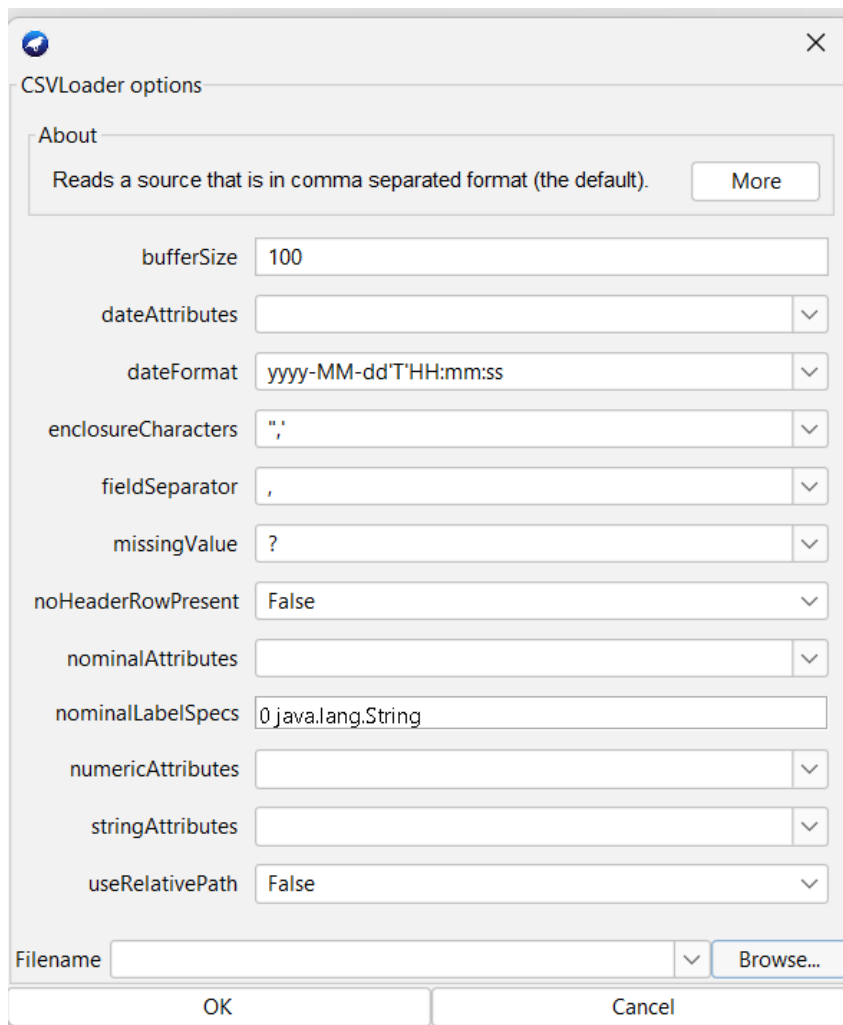
Bước 3: Tiếp theo sẽ nạp dữ liệu vào, nhấp phải chuột vào và chọn “Configure”.



Hình 4. 25: Chọn Configure để nạp dữ liệu

Sử dụng thuật toán K-Means để phân tích khách hàng dựa trên hành vi mua hàng trên TiktokShop

Lúc này một bảng hiện lên, và tại vị trí Filename, bấm vào “Browse” và chọn vào file cần tải lên:



CSVLoader options

About

Reads a source that is in comma separated format (the default). [More](#)

bufferSize 100

dateAttributes

dateFormat yyyy-MM-dd'T'HH:mm:ss

enclosureCharacters ;

fieldSeparator ,

missingValue ?

noHeaderRowPresent False

nominalAttributes

nominalLabelSpecs 0 java.lang.String

numericAttributes

stringAttributes

useRelativePath False

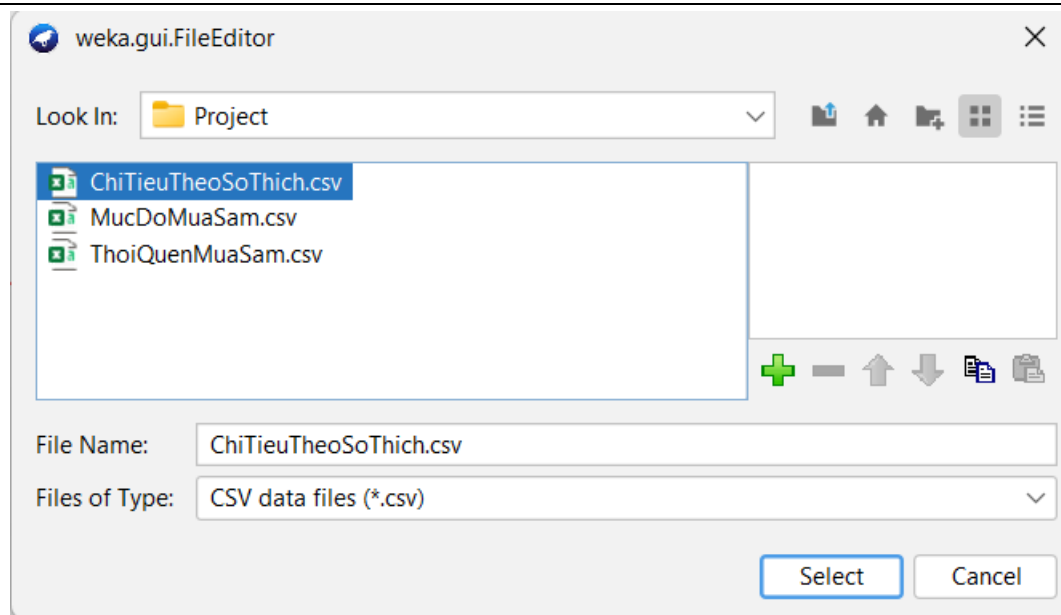
Filename [Browse...](#)

OK Cancel

Hình 4. 26: Chọn Browse

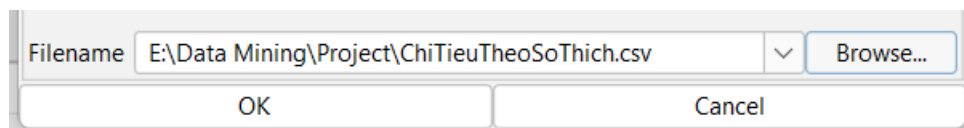
Sau khi chọn được file đã tải lên, thì bấm Select:

Sử dụng thuật toán K-Means để phân tích khách hàng dựa trên hành vi mua hàng trên TiktokShop



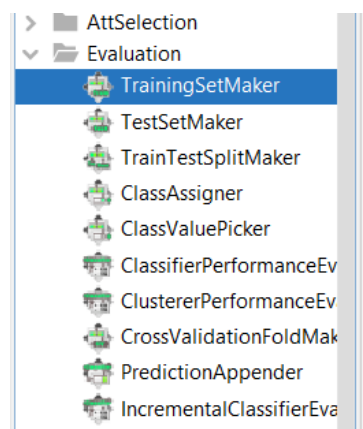
Hình 4. 27: Chọn vào file cần chạy thuật toán

Lúc này File đã được hiển thị, bấm tiếp vào OK:



Hình 4. 28: File đã được hiển thị

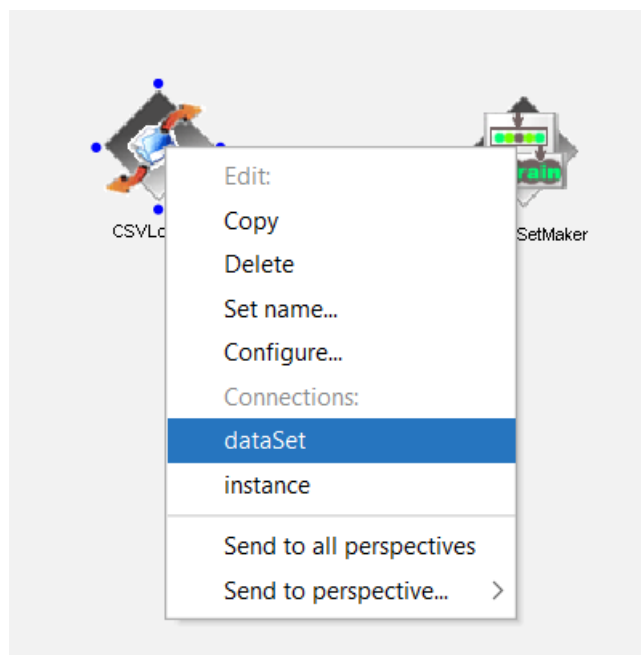
❖ **Bước 4:** Tiếp theo tại Evaluation, ta sẽ chọn vào TrainingSetMaker và bỏ vào:



Hình 4. 29: Chọn vào TrainingSetMaker tại Evaluation

Sử dụng thuật toán K-Means để phân tích khách hàng dựa trên hành vi mua hàng trên TiktokShop

- ❖ **Bước 5:** Để truyền dữ liệu từ file csv, chuột phải vào “CSVLoader”, chọn vào “dataSet”:



Hình 4. 30: Chọn vào dataSet

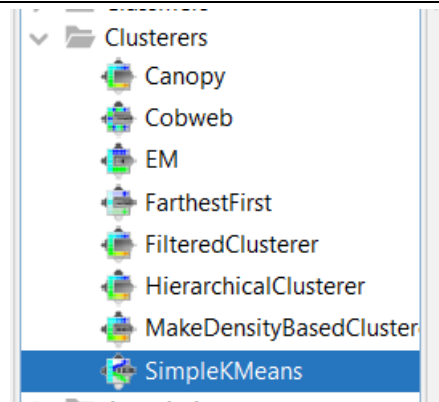
Sau đó nối đến “TrainingSetMaker”:



Hình 4. 31: Truyền dữ liệu đến TrainingSetMaker

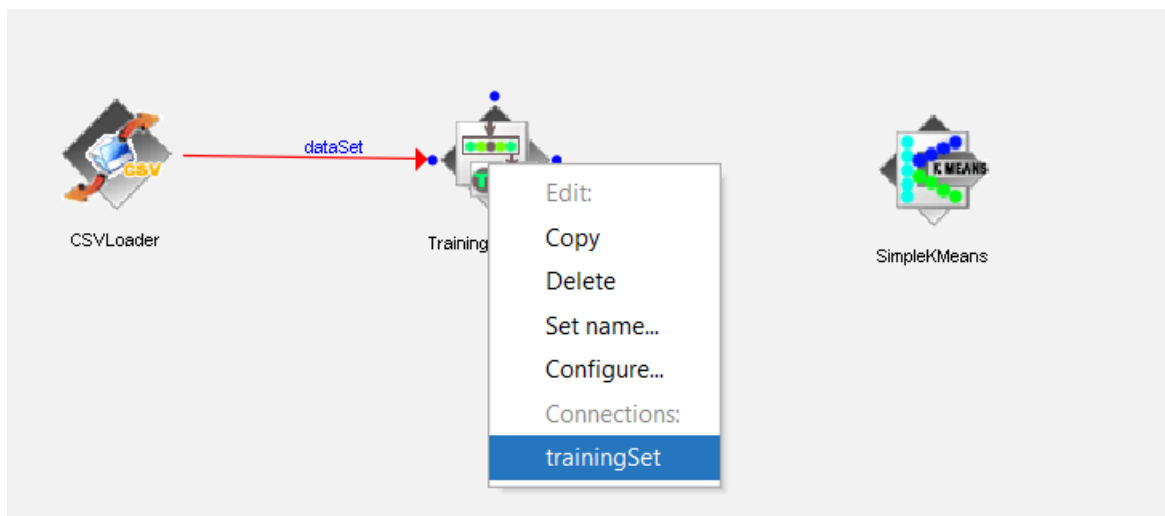
- ❖ **Bước 6:** Tại “Clusterers” nơi chứa những thuật toán phân cụm, và chọn vào “SimpleKMeans” để sử dụng thuật toán Kmeans:

Sử dụng thuật toán K-Means để phân tích khách hàng dựa trên hành vi mua hàng trên TiktokShop



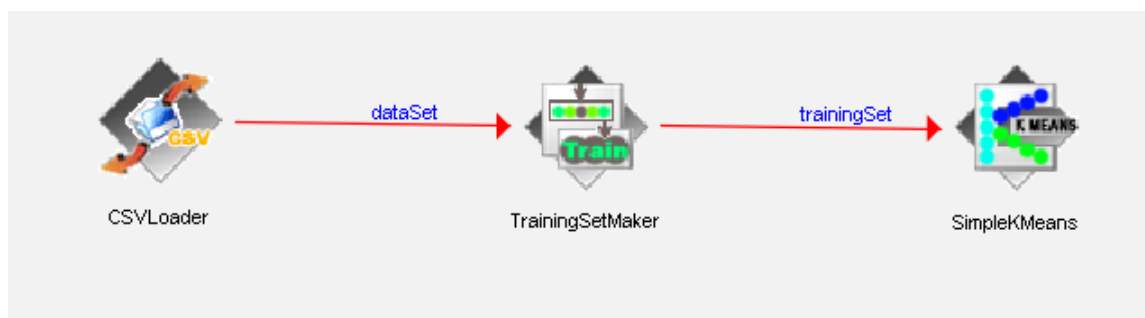
Hình 4. 32: Chọn thuật toán SimpleKMeans tại Clusterers

❖ **Bước 7:** Tại “TrainingSetMaker”, sẽ chuột phải vào và chọn “trainingSet”:



Hình 4. 33: Chọn TrainingSet

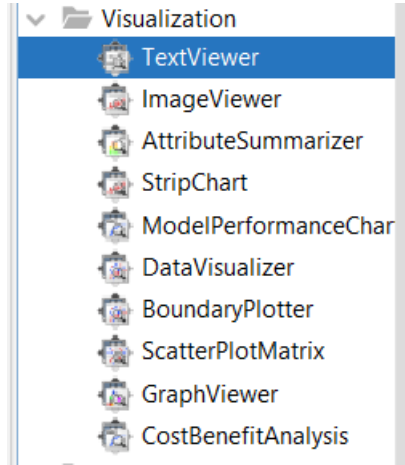
Sau đó đưa nó vào thuật toán “SimpleKMeans”:



Hình 4. 34: Truyền dữ liệu vào thuật toán SimpleKmeans

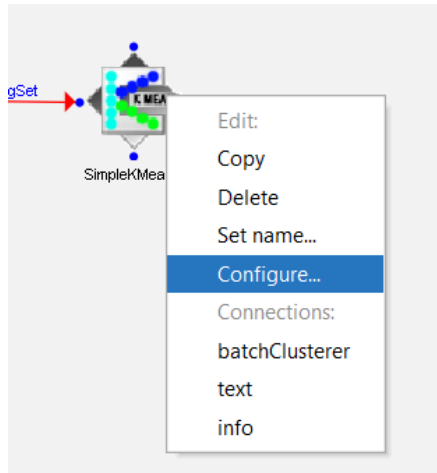
Sử dụng thuật toán K-Means để phân tích khách hàng dựa trên hành vi mua hàng trên TiktokShop

- ❖ **Bước 8:** Vào “Visualization” chọn vào “TextViewer” để truyền dữ liệu kết quả từ thuật toán vào:



Hình 4. 35: Tạo TextViewer để chứa kết quả thuật toán

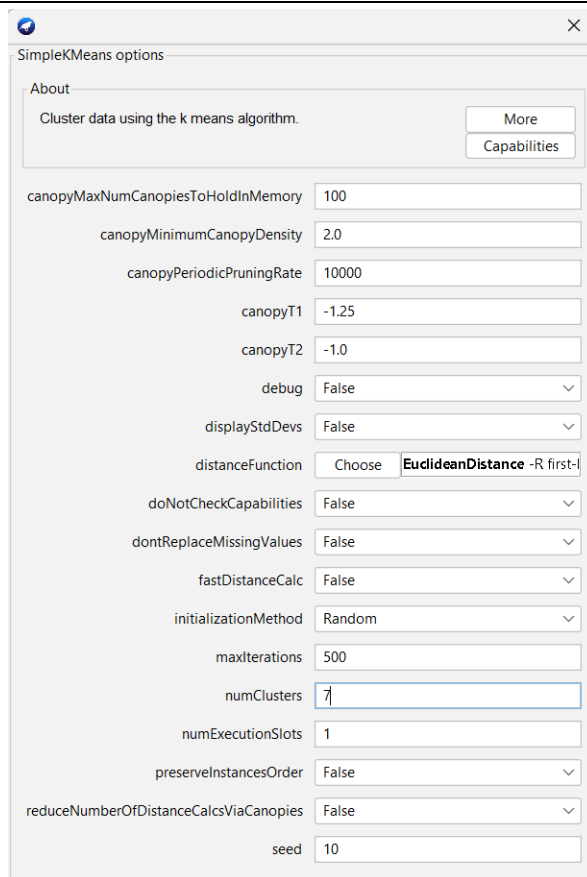
- ❖ **Bước 9:** Tiếp đến ta sẽ cài đặt cho thuật toán bằng cách chuột phải vào “SimpleKmeans”, và chọn “Configure”:



Hình 4. 36: Chọn Configure để thiết lập thuật toán

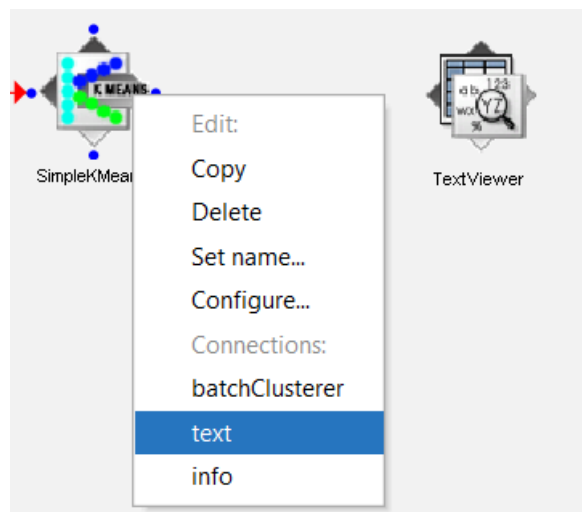
Ở chỗ tùy chỉnh này, có thể thay đổi cách đo khoảng cách tại “distanceFunction” và thay đổi số cụm tại “numClusters”. Ở đây, chọn đo khoảng cách “Euclidean” và số cụm là 7:

Sử dụng thuật toán K-Means để phân tích khách hàng dựa trên hành vi mua hàng trên TiktokShop



Hình 4. 37: Chọn giá trị đo và số cụm

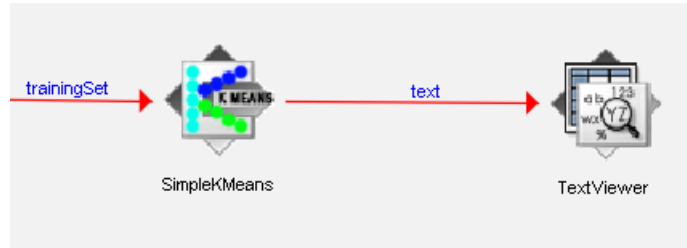
❖ **Bước 10:** Tại Thuật toán SimpleKMeans, sẽ chuột phải vào “Text”:



Hình 4. 38: Chọn vào text

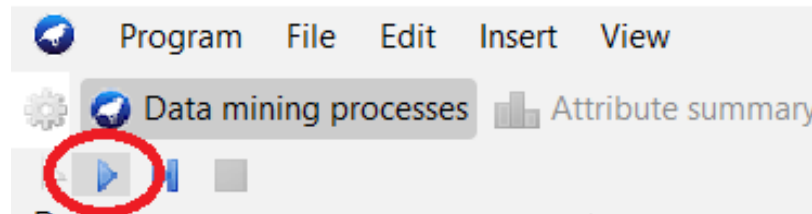
Sử dụng thuật toán K-Means để phân tích khách hàng dựa trên hành vi mua hàng trên TiktokShop

Sau đó dẫn đến “TextViewer”:



Hình 4. 39: Truyền kết quả vào TextViewer

❖ **Bước 11:** Để chạy thuật toán, ta sẽ chọn vào biểu tượng này ở trên bên trái:



Hình 4. 40: Chạy thuật toán

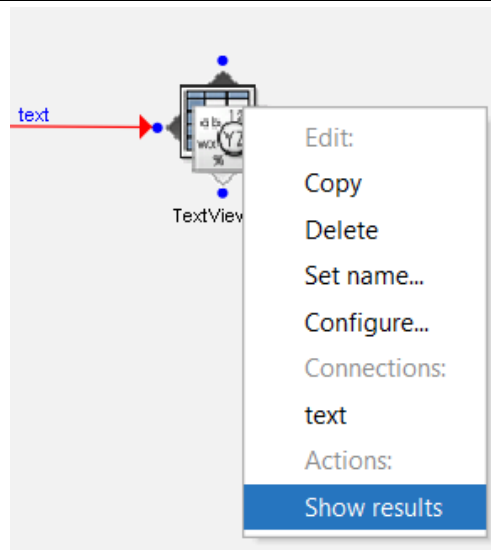
Khi thuật toán chạy thành công, không bị lỗi gì thì sẽ có thông báo như sau:

Status Log				
Component	Parameters	Time	Status	
[KnowledgeFlow]			OK.	
CSVLoader	-format "yyyy-MM-dd\T\HH:mm:ss" -M ? -B 100 -E "\\\\"	-	Finished.	
TrainingSetMaker		-	Finished.	
SimpleKMeans	-init 0 -max-candidates 100 -periodic-pruning 10000 -min-	-	Finished.	
TextViewer		-	Finished.	

Hình 4. 41: Hiển thị thông báo chạy thành công

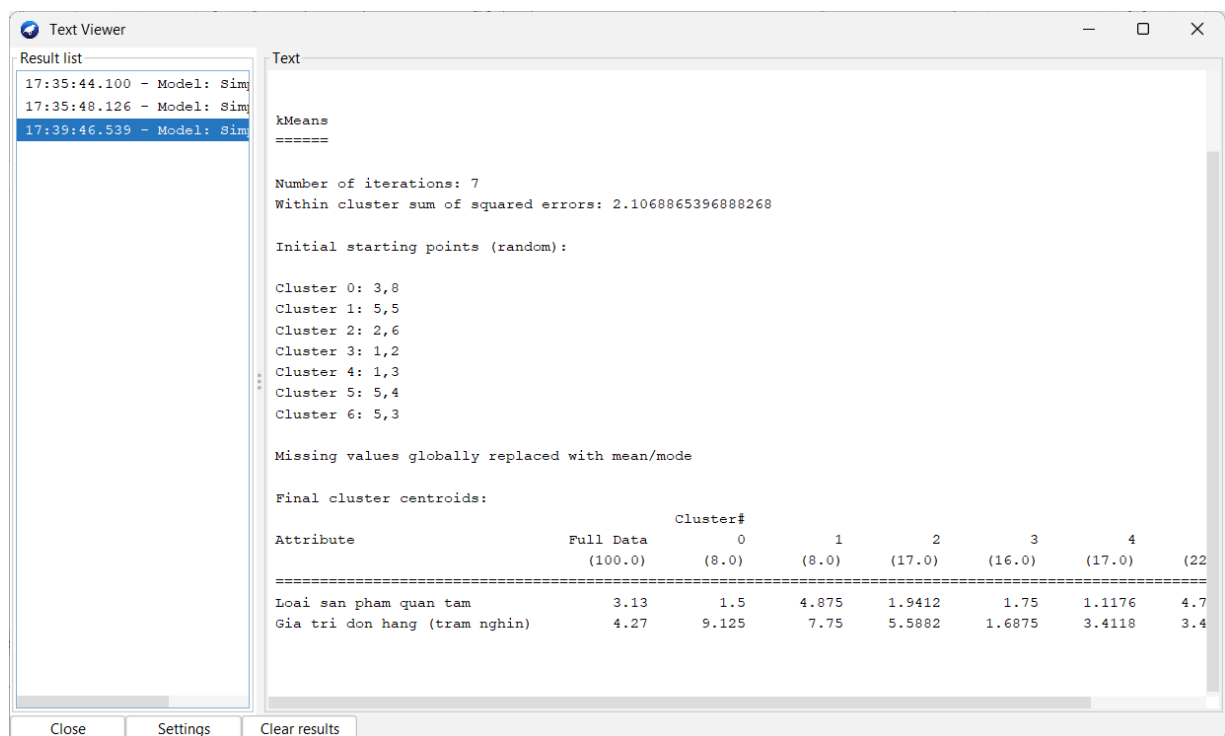
❖ **Bước 12:** Để xem được kết quả của thuật toán, ta sẽ chuột phải vào “TextViewer”, chọn vào “Show results”:

Sử dụng thuật toán K-Means để phân tích khách hàng dựa trên hành vi mua hàng trên TiktokShop



Hình 4. 42: Chọn Show results để hiển thị kết quả thuật toán

Lúc này thông tin về thuật toán sẽ được hiển lên:



Hình 4. 43: Màn hình hiển thị kết quả của thuật toán

Sử dụng thuật toán K-Means để phân tích khách hàng dựa trên hành vi mua hàng trên TiktokShop

Giải thích kết quả:

Giải thuật K-means đã thực hiện được 7 vòng lặp để tìm ra được các trung tâm cụm (cluster centers) tối ưu nhất cho dữ liệu đầu vào.

Tổng bình phương sai số (Within cluster sum of squared errors) của các điểm dữ liệu tính từ trung tâm của các cụm là 2.1068865396888268. Giá trị này càng nhỏ thì chất lượng phân cụm càng tốt.

Mỗi cụm (cluster) được mô tả bởi một tập hợp các điểm dữ liệu có sự tương đồng với nhau. Kết quả trả về hiển thị trung tâm của mỗi cụm được xác định bằng cách chạy giải thuật K-means với một bộ khởi tạo ngẫu nhiên (Initial starting points) ban đầu. Cụ thể, trung tâm của 7 cụm với bộ khởi tạo ngẫu nhiên ban đầu lần lượt là:

- + Cluster 0: 3,8
- + Cluster 1: 5,5
- + Cluster 2: 2,6
- + Cluster 3: 1,2
- + Cluster 4: 1,3
- + Cluster 5: 5,4
- + Cluster 6: 5,3

Kết quả hiển thị “Final Cluster Centroids” cho thấy giá trị trung bình của các thuộc tính trong mỗi cluster. Cụ thể, trong bảng này, có 7 cluster được đánh số từ 0 đến 6. Đối với mỗi cluster, bảng hiển thị giá trị trung bình của hai thuộc tính “Loại sản phẩm quan tâm” và “Giá trị đơn hàng (trăm nghìn đồng)”. Ví dụ, Cluster 0 có giá trị trung bình “Loại sản phẩm quan tâm” là 1,5 và “Giá trị đơn hàng” là 912.500 đồng. Kết quả này cung cấp thông tin về cách các khách hàng được phân loại thành các nhóm dựa trên các thuộc tính được sử dụng trong phân cụm.

Sử dụng thuật toán K-Means để phân tích khách hàng dựa trên hành vi mua hàng trên TiktokShop

4.8 Chạy thuật toán K-Means trên R

Để có thể chạy thuật toán Kmeans và trực quan hóa dữ liệu trong R, sẽ có rất nhiều thư viện khác nhau hỗ trợ cho việc này, nhưng ở đây nhóm sẽ chọn 4 thư viện chính là “stats”, “dplyr”, “ggplot2”, “ggfortify”.

- ❖ **Bước 1:** Mở R, nếu chưa cài đặt các thư viện trên, thì đầu tiên cần chạy các lệnh sau để tiến hành cài thư viện:

```
install.packages("stats")
```

```
install.packages("dplyr")
```

```
install.packages("ggplot2")
```

```
install.packages("ggfortify")
```

- ❖ **Bước 2:** Sau khi cài đặt thành công, bạn có thể sử dụng thư viện bằng cách gọi hàm trong R:

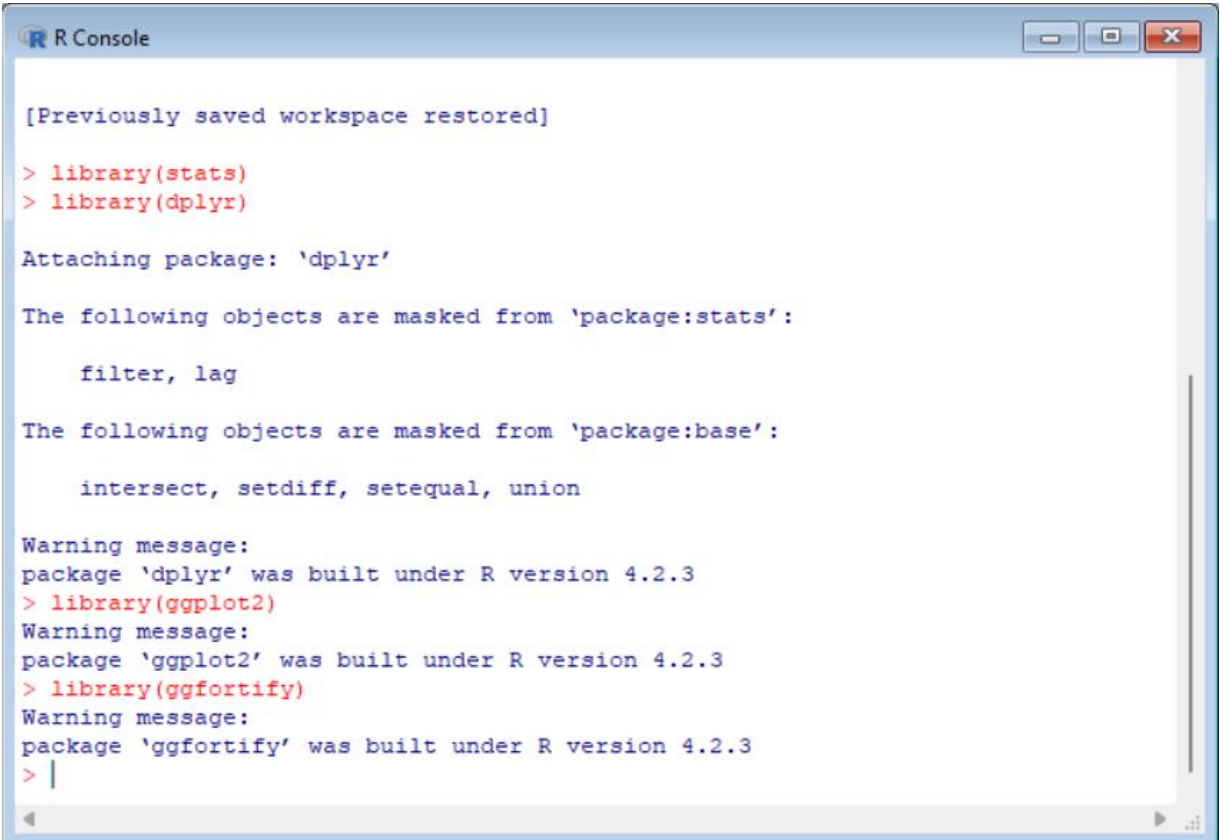
```
library(stats)
```

```
library(dplyr)
```

```
library(ggplot2)
```

```
library(ggfortify)
```

Sử dụng thuật toán K-Means để phân tích khách hàng dựa trên hành vi mua hàng trên TiktokShop



```
R Console

[Previously saved workspace restored]

> library(stats)
> library(dplyr)

Attaching package: 'dplyr'

The following objects are masked from 'package:stats':

  filter, lag

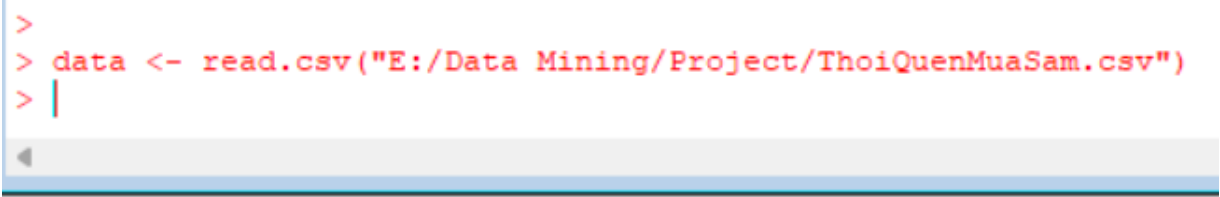
The following objects are masked from 'package:base':

  intersect, setdiff, setequal, union

Warning message:
package 'dplyr' was built under R version 4.2.3
> library(ggplot2)
Warning message:
package 'ggplot2' was built under R version 4.2.3
> library(ggfortify)
Warning message:
package 'ggfortify' was built under R version 4.2.3
> |
```

Hình 4. 44: Gọi hàm trong R

- ❖ **Bước 3:** Đọc tệp dữ liệu ThoiQuenMuaSam.csv (Tệp dữ liệu đã tạo ở bước tiền xử lý dữ liệu trong giai đoạn “Chuẩn bị dữ liệu”) bằng hàm `read.csv` và lưu vào biến `data`:



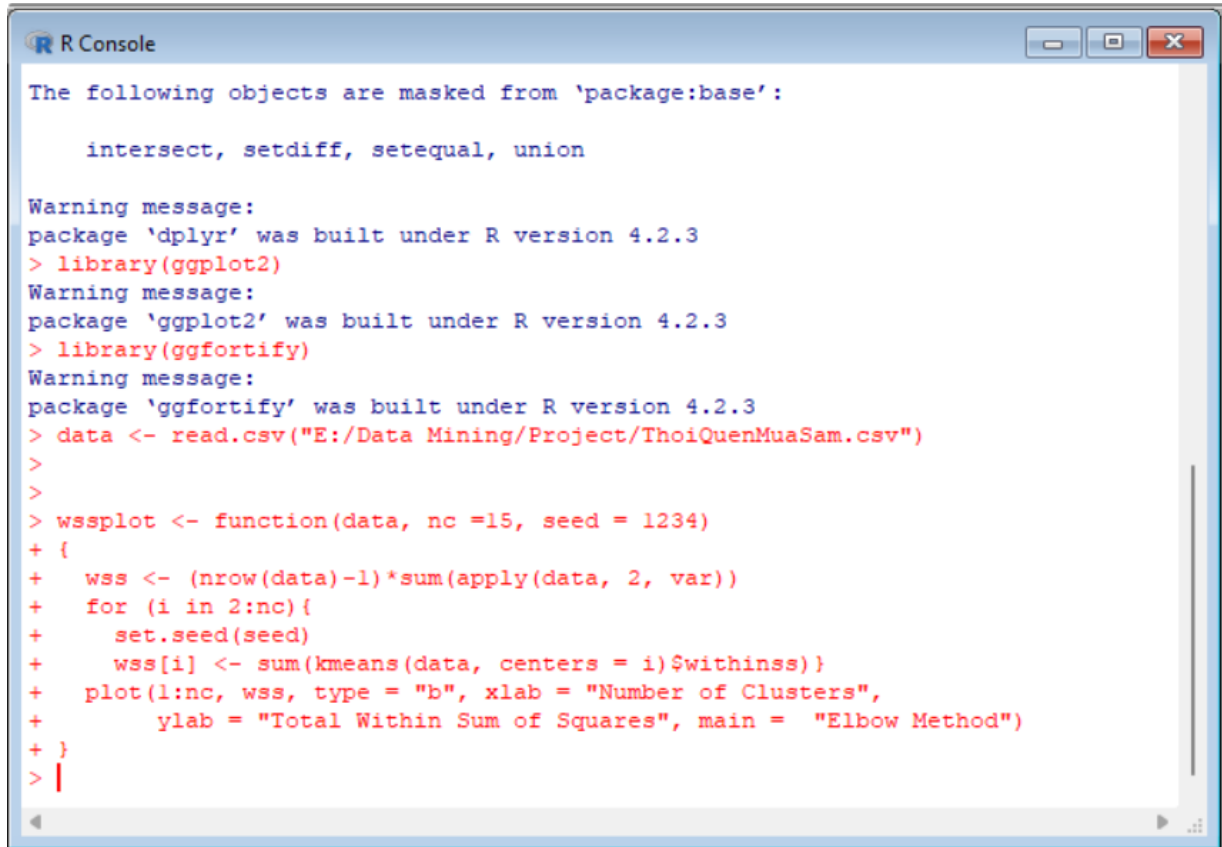
```
>
> data <- read.csv("E:/Data Mining/Project/ThoiQuenMuaSam.csv")
> |
```

Hình 4. 45: Đọc dữ liệu bằng hàm `read.csv` và lưu vào biến `data`

- ❖ **Bước 4:** Viết hàm `wssplot` để vẽ biểu thị biểu diễn sự thay đổi của tổng bình phương khoảng cách từ các điểm đến tâm cụm, tương ứng với số lượng cụm khác

Sử dụng thuật toán K-Means để phân tích khách hàng dựa trên hành vi mua hàng trên TiktokShop

nhau (số lượng cụm được thử nghiệm từ 1 đến nc) của thuật toán K-means. Kết quả được lưu vào biến wss.



```
R Console

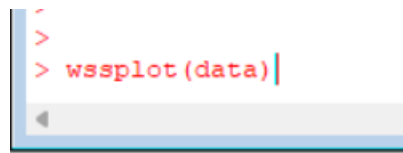
The following objects are masked from 'package:base':

  intersect, setdiff, setequal, union

Warning message:
package 'dplyr' was built under R version 4.2.3
> library(ggplot2)
Warning message:
package 'ggplot2' was built under R version 4.2.3
> library(ggfortify)
Warning message:
package 'ggfortify' was built under R version 4.2.3
> data <- read.csv("E:/Data Mining/Project/ThoiQuenMuaSam.csv")
>
>
> wssplot <- function(data, nc =15, seed = 1234)
+ {
+   wss <- (nrow(data)-1)*sum(apply(data, 2, var))
+   for (i in 2:nc){
+     set.seed(seed)
+     wss[i] <- sum(kmeans(data, centers = i)$withinss)}
+   plot(1:nc, wss, type = "b", xlab = "Number of Clusters",
+        ylab = "Total Within Sum of Squares", main = "Elbow Method")
+ }
> |
```

Hình 4. 46: Viết hàm “wssplot”

- ❖ **Bước 5:** Gọi hàm “wssplot” vừa tạo ở trên để vẽ biểu đồ, và dữ liệu là từ biến “data”.



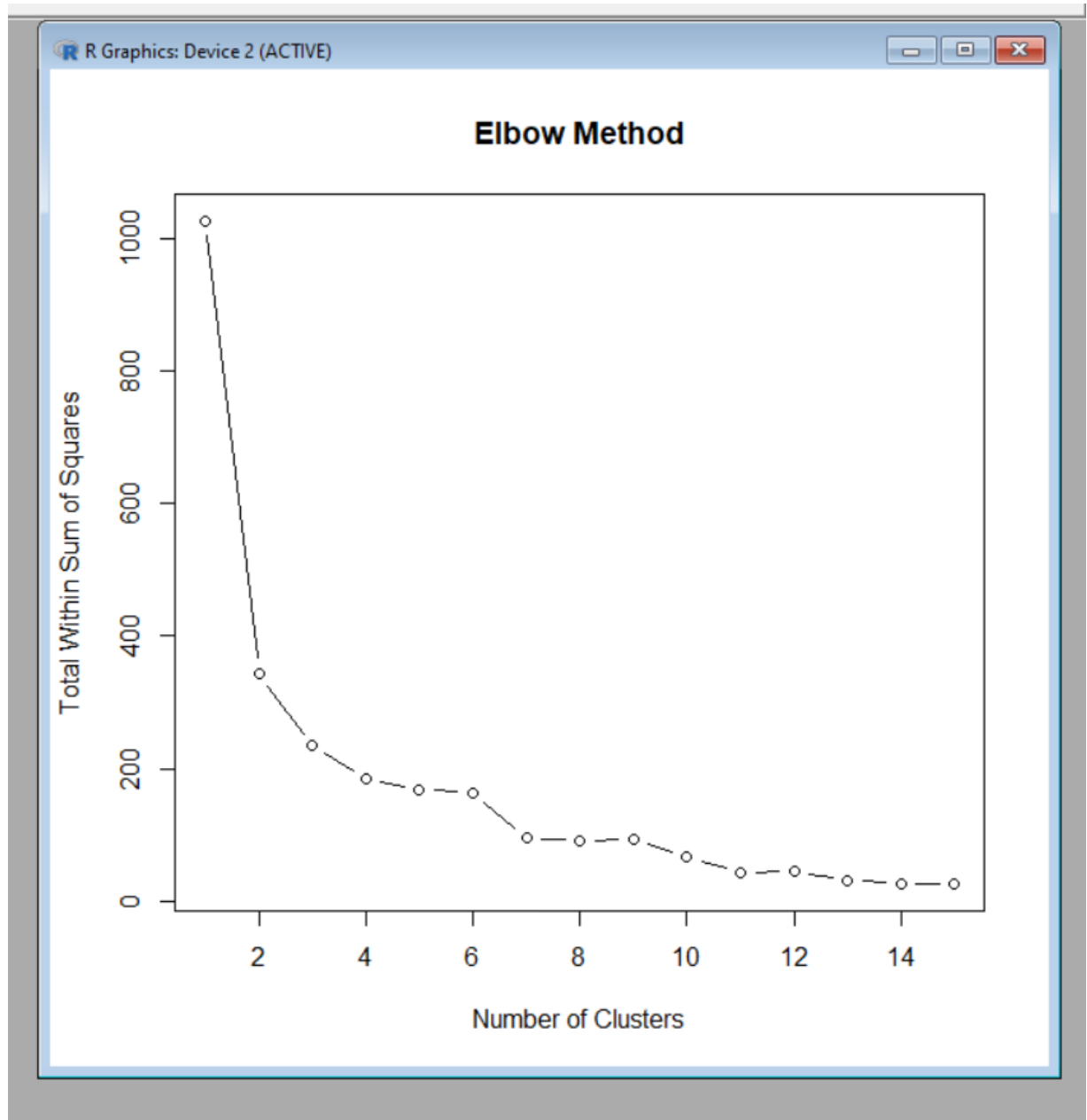
```
>
>
> wssplot(data) |
```

Hình 4. 47: Gọi hàm “wssplot”

Sau khi chạy thì sẽ ảnh biểu đồ, sẽ hiện lên. Tại đây, có thể sử dụng phương pháp “Elbow” để xác định số cụm của k.

Sử dụng thuật toán K-Means để phân tích khách hàng dựa trên hành vi mua hàng trên TiktokShop

Dựa vào biểu đồ ta có thể thấy tại số cụm $k = 4$. Khi ta tăng k lên thì tốc độ suy giảm của hàm biến dạng đường như không đáng kể so với trước đó. Do đó sẽ chọn $k = 4$.



Hình 4. 48: Sử dụng “Elbow Method” để xác định số cụm dựa vào biểu đồ được hiện ra

- ❖ **Bước 6:** Chạy thuật toán K-Means trên data với số lượng cụm là 4 và lưu vào biến KM.

Sử dụng thuật toán K-Means để phân tích khách hàng dựa trên hành vi mua hàng trên TiktokShop

```
> wssplot(data, 1:10)
> 
> KM = kmeans(data, 4)
> |
```

Hình 4. 49: Chạy thuật toán K-Means với số cụm là 4, lưu vào biến KM

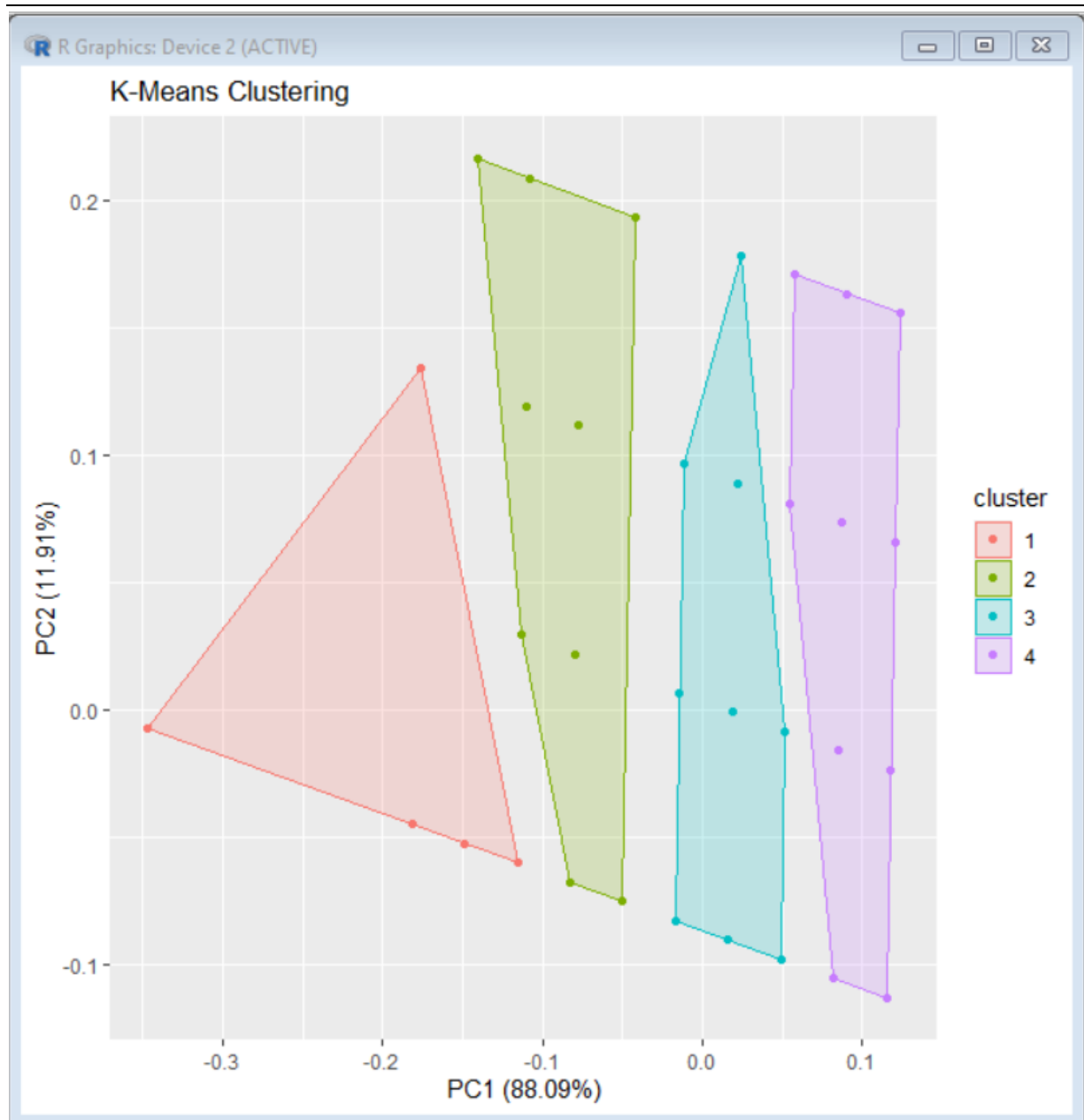
- ❖ **Bước 7:** Sử dụng hàm autoplot của ggfortify để vẽ biểu đồ scatter plot của dữ liệu mydata theo cụm đã được phân chia bởi thuật toán K-means. Biểu đồ có khung (frame) và màu sắc được đặt theo thuật toán phân cụm.

```
> autoplot(KM, data, frame = TRUE, main = "K-Means Clustering")
> |
```

Hình 4. 50: Sử dụng hàm "autoplot" của thư viện "ggfortify" để vẽ biểu đồ trực quan

Biểu đồ trực quan hóa dữ liệu sẽ vẽ như sau:

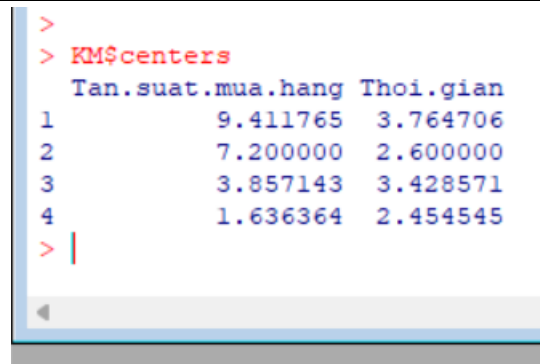
Sử dụng thuật toán K-Means để phân tích khách hàng dựa trên hành vi mua hàng trên TiktokShop



Hình 4. 51: Dữ liệu được trực quan hóa với 4 cụm

- ❖ **Bước 8:** Có thể chạy lệnh như dưới để trả về các tâm của các cụm đã được phân chia bởi K-means và lưu vào `KM$centers`.

Sử dụng thuật toán K-Means để phân tích khách hàng dựa trên hành vi mua hàng trên TiktokShop



```
>
> KM$centers
Tan.suat.mua.hang Thoi.gian
1          9.411765  3.764706
2          7.200000  2.600000
3          3.857143  3.428571
4          1.636364  2.454545
> |
```

Hình 4. 52: Dữ liệu tâm các cụm được trả về

Giải thích kết quả:

Tại cụm 1: Nhóm khách trong cụm này, có tần suất mua 9 lần/tháng, và vào trong khoảng thời gian là ban đêm.

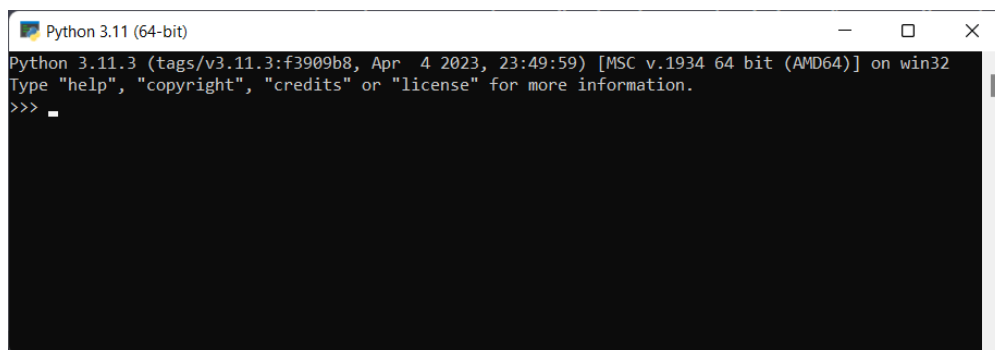
Tại cụm 2: Đối với nhóm khách hàng trong cụm đây, mỗi tháng họ sẽ mua hàng trên TiktokShop 7 lần và vào buổi chiều.

Tại cụm 3: Những khách hàng ở cụm này, sẽ mua hàng với tần suất là 4 lần mỗi tháng, và sẽ mua vào lúc chiều

Tại cụm thứ 4: Những khách hàng trong cụm này sẽ có điểm chung, mua hàng vào thời gian là buổi trưa và tần suất 2 lần mỗi tháng.

4.9 Chạy thuật toán K-Means trên Python

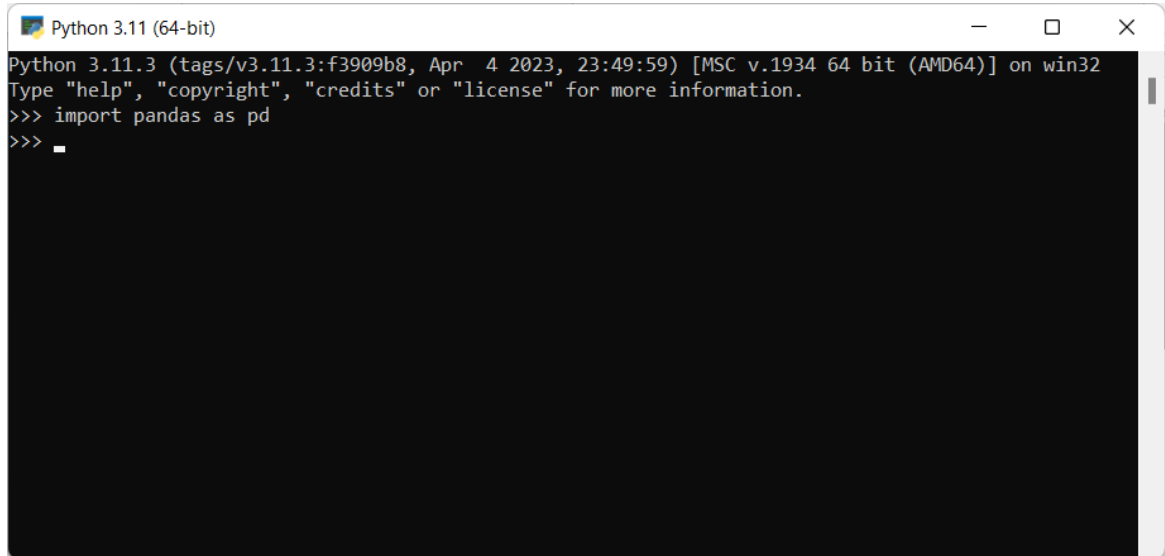
❖ Bước 1: Mở Python lên:



Hình 4. 53: Mở Python

Sử dụng thuật toán K-Means để phân tích khách hàng dựa trên hành vi mua hàng trên TiktokShop

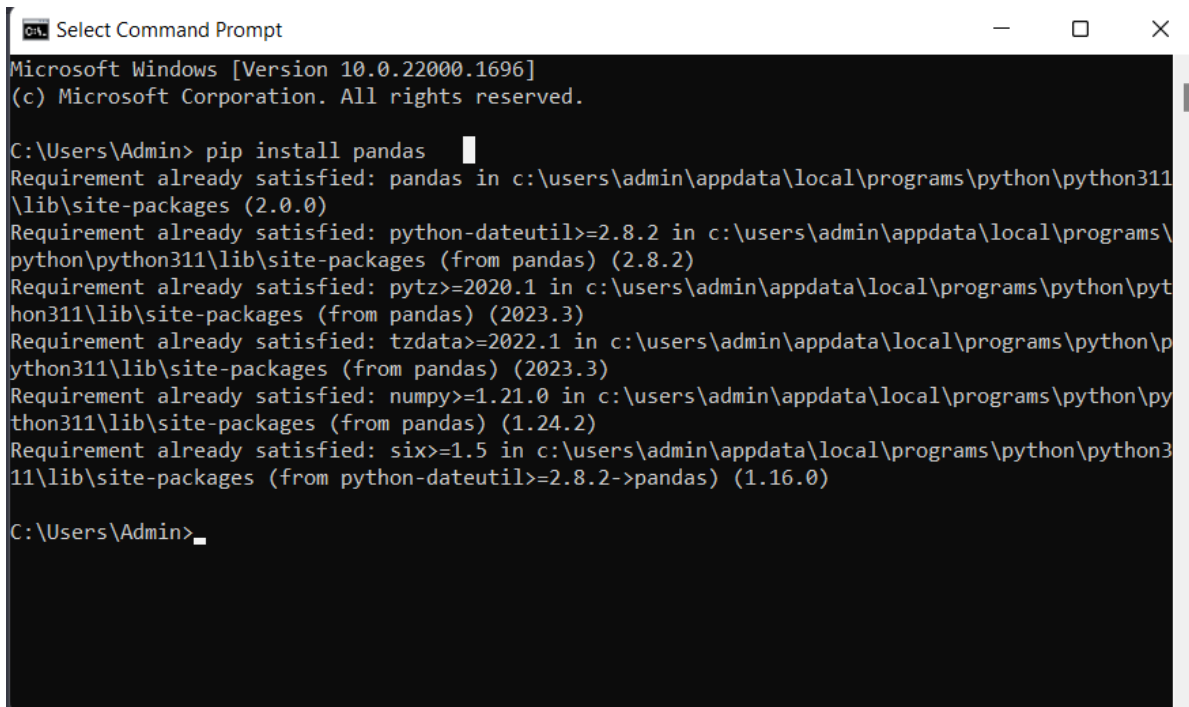
❖ **Bước 2:** Import thư viện “Pandas” vào trong Python với lệnh sau:



```
Python 3.11 (64-bit)
Python 3.11.3 (tags/v3.11.3:f3909b8, Apr  4 2023, 23:49:59) [MSC v.1934 64 bit (AMD64)] on win32
Type "help", "copyright", "credits" or "license" for more information.
>>> import pandas as pd
>>> _
```

Hình 4. 54: Import thư viện “Pandas” vào Python

Nếu chưa cài đặt thư viện pandas, có thể sử dụng pip để cài đặt. Để cài đặt pandas, hãy mở command prompt, gõ lệnh sau và enter:



```
Select Command Prompt
Microsoft Windows [Version 10.0.22000.1696]
(c) Microsoft Corporation. All rights reserved.

C:\Users\Admin> pip install pandas
Requirement already satisfied: pandas in c:\users\admin\appdata\local\programs\python\python311\lib\site-packages (2.0.0)
Requirement already satisfied: python-dateutil>=2.8.2 in c:\users\admin\appdata\local\programs\python\python311\lib\site-packages (from pandas) (2.8.2)
Requirement already satisfied: pytz>=2020.1 in c:\users\admin\appdata\local\programs\python\python311\lib\site-packages (from pandas) (2023.3)
Requirement already satisfied: tzdata>=2022.1 in c:\users\admin\appdata\local\programs\python\python311\lib\site-packages (from pandas) (2023.3)
Requirement already satisfied: numpy>=1.21.0 in c:\users\admin\appdata\local\programs\python\python311\lib\site-packages (from pandas) (1.24.2)
Requirement already satisfied: six>=1.5 in c:\users\admin\appdata\local\programs\python\python311\lib\site-packages (from python-dateutil>=2.8.2->pandas) (1.16.0)

C:\Users\Admin>
```

Hình 4. 55: Cài đặt thư viện “Pandas” bằng pip trong “cmd”

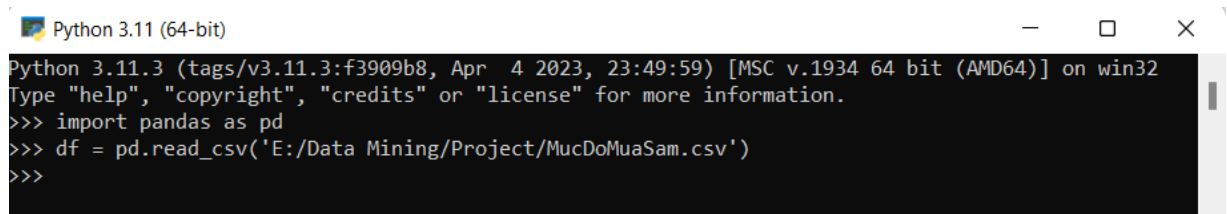
Sử dụng thuật toán K-Means để phân tích khách hàng dựa trên hành vi mua hàng trên TiktokShop

Pandas là một thư viện mã nguồn mở được sử dụng trong Python để xử lý và phân tích dữ liệu.

Pandas cung cấp các cấu trúc dữ liệu linh hoạt như Series (dữ liệu một chiều) và DataFrame (dữ liệu hai chiều) để phân tích và xử lý dữ liệu, cũng như cung cấp nhiều chức năng để thực hiện các thao tác trên dữ liệu như lọc, sắp xếp, tính toán và ghép nối dữ liệu.

Pandas cũng cho phép đọc và ghi dữ liệu từ nhiều định dạng tập tin như CSV, Excel, SQL và nhiều hơn nữa. Pandas là một thư viện rất hữu ích cho các nhà khoa học dữ liệu và các nhà phân tích dữ liệu để thực hiện các nhiệm vụ xử lý dữ liệu.

- ❖ **Bước 3:** Đọc file “MucDoMuaSam.csv” bằng hàm `read_csv` (dùng để đọc file csv) của thư viện “Pandas”:



```
Python 3.11 (64-bit)
Python 3.11.3 (tags/v3.11.3:f3909b8, Apr 4 2023, 23:49:59) [MSC v.1934 64 bit (AMD64)] on win32
Type "help", "copyright", "credits" or "license()" for more information.
>>> import pandas as pd
>>> df = pd.read_csv('E:/Data Mining/Project/MucDoMuaSam.csv')
>>>
```

Hình 4. 56: Đọc file bằng hàm `read_csv` của thư viện “Pandas”

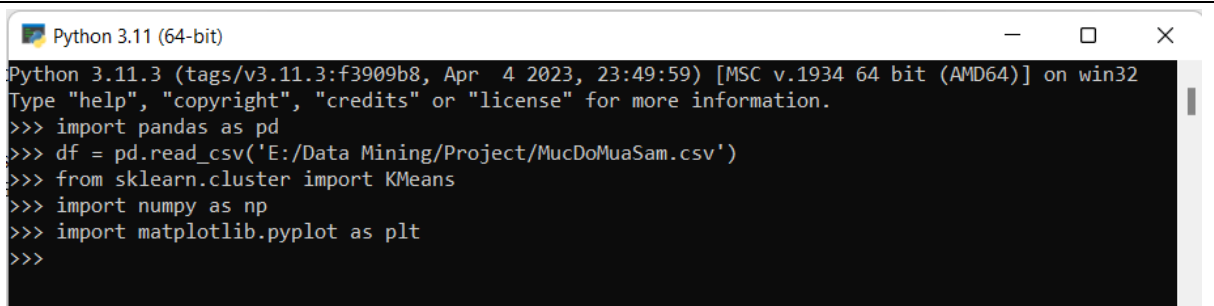
- ❖ **Bước 4:** Tiếp theo, Import các thư viện “sklearn” (để áp dụng thuật toán Kmeans), “numpy” (để xử lý dữ liệu và tính toán), “matplotlib” (để trực quan hóa dữ liệu với biểu đồ).

Nếu chưa có thì có thể mở cmd và sử dụng pip để cài đặt. Như bước 2, đã hướng dẫn cài đặt pandas. Các lệnh cài đặt như sau:

- Để cài đặt “sklearn”: `pip install scikit-learn`
- Để cài đặt “numpy”: `pip install numpy`
- Để cài đặt “matplotlib”: `pip install matplotlib`

Sau khi cài đặt xong, có thể Import vào Python như sau:

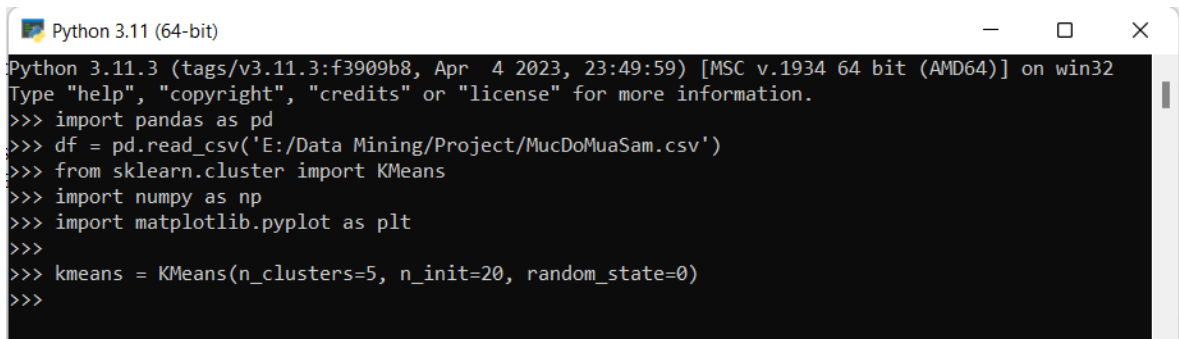
Sử dụng thuật toán K-Means để phân tích khách hàng dựa trên hành vi mua hàng trên TiktokShop



```
Python 3.11 (64-bit)
Python 3.11.3 (tags/v3.11.3:f3909b8, Apr 4 2023, 23:49:59) [MSC v.1934 64 bit (AMD64)] on win32
Type "help", "copyright", "credits" or "license" for more information.
>>> import pandas as pd
>>> df = pd.read_csv('E:/Data Mining/Project/MucDoMuaSam.csv')
>>> from sklearn.cluster import KMeans
>>> import numpy as np
>>> import matplotlib.pyplot as plt
>>>
```

Hình 4. 57: Import thư viện “sklearn”, “numpy”, “matplotlib” vào Python

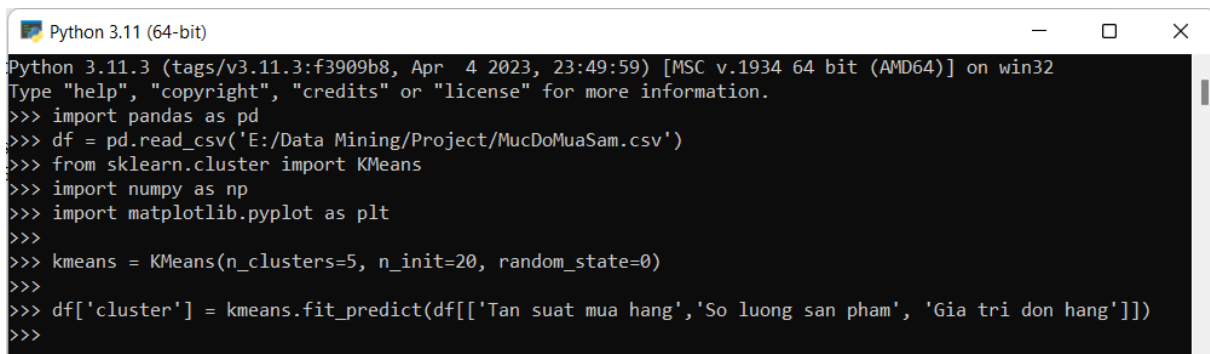
- ❖ **Bước 5:** Khởi tạo một đối tượng KMeans với 5 cụm ($n_clusters = 5$), thực hiện 20 lần khởi tạo cụm khác nhau ($n_init=20$), và thiết lập seed cho quá trình ngẫu nhiên ($random_state=0$).



```
Python 3.11 (64-bit)
Python 3.11.3 (tags/v3.11.3:f3909b8, Apr 4 2023, 23:49:59) [MSC v.1934 64 bit (AMD64)] on win32
Type "help", "copyright", "credits" or "license" for more information.
>>> import pandas as pd
>>> df = pd.read_csv('E:/Data Mining/Project/MucDoMuaSam.csv')
>>> from sklearn.cluster import KMeans
>>> import numpy as np
>>> import matplotlib.pyplot as plt
>>>
>>> kmeans = KMeans(n_clusters=5, n_init=20, random_state=0)
>>>
```

Hình 4. 58: Khởi tạo đối tượng kmeans

- ❖ **Bước 6:** Thực hiện phân cụm dữ liệu trên các đặc trưng “Tần suất mua hàng”, “Số lượng sản phẩm” và “Giá trị đơn hàng”. Và gán nhãn cụm cho mỗi mẫu trong DataFrame “df”. Nhãn cụm được gán vào cột cluster của DataFrame:



```
Python 3.11 (64-bit)
Python 3.11.3 (tags/v3.11.3:f3909b8, Apr 4 2023, 23:49:59) [MSC v.1934 64 bit (AMD64)] on win32
Type "help", "copyright", "credits" or "license" for more information.
>>> import pandas as pd
>>> df = pd.read_csv('E:/Data Mining/Project/MucDoMuaSam.csv')
>>> from sklearn.cluster import KMeans
>>> import numpy as np
>>> import matplotlib.pyplot as plt
>>>
>>> kmeans = KMeans(n_clusters=5, n_init=20, random_state=0)
>>>
>>> df['cluster'] = kmeans.fit_predict(df[['Tần suất mua hàng', 'Số lượng sản phẩm', 'Giá trị đơn hàng']])
>>>
```

Hình 4. 59: Thực hiện phân cụm dựa trên các thuộc tính

Sử dụng thuật toán K-Means để phân tích khách hàng dựa trên hành vi mua hàng trên TiktokShop

❖ **Bước 7:** In ra các tọa độ của các tâm cụm được tìm thấy bởi thuật toán Kmeans:

```
>>> print(kmeans.cluster_centers_)
[[3.31034483  3.44827586  4.34482759]
 [8.6         3.9         8.9        ]
 [8.27272727  2.22727273  3.13636364]
 [2.5         1.30769231  2.15384615]
 [2.15384615  1.53846154  6.69230769]]
>>> _
```

Hình 4. 60: In tọa độ tâm cụm

Đây chính là tọa độ tâm cụm được thực hiện bởi thực toán Kmeans, với các giá trị theo thứ tự “Tần suất mua hàng”, “Số lượng sản phẩm”, “Giá trị đơn hàng”.

❖ **Bước 8:** Để thực hiện trực quan hóa dữ liệu, đầu tiên cần tạo 5 màu và gán màu sắc tương ứng cho từng cụm:

```
[2.5         1.30769231  2.15384615]
[2.15384615  1.53846154  6.69230769]]
>>>
>>> colors = ['#001949', '#5567C9', '#8D9AC5', '#FFC2B8', '#E17A8D']
>>> df['c'] = df.cluster.map({0:colors[0], 1:colors[1], 2:colors[2], 3:colors[3], 4:colors[4]})
>>> _
```

Hình 4. 61: Gán màu tương ứng từng cụm

❖ **Bước 9:** Tiếp theo, sử dụng thư viện “matplotlib” để tạo ra một hình ảnh (figure) với kích thước là (26, 6) và thêm một trục 3D (axis) vào vị trí thứ nhất trong lưới 1x3 (trong số 3 trục) của hình ảnh đó để hiển thị dữ liệu 3 chiều.

```
Select Python 3.11 (64-bit)
>>> df['cluster'] = kmeans.fit_predict(df[['Tan suat mua hang', 'So luong san pham', 'Gia tri don hang']])
>>>
>>> print(kmeans.cluster_centers_)
[[3.31034483  3.44827586  4.34482759]
 [8.6         3.9         8.9        ]
 [8.27272727  2.22727273  3.13636364]
 [2.5         1.30769231  2.15384615]
 [2.15384615  1.53846154  6.69230769]]
>>>
>>> colors = ['#001949', '#5567C9', '#8D9AC5', '#FFC2B8', '#E17A8D']
>>> df['c'] = df.cluster.map({0:colors[0], 1:colors[1], 2:colors[2], 3:colors[3], 4:colors[4]})
>>>
>>> fig = plt.figure(figsize=(26 , 6))
>>> axis = fig.add_subplot(131, projection='3d')
>>> _
```

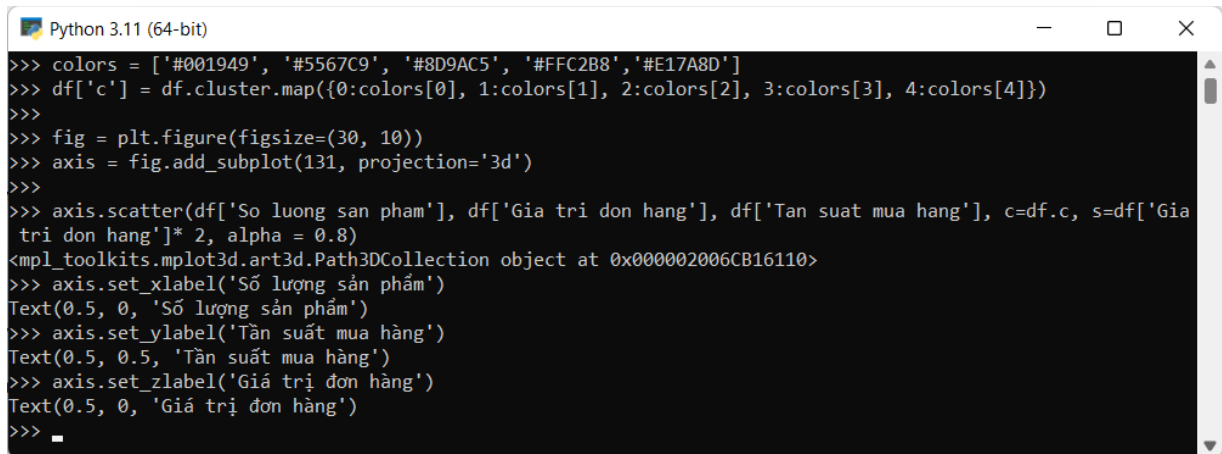
Hình 4. 62: Tạo hình ảnh và thêm một trục 3D

Sử dụng thuật toán K-Means để phân tích khách hàng dựa trên hành vi mua hàng trên TiktokShop

❖ **Bước 10:** Truyền dữ liệu vào hình ảnh 3D và thiết lập các trục tương ứng với các thuộc tính của dữ liệu:

- `axis.scatter()`: tạo ra một biểu đồ scatter plot với các trục tương ứng với các thuộc tính của dữ liệu, với các giá trị tương ứng được truyền vào từ dataframe `df`.
- `c=df.c`: đặt giá trị màu của từng điểm trên biểu đồ bằng cách sử dụng thuộc tính “cluster” trong dataframe “`df`”, sau đó dựa vào giá trị này để truy xuất giá trị màu tương ứng trong danh sách colors đã được định nghĩa trước đó.
- `s=df['Giá trị đơn hàng']* 2`: đặt kích thước của từng điểm trên biểu đồ dựa vào thuộc tính “Giá trị đơn hàng” của dữ liệu, Nhưng do các điểm dữ liệu quá nhỏ, nên sẽ gấp đôi lên cho dễ dàng nhìn thấy.
- `alpha = 0.8`: đặt mức độ trong suốt của từng điểm trên biểu đồ.

Sau đó, thông qua các hàm `.set_xlabel()`, `.set_ylabel()`, `.set_zlabel()`, để đặt tên cho các trục x, y, z tương ứng.



```
>>> colors = ['#001949', '#5567C9', '#8D9AC5', '#FFC2B8', '#E17A8D']
>>> df['c'] = df.cluster.map({0:colors[0], 1:colors[1], 2:colors[2], 3:colors[3], 4:colors[4]})
>>>
>>> fig = plt.figure(figsize=(30, 10))
>>> axis = fig.add_subplot(131, projection='3d')
>>>
>>> axis.scatter(df['Số lượng sản phẩm'], df['Tần suất mua hàng'], df['Giá trị đơn hàng'], c=df.c, s=df['Giá
trị đơn hàng']* 2, alpha = 0.8)
<mpl_toolkits.mplot3d.art3d.Path3DCollection object at 0x000002006CB16110>
>>> axis.set_xlabel('Số lượng sản phẩm')
Text(0.5, 0, 'Số lượng sản phẩm')
>>> axis.set_ylabel('Tần suất mua hàng')
Text(0.5, 0.5, 'Tần suất mua hàng')
>>> axis.set_zlabel('Giá trị đơn hàng')
Text(0.5, 0, 'Giá trị đơn hàng')
>>>
>>>
```

Hình 4. 63: Truyền dữ liệu vào hình ảnh và thiết lập các trục tương ứng các thuộc tính

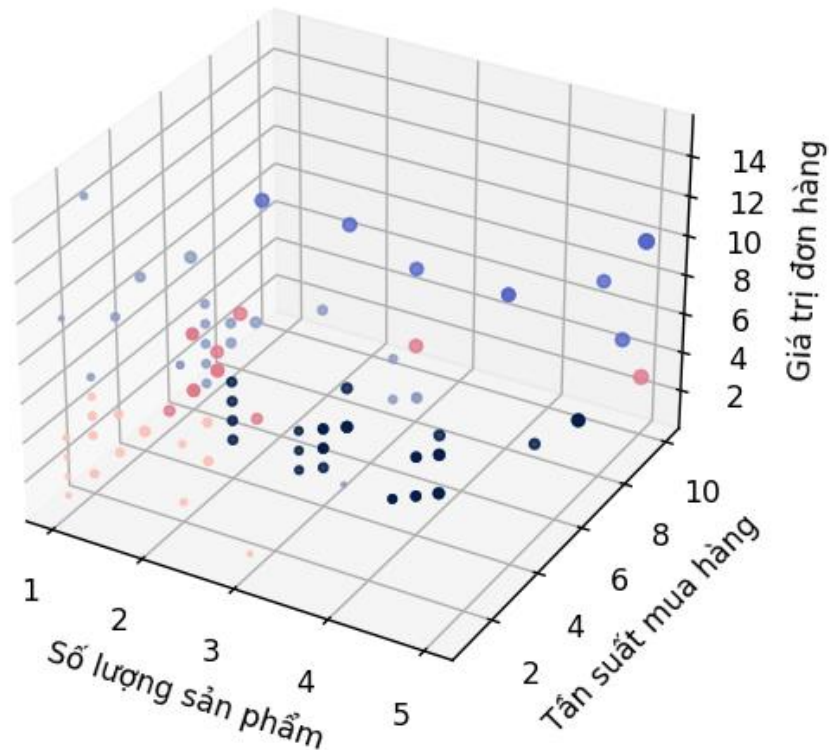
Sử dụng thuật toán K-Means để phân tích khách hàng dựa trên hành vi mua hàng trên TiktokShop

❖ **Bước 11:** Hiện thị hình ảnh đã được vẽ trên màn hình:

```
Select Python 3.11 (64-bit)
>>>
>>> fig = plt.figure(figsize=(30, 10))
>>> axis = fig.add_subplot(131, projection='3d')
>>>
>>> axis.scatter(df['Số lượng sản phẩm'], df['Giá trị đơn hàng'], df['Tần suất mua hàng'], c=df.c, s=df['Giá trị đơn hàng']* 2, alpha = 0.8)
<matplotlib.collections.Path3DCollection object at 0x000002006CB16110>
>>> axis.set_xlabel('Số lượng sản phẩm')
Text(0.5, 0, 'Số lượng sản phẩm')
>>> axis.set_ylabel('Tần suất mua hàng')
Text(0.5, 0.5, 'Tần suất mua hàng')
>>> axis.set_zlabel('Giá trị đơn hàng')
Text(0.5, 0, 'Giá trị đơn hàng')
>>>
>>> plt.show()
```

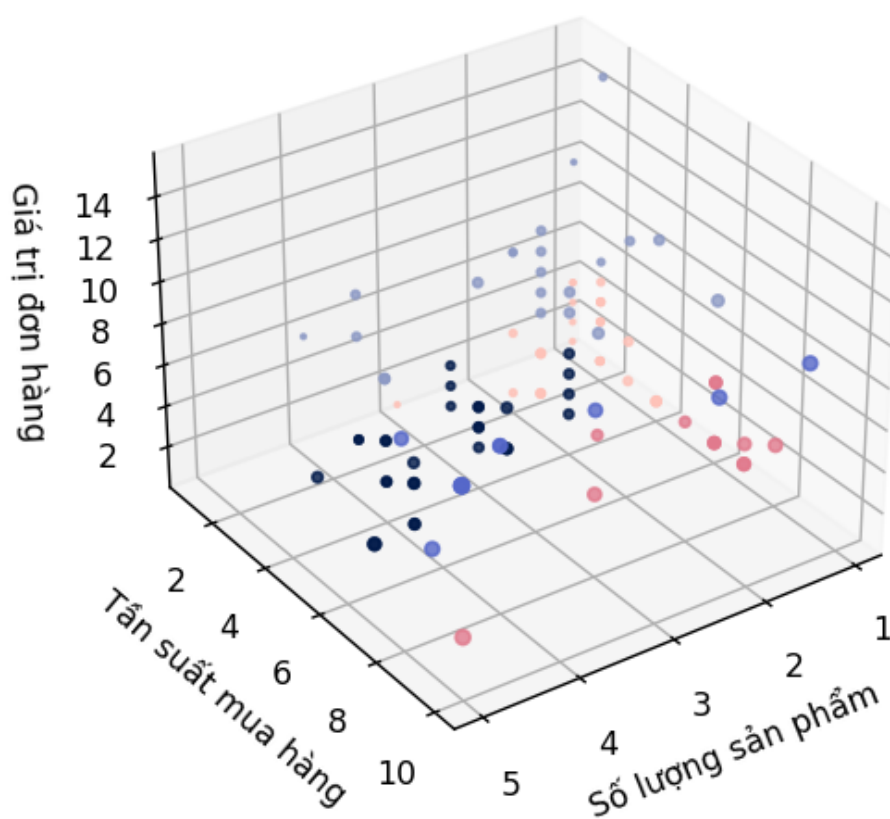
Hình 4. 64: Lệnh hiển thị hình ảnh

Và đây là hình đồ thị 3D đã được vẽ ra:



Hình 4. 65: Đồ thị 3D trực quan hóa 1

Sử dụng thuật toán K-Means để phân tích khách hàng dựa trên hành vi mua hàng trên TiktokShop



Hình 4. 66: Đồ thị 3D trực quan hóa 2

CHƯƠNG 5. KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

5.1 Kết luận

5.1.1 Những kết quả đạt được

- Tìm và hiểu được các lý thuyết liên quan đến môn học.
 - + Kho dữ liệu.
 - + Data Warehouse Schema.
 - + Mô hình dữ liệu đa chiều
 - + Khai phá dữ liệu: khái niệm, quy trình khai phá dữ liệu, phương pháp và ứng dụng khai phá dữ liệu.
- Nêu được phương pháp khai phá trong đề tài, và cho ví dụ cụ thể và minh họa bằng từng bước thực hiện.
- Hiểu và ứng dụng được thuật toán K-Means.
- Có thực hiện các khảo sát để lấy dữ liệu thông tin về hành vi mua hàng của khách hàng, hiểu được dữ liệu, chuẩn bị dữ liệu, tiền xử lý dữ liệu nhằm loại bỏ những giá trị gây nhiễu cho quá trình phân tích thuật toán.
- Ứng dụng được các công cụ để chạy thuật toán K-Means và trực quan hóa các dữ liệu thuật toán:
 - + Weka.
 - + Weka KnowledgeFlow.
 - + R.
 - + Python.

5.1.2 Hạn chế

- Do tệp dữ liệu chỉ dừng lại ở 100 khảo sát của các khách hàng, nên có thể dữ liệu thuật toán chưa hoàn toàn chính xác, vẫn có sai số nhỏ có thể xảy ra.
- Bài toán đặt ra vẫn chưa cụ thể, rõ ràng.

Sử dụng thuật toán K-Means để phân tích khách hàng dựa trên hành vi mua hàng trên TiktokShop

5.2 Hướng phát triển

5.2.1 Hướng khắc phục những hạn chế

- Để khắc phục hạn chế đầu tiên, chúng ta cần có thêm dữ liệu từ nhiều khách hàng hơn để tăng tính chính xác cho phân tích. Có thể sử dụng các kênh khác nhau để thu thập dữ liệu, chẳng hạn như khảo sát trực tuyến hoặc phân tích dữ liệu từ các nền tảng mạng xã hội khác nhau để thu thập dữ liệu về hành vi mua hàng của khách hàng.
- Để khắc phục hạn chế thứ hai, chúng ta cần phải làm rõ và cụ thể hóa mục tiêu và câu hỏi nghiên cứu của bài toán.

5.2.2 Hướng mở rộng của đề tài

Để mở rộng đề tài, có thể kết hợp thuật toán Kmeans với các phương pháp học máy khác như học sâu, mạng neuron, máy vector hỗ trợ (SVM), và Random Forest để tăng độ chính xác và khả năng dự báo. Ngoài ra, cũng có thể kết hợp với phân tích định tính và dùng các phương pháp phân tích khác như phân tích thành phần chính (PCA), phân tích đường cong khả năng sinh (ROC), để đánh giá tính hiệu quả của thuật toán phân tích khách hàng trên nền tảng Tiktokshop.

TÀI LIỆU THAM KHẢO

- [1]. ThS Thái Thị Ngọc Lý, *Bài giảng Khai phá dữ liệu*, Khoa Công nghệ thông tin, Trường Đại Học Tài Chính – Marketing.
- [2]. Jake VanderPlas (2016), *Python Data Science Handbook*, O'Reilly Media.
- [3]. Hadley Wickham & Garrett Grolemund (2017), *R for Data Science*, O'Reilly Media.
- [4]. Sebastian Raschka & Vahid Mirjalili (2017), *Python Machine Learning*, Packt Publishing.